Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze Biotecnologiche, Biocomputazionali, Farmaceutiche e Farmacologiche

Ciclo XXXIV

Settore Concorsuale: 05/E1 Biochimica generale

Settore Scientifico Disciplinare: BIO/10 Biochimica

Machine Learning Tools for Protein Annotation:
the cases of transmembrane β-barrel and myristoylated proteins

Presentata da: Giovanni Madeo

Coordinatore Dottorato:

Prof.ssa Maria Laura Bolognesi

Supervisore:

Prof. Pier Luigi Martelli

Correlatore:

Prof. Castrense Savojardo

Esame finale anno 2022

Abstract

Biology is now a member of the so-called “Big Data Sciences” thanks to technological advancements that allow to fully characterize the macromolecular content of a cell or a collection of cells, generating a huge amount of data with time and costs constantly decreasing. This opens interesting perspectives, but only a small portion of this data may be experimentally characterized. From this derives the demand of accurate and efficient computational tools for automatic annotation of biological molecules. This is even more true when dealing with membrane proteins, on which my research project is focused leading to the development of two machine learning-based methods (both made available through web server): BetAware-Deep and SVMyr.

BetAware-Deep is a tool for the detection and topology prediction of transmembrane beta-barrel proteins found in the outer membrane of Gram-negative bacteria. These proteins are of particular interest, being involved in many biological processes and primary candidates as drug targets. BetAware-Deep exploits the combination of a deep learning framework (bidirectional long short-term memory) and a probabilistic graphical model (grammatical-restrained hidden conditional random field). Moreover, it introduced a modified formulation of the hydrophobic moment, designed to include the evolutionary information. BetAware-Deep outperformed all the available methods in topology prediction and reported high scores in the detection task.

Glycine myristoylation in Eukaryotes is the binding of a myristic acid on an N-terminal glycine. SVMyr is a fast method based on support vector machines designed to predict this modification co- and post-translationally in dataset of proteomic scale. It uses as input octapeptides and exploits computational scores derived from experimental examples and mean physicochemical features. SVMyr outperformed all the available methods for co-translational myristoylation prediction. In addition, it allows (as a unique feature) the prediction of post-translational myristoylation.

Both the tools here described are designed having in mind best practices for the development of machine learning-based tools outlined by the bioinformatics community. Moreover, they are made available via user-friendly web servers. All this make them valuable tools for filling the gap between sequential and annotated data.

Summary

1. Introduction

The last few decades have been characterized by an ever-growing interest in the so-called "Omics sciences". This word refers to the branches of biochemistry and molecular biology aiming to collectively characterize biological systems by detecting and quantifying a variety of molecules, such as genes (genomics), proteins (proteomics), messenger RNA (transcriptomics) or metabolites (metabolomics). Omics sciences are also devised to the understanding of the interactions among the different molecules present in the cell (protein-protein/protein-nucleic acids/protein-small molecule interactomics), to provide a thorough description of the mechanisms at the basis of complex biological processes. This knowledge paves the way to an enormous number of applications in different fields, such as precision medicine, novel drug discovery, drug repurposing, genetic selection in agri-food productive systems.

Technological advancements in high-throughput techniques adopted in omics sciences allow the production of a huge amount of data with constantly reducing cost and time. This has brought in biology the concept of Big Data, which may be defined as the availability of large, complex, diverse and multi-dimensional, structured or unstructured datasets [1].

The advent of Big Data in biology opens new opportunities, as well as new challenges. In fact, our ability of producing data in biology is now exceeding our ability of storing, analyzing, and integrating them [1], and this gap is expected to increase in the next years [2]. Moreover, issues raised by Big Data in biology are not only relative to data size, but also to their growing complexity [3]. On the other hand, the access to such data in a cost- and time-efficient manner paves the way to precision medicine, namely the customization of medical treatment for individual patients, and other applications, as mentioned. All these considerations make clear that we need to put efforts in developing efficient tools able to deal with the data deluge we are facing.

One of the main issues concerning Omics Big Data concerns molecules still uncharacterized, which need to be annotated *i.e.*, endowed with structural and functional information. Out of 225,578,953 proteins collected in version 2021_04 (Sep 2021) of UniProtKB (565,928 manually annotated in SwissProt and 225,013,025 automatically annotated in TrEMBL), only 54,943 (30,970 from SwissProt, 23,973 from TrEMBL) are endowed with an experimental structure, at least partial, in the Protein Data Bank (PDB). Therefore, experimental structural information is available for only 0.02% of proteins (5.5% in the SwissProt set). When looking at functional annotations, 1,214,501 UniProtKB entries report Gene Ontology (GO) terms, endowed with a "manual assertion" evidence code. The rate of curated functional annotation is therefore 0.5%. When analyzing the SwissProt curated section, the rate increases to 25.5% (144,482 proteins). The gold-standard approaches for functional/structural annotation of biological molecules consist in wet-lab experiments. However, these are often costly and time-consuming, and hence not suitable of keeping up with high-throughput techniques producing tons of data every day. The picture here described makes clear the need for computational methods for fast and reliable structural and functional annotation of large datasets of biomolecules.

My PhD project focused on "Innovative Methods for the Analysis of 'Omics' Big Data" and fits in this context. This project was funded by a scholarship provided by the region Emilia-Romagna under the theme "Human Resources for a Digital Economy: Big Data". The main goal of the project was the improvement of computational tools for structural and functional annotation of biological macromolecules. Such tools, once developed and benchmarked, are made available via user-friendly and accessible web server, also designed thanks to the software engineering skills I have acquired during my internship at BioDec company (http://www.biodec.com/it).

My research work focused on the annotation of Membrane Proteins (MPs), which are proteins of great interest, having many important functions and being primary candidates as drug targets. From

this derives the interest toward proper characterization of MPs *in silico*, also given the methodological limitations that hamper a large-scale experimental characterization of such proteins.

In the next sections, I present the biological background (Section 1) on biological membranes and their constituents (proteins and lipids), briefly describing the cell membrane, the bacterial outer membrane, and the membrane-bound organelles localized inside the cell. In Section 2, I highlight the motivation of interest on MPs annotation, as a central problem in computational biology. Finally, I describe the two novel methods that constitute the main subjects of my research project. The first one regards prokaryotic Transmembrane Beta-Barrel (TMBB) proteins (Section 3), for which we developed a deep learning-based tool, named BetAware-Deep [4], designed for their detection in proteomes and the prediction of their topology. The second one addresses the problem of glycine myristoylation in proteins (Section 4), a type of lipidation occurring in Eukaryotes, whose annotation has been tackled with our SVMyr [5], a method based on Support Vector Machines (SVMs) submitted for publication.

1.1. Biological Background

1.1.1. Membrane Lipids

A biological membrane is a fundamental structure that encloses cells and cell compartments, and defines volumes with peculiar compositions and mediated interactions with the environment [6,7]. It is constituted by a lipid bilayer, namely two layers of lipid molecules. The main lipidic component of all membranes are phospholipids. Each of these molecules have a polar or hydrophilic “head”, namely a phosphate group, and two apolar or hydrophobic “tails”, two fatty acids. Given this feature, known as amphipathicity, phospholipids in an aqueous solvent spontaneously form the lipid bilayer, which is a favorable conformation exposing polar heads to the solvent, while burying apolar tails in the inner part of the bilayer [7]. In addition to phospholipids, in cell membrane contains glycolipids

and sterols. The former class of lipids is characterized by a carbohydrate bonded in the polar portion and its roles are linked to protein stability and cellular recognition. The latter class consists of a group of steroids with a distinctive shape formed by four rings and a hydroxy group, with the function of strengthening the membrane, reduce its permeability, and modulate its flexibility [6].

1.1.2. Membrane Proteins

So far, we have described the structure and function of membranes neglecting MPs, which deserve a separate discussion. Indeed, MPs account for about the 50% of the volume in most membranes and are responsible for most of their functions, acting as receptors, transporters, anchors, and enzymes [8]. Receptors are proteins able of binding a ligand (a small molecule or another proteins), mediating a cellular response upon it. Transporters mediate the movement of molecules and ions through the membrane, in an active or passive way, using the energy derived from biochemical reactions or the electrochemical gradient. Anchorage proteins mediate the joining among different cells and the adhesion to tissues. Finally, enzymes are proteins which catalyze reactions.

The central role of MPs is also confirmed by their abundance in the proteome of most organisms, accounting for 20-30% of all protein types expressed in the genome [9]. MPs are crucial for defining the specific functions of different membranes in the cell.

MPs may be classified in two major groups [8]: peripheral and integral. Peripheral MPs have transient interactions with membranes, while integral MPs permanently interact with them. Integral MPs are further divided in transmembrane and lipid-anchored MPs. In the former instance, the protein spans the membrane once (single-pass MP) or multiple times (multi-pass MPs), by means of protein segments folded in α-helical or β-strand conformations [10]. In the latter instance, the MP is covalently bound to a lipid that, integrated into the lipid bilayer, serves as an anchor. According to their lipid group, three types of lipid-anchored proteins exist: glycosylphosphatidylinositol (GPI)-

linked, prenylated, and fatty acylated proteins [11]. In figure 1, classification of MPs is graphically summarized.

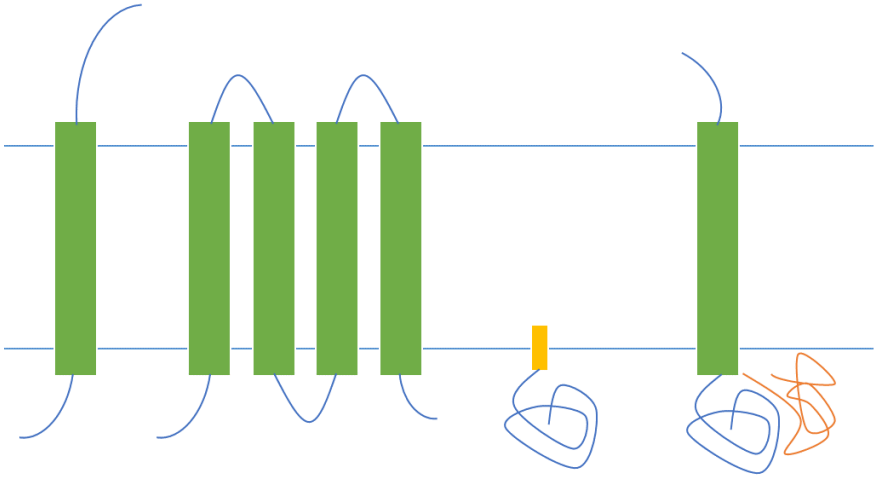

Figure 1. Examples of membrane proteins. From left to right: single-pass membrane protein; multi-pass membrane protein; lipidated protein; peripheral membrane protein (in orange).

1.1.3. Biological Membranes

One of the main characteristics shared by all living cells is the presence of a cell membrane, also called plasma membrane [7]. The main function performed by cell membranes is to create an internal environment having a different composition with respect to the external environment. This is realized by acting as a selective barrier, which controls the access and the expulsion of small molecules and ions. Nevertheless, cell membranes have a wider range of functions. In fact, they are involved in the process of cell adhesions, cell signaling, and they act as attachment surface for cell wall and cytoskeleton [6,7].

In Gram-negative bacteria, a second membrane is present, known as outer membrane, separated from the inner membrane by a space called periplasm. This lipid bilayer surrounds a thin peptidoglycans layer, and together they form the Gram-negative cell wall [12]. On the opposite, Gram-positive bacteria have a cell wall composed exclusively by peptidoglycans forming a thick layer.

The bacterial outer membrane has peculiar features that make it distinct from the cell membrane. One of the key features of outer membranes is the presence of the Outer Membrane Proteins (OMPs), transmembrane proteins which cross the membrane with beta-strand segments forming a closed structural motif resembling a barrel (transmembrane beta-barrel proteins, TMBB).

In Eukaryotes, TMBBs are also present in the outer membranes of mitochondria and plastids (in plants) [13]. Notably, these organelles are enclosed into two membrane systems (inner and outer membranes) and, following the endosymbiotic hypothesis, they have evolved from bacteria that ended up inside of other cells (host cells). Mitochondria are organelles principally involved in the production of energy for the cell, in the form of adenosine triphosphate (ATP). Plastids are organelles found in plants, with functions including photosynthesis (chloroplasts), synthesis and storage of pigments (chromoplasts), and storage of amylum (leucoplasts).

1.2. Relevance of Membrane Proteins Annotation

MPs are a class of protein of particular interest. First of all, they perform a wide range of functions and are part of many biological processes [8]. This makes them crucial for the survival of the cell. For example, transporters are fundamental for the maintenance of a different composition in the cell with respect to the surrounding environment, but also for the intake of essential molecules and the expulsion of toxic metabolites from the cell. As a further example, receptors control the cell (or tissue) response to external stimuli mediated by molecules, such as hormones and neurotransmitters. All these without even mention enzymes associated to membrane, which include a wide range of classes, such as (but not limited to): oxidoreductases, hydrolases, lyases, isomerases, transferases, and ligases.

Given the variety of molecular functions and biological processes associated to MPs, they are linked to many diseases as well, and fundamental in human health, as reviewed also in [6]. For example, cystic fibrosis arises from a mutation in the *CFTR* gene, which results in the misfolding of a Cl^- anion

channel. Moreover, MPs may be recognized by viruses, which allows them to target specific cell tissues. As a further example, TMBBs localized in the outer membrane of Gram-negative bacteria are promising target for developing antibiotics, and at the same time they are involved in mechanisms of antibiotics resistance.

MPs are targeted by some 60% of the approved drugs [14]. This is not surprising, given the wide range of functions performed by MPs and the fact that they are implied in diverse diseases. Moreover, being localized on the membrane, they are reached more easily by drugs. In fact, delivering drugs inside the cell is generally a hard issue to overcome.

Despite their relevance and the medical interest, MPs are still underrepresented in the Protein Data Bank (PDB) [15], the database collecting experimentally determined structures of macromolecules. This is principally due to technical issues encountered in the process of structural characterization. First of all, most MPs are found naturally in small quantities in membranes, and it is difficult to purify them. At the same time, it is difficult to overexpress MPs in host organisms, due to toxicity [16]. The second problem is given by the hydrophobic nature of MPs, which prevents them from being solubilized and then concentrated to crystals [17]. Detergents are required for their solubilization, but they can disrupt their structure. Moreover, the process is costly, and it is not always easy to select the right detergent for the problem at hand [17]. As reviewed in [17], much effort is spent in trying to overcome all these issues. This is done, for example, trying to make the protein soluble by substituting hydrophobic residues with hydrophilic ones [18], using stealthy artificial membranes [19,20], or combining high-resolution solid-state NMR spectroscopy with electron cryotomography [21]. The application and the further refinement of these techniques will probably give access to a larger number of structural data for MPs, but for the time being our knowledge in this field is quite limited.

Given all these considerations, the availability of accurate computational methods for MPs annotation is crucial to expand our understanding, endowing with functional and/or structural information protein sequences coming from high-throughput Omics studies.

1.3. Machine-Learning for Bioinformatics

Today, our ability of producing data using high-throughput sequencing technique is exceeding our capability of experimentally characterize them. For this reason, it is crucial to put effort in developing automated computational tools for the annotation and analysis of biological macromolecules. This task is principally addressed via the application of supervised machine-learning algorithms. These are defined as algorithms to build a model of association between an input and an output starting from data through a process called training. During this process, examples (*i.e.*, data associated with desired outputs) are provided to the method and the training algorithm adjusts the value of a (large) number of internal parameters to optimally reproduce the associations present in the provided examples. The agreement between the desired outputs and the outputs computed from the algorithm is measured with a cost (or loss) function. The result of the training procedure is a set of parameters that encode the model that best fits data in the training set. If the training set is accurate and large enough, the model generalizes the rules of association and can be applied to new independent inputs, making predictions on them.

Given the above definition and description, the difference between machine-learning algorithms and classical algorithms should be evident. In fact, in the former case we define a set of data and desired answers, from which the machine derives a set of rules. In the latter case, we input data and rules, obtaining answers based on them. Thus, machine learning overcome the intrinsic limitations of classical programming which needs to be programmed by hand, a condition incompatible with a large category of complex problems for which a clear mapping between inputs and desired outputs is unknown.

1.3.1. Supervised Classification

In bioinformatics, machine-learning methods are widely applied in different fields, including (but not limited to): genomics, proteomics, systems biology, the study of evolution, text mining, management

of complex experimental data (*e.g.*, microarray assays), primer design [22]. Different machine-learning frameworks have been defined so far, each well-suited for tackling different types of predictive tasks [22]. In the context of this thesis project, I mainly adopted algorithms for supervised classification and probabilistic graphical models for the annotation of biological sequences.

A classification problem is characterized by a collection of instances associated with classes, which are assigned given their features and a set of classification rules. Supervised classification methods are applied to automatically derive these rules starting from a collection of labelled examples in the training phase.

There are several methods belonging to this class of machine-learning algorithms: Bayesian classifiers [23], logistic regression [24], classification trees [25], nearest-neighbor classifiers [26], artificial neural networks [27], SVMs [28]. The last two methods are of particular interest for this work.

Artificial neural networks are based on a simple computing unit called neuron (Figure 2) [27]. Each neuron receives several inputs and integrates them computing an activation as a weighted sum with a threshold bias. The activation is transformed with a nonlinear transfer function. Thresholds of each neuron and weights connecting neurons are the trainable parameters of the network.

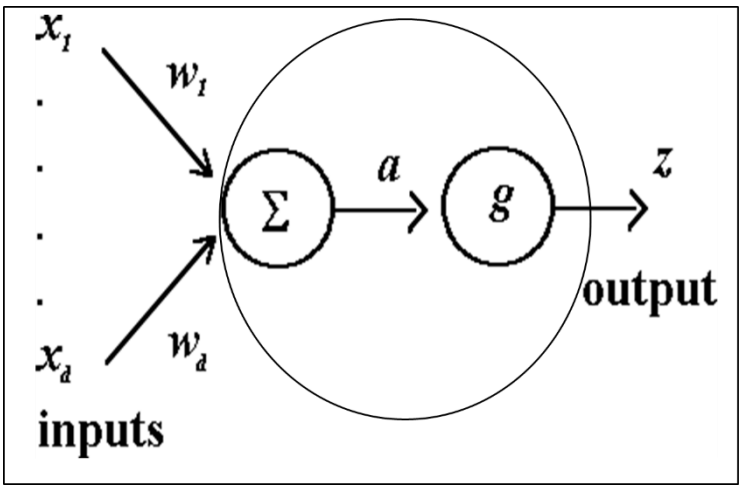

Figure 2. Schematization of the McCulloch-Pitts neuron. Inputs (x), each one with a given weight (w), are summed computing an activation (a). This is transformed by the transfer function (g) to give the output (z).

In multilayer perceptrons, neurons are organized in layers forward-linked to each other through weighted directed connections. Thus, the signal flows from the input layer to intermediate hidden layers, to finally reach the output layer.

Artificial neural networks may be used to build very complex and multi-layered architectures, which fall in the field of deep learning [29]. This family of machine-learning methods has gained great popularity in the last years, and it is currently used in a wide range of applications. For example, convolutional neural networks are an essential deep learning tool, that is broadly used in image and video processing. Another important class of methods in the deep learning field is constituted by recurrent neural networks, in which feedback loops are introduced. This makes them suitable for the analysis of sequential data. The most famous and used examples of it are Long Short-Term Memory (LSTM) models [30] and gated recurrent units [31].

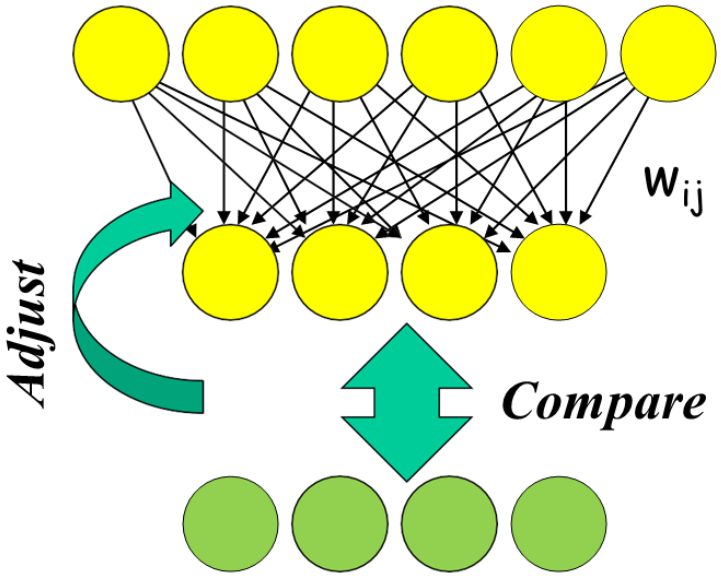

Figure 3. Schematization of the training procedure adopted for the training of neural networks. The output of the network is compared with desired outputs (green neurons), then the connection weights (w), on which the output depends are adjusted. This procedure is iterated to have outputs as close as possible to the desired ones.

Neural networks are trained with a procedure called gradient descent. This is an iterative optimization algorithm used to find the local minimum of the loss function, by moving in each step in the opposite direction of the gradient. Due to the complexity of the loss function, the algorithm cannot ensure to

find the optimal solution and usually requires a high number of iterations. Gradient descent is used in combination with a backpropagation algorithm [32], which is used to efficiently compute the gradient itself with respect to a loss function exploiting the connected architecture of the network (Figure 3).

An alternative popular approach of machine learning is Support Vector Machines (SVMs) that adopt a geometric approach (Figure 4). Indeed, SVMs define a separating hyperplane given a set of examples mapped in the feature space. The optimal hyperplane is obtained as the one that maximizes the margin *i.e.*, the highest distance among the separating hyperplane and the closest examples, called support vectors. Once the separating hyperplane is defined, new examples can be classified given the side (thus, the class) in which they fall.

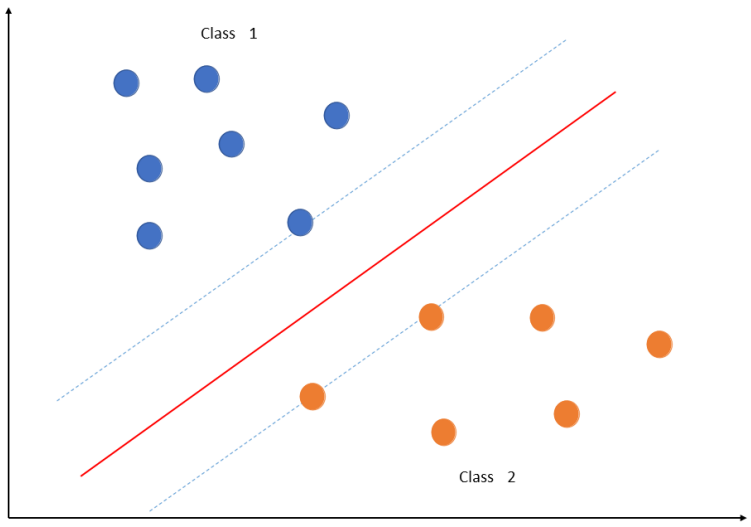

Figure 4. Schematization of an SVM. Blu circles are examples belonging to the class 1, orange circles belonging to the class 2. Red line represents the separating hyperplane, while blue lines indicate the margin. Circles from both classes lying on the margin lines are the support vectors.

SVMs can perform also nonlinear separation by means of the kernel techniques (Figure 5). Briefly, general functions are used to remap input data into a higher-dimensional space (feature space) where a better separation can be obtained. Thanks to the mathematical formulation of the SVM score function (dual Lagrangian), only the scalar product in the feature space must be known (kernel),

avoiding the explicit transformation of points in the feature space. Differently from neural networks, the training of SVMs is not iterative and ensures to reach the optimal solution.

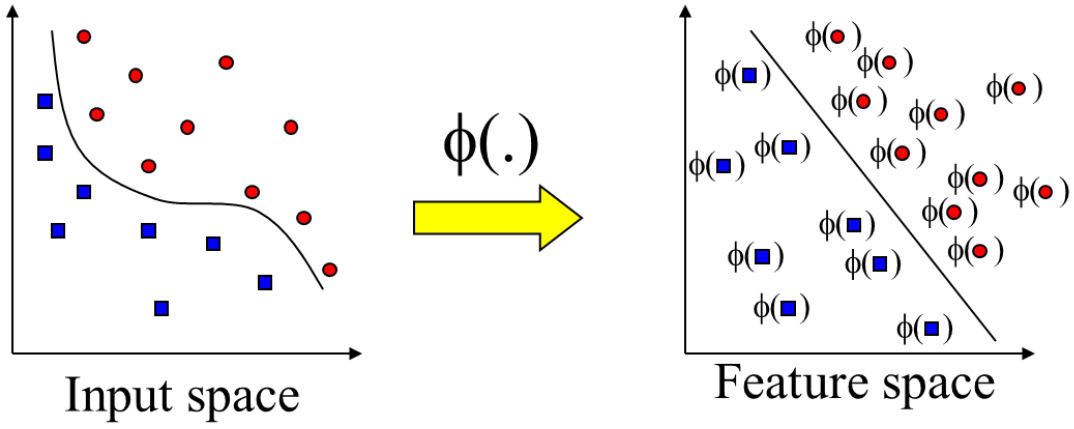

Figure 5. Kernel trick: examples non-linearly separable in the input space are mapped through the function Φ in a higher-dimensional, the so-called feature space, where they are linearly separable.

1.3.2. Probabilistic Graphical Models

Probabilistic graphical models are machine-learning methods for which conditional dependence between variables may be represented by means of a graph (Figure 6). These methods are further classified as generative or discriminative models. Generative models, such as hidden Markov models [33], estimate a joint probability distribution over inputs and outputs. Instead, discriminative models (*e.g.*, hidden conditional random fields [34]), directly model a conditional distribution, avoiding the computation of a marginal probability.

Hidden Markov models are Markov chains for which the observable is not the sequence of states (hidden path) but the probabilistic emission of characters. A Markov chain respects the so-called Markov rule *i.e.*, transition probabilities from a state to another depends only on the current state. Each state generates events (which are observable, contrary to states) with a given set of emission probabilities, specific for each state. The training of a hidden Markov model, for setting both transition and emission probabilities, is usually realized using the Baum-Welch algorithm. Once trained on a set of known data, for example proteins sharing the same family, HMM can recognize

other sequences belonging to the same family and can align them to the model states, allowing to relabel each position of the sequence with the state that most probably generate it given the trained parameters (Viterbi path).

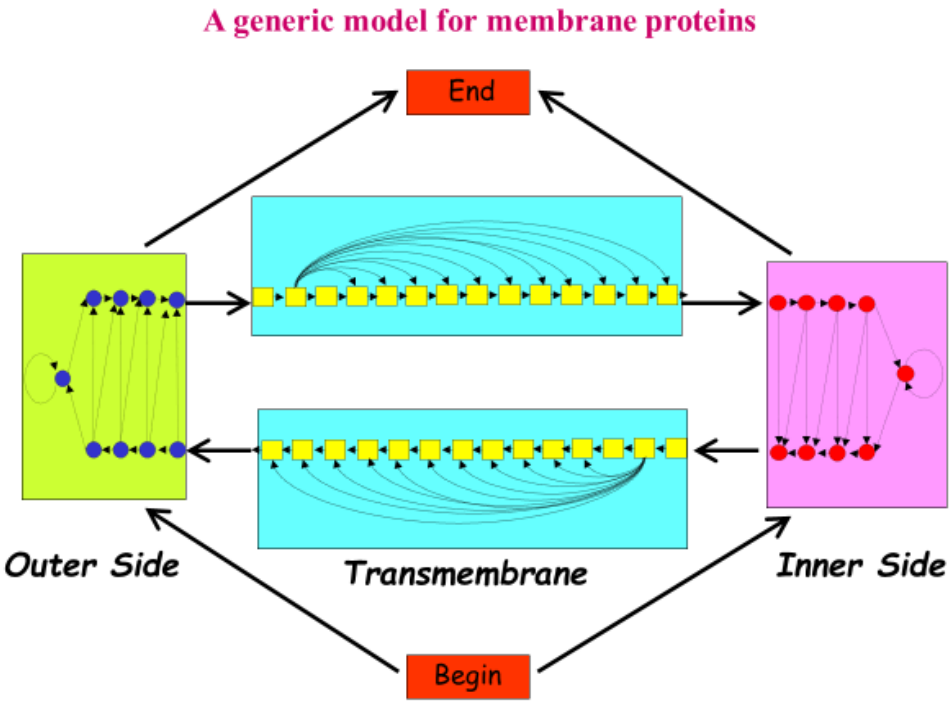

Figure 6. Example of a probabilistic graphical model designed for membrane proteins, in which states are represented as circles or squares, and arrows represent transitions between two states.

Hidden conditional random fields are similar to hidden Markov models, sharing the same basic architecture. Anyway, they do not compute a joint probability, which requires strict independence assumptions to be calculated, thus hampering the modeling of long-range interactions and/or multiple interacting features. At the contrary, hidden conditional random fields allow to make the computation of the transition probability depending also on previous states, rather than only on the current state (as stated by the Markov rule). Thus, they overcome one of the main limitations imposed by generative models and allow to model non-independent observations, that may overlap in space and time, which is the case in many applications, including bioinformatics ones.

1.3.3. Implementation and Validation of Machine-Learning Methods

The increasing computational power allowed made possible by advancement in computer technology allow to implement machine-learning models of ever-increasing complexity, opening new frontiers in the field of bioinformatics, as the advent of AlphaFold proves [35].

However, it must be considered that the success of a predictive method strongly depends on the careful choice of data for training and testing the method and on the rigorous application of validation procedures that prevent overfitting, in particular when the machine-learning method is complex in terms of training procedure and high number of trainable parameters.

The curation of the set of examples is crucial in the training phase [36]. A big effort must be placed on collecting a dataset that: i) contains only highly accurate data, possibly coming from experiments; ii) avoids biases towards clusters of similar data, that could unbalance the training procedure; iii) provides a large representation of the available examples. In a classification problem, examples of the positive and the negative class must be selected with the same criteria. A second aspect that deserves attention is the choice of the most informative representation of the input examples, the choice of the most relevant features for the problem at hand and of the most convenient feature encoding, ensuring not to introduce spurious correlations while preserving compactness.

Another aspect of the implementation of machine-learning methods that deserves the highest attention is the validation procedure to assess the generalization ability of the trained model [36]. To avoid overestimating the method performance, it is necessary to test it on a dataset of known examples as dissimilar as possible from the data used during the training phase. Besides the trainable parameters (*e.g.*, weights in neural networks, transition and emission probabilities in hidden Markov models), machine-learning models are usually defined by a set of hyperparameters (*e.g.*, network architecture, loss function, kernel, optimization procedure, learning rates and many others). The value of hyperparameters is not optimized by the training algorithm itself and the search of suitable

parametrization usually requires performing different training runs for exploring the hyperparameter space e.g., via a grid-search procedure. Again, to avoid overfitting, data used to search for optimal hyperparameters must be not used to evaluate the method.

Therefore, in general three sets of well annotated data are needed to train and test a machine-learning method: a training set for optimizing the trainable parameters by means of the training algorithm, a validation set to perform the choice of the best hyperparameters and a testing set, exclusively used to report the method performance.

A common strategy to adopt the three-set schema (training/validation/testing) while using all the available data for reporting performances is cross-validation. In cross validation, a dataset is split in *n* subsets. For each run, one subset is selected for validation, one for testing, and the remaining subsets (*n-2)* are used for training. The procedure is repeated *n* times rotating the choice of the subsets to use in each role (Figure 7).

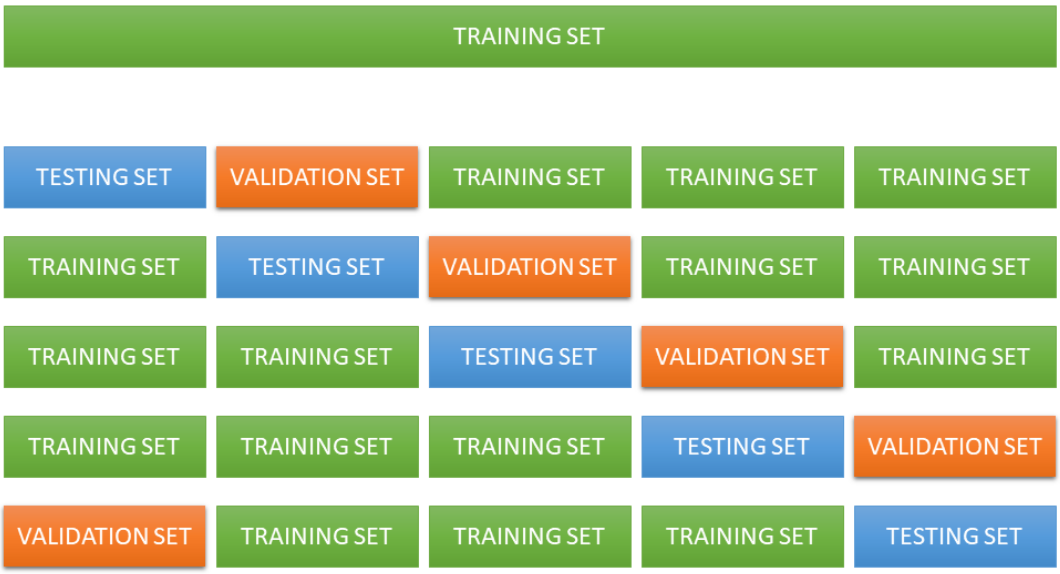

Figure 7. Cross-validation procedure: the training set is split in n subsets (in this case five). The procedure is repeated n times, so that each subset take turn being testing and validation set. The remaining subsets serve as training sets.

Another key issue is to evaluate the performance of methods by adopting convenient scores [36]. In particular, in classification problems is important to consider and balance errors in both positive and negative classes. The examples predicted in the positive class (P) can be correct (True Positive, TP) and wrong (False Positive, FP). Analogously examples predicted in the negative class (N) are either true negative (TN) or false negatives (FN). These numbers form the so-called confusion matrix and must be analyzed in their complexity. Pairs of complementary indexes such as Sensitivity (TP/(TP+FN), i.e., the ability to recognize examples in the positive class, and Precision (TP/(TP+FP), i.e., the probability of correct prediction, must be reported. If the classification depends on a variable threshold, the complementary index dependences can be plotter in a Receiver Operating Curve (ROC). Alongside these scores, measures that integrate all the information, such as the Matthews' correlation coefficient (MCC) or the F1 score, must also be computed and reported. The analysis of only partial aspects of the prediction might possibly lead to misinterpretation of the prediction performance.

Sequence labelling methods (such as those for annotating transmembrane segments on a sequence) require the evaluation of supplementary indexes that assess the prediction along the sequence, besides the efficiency in predicting single points. One of these indexes is the segment overlap score (SOV) that measures the superimposition between predicted and real segments.

1.3.4. Bioinformatics Methods for the Community

Predictive methods developed by bioinformaticians are routinely released so they can be used from researchers in life science to address practical problems. Several solutions are adopted: the release of the source code (routinely in public repositories stored on hosting and versioning services like GitHub or GitLab), the release of a containerized version of the package (a virtualization of the application and its dependencies that facilitate the deployment on different systems) developed by means of technologies like Docker (https://www.docker.com) or Singularity (https://apptainer.org), and/or the implementation of publicly accessible web servers. The last solution facilitates the access to the

resources to researcher lacking technical skills or computational resources required for installing the application. The systems developed in this thesis have been released through publicly available web servers whose implementation required the application of guidelines ensuring their security, maintainability, reproducibility, and usability.

Moreover, particular attention has been dedicated to ensuring their interoperability of applications, following the guidelines of ELIXIR, the European infrastructure for bioinformatics. To this aim, wherever possible we adopted standard ontologies, identifier resolvers (identifiers.org) [37], and schemas (Bioschemas) [38]. This allows to integrate the developed tools in an ecosystem of computational resources sharing standards, formats and ontologies greatly improving the usability and the FAIRness of tools (FAIR: Findable, Accessible, Interoperable, Reusable) [39]. Finally, to ensure accessibility and findability, the tools were inserted into Bio.tools (https://bio.tools), the official ELIXIR comprehensive repository for bioinformatics software and databases.

1.4. Prokaryotic Transmembrane Beta-Barrel Proteins Annotation

1.4.1. Biological Background and Motivations

TMBB proteins, or OMPs, are integral MPs localized in the outer membrane of Gram-negative bacteria, mitochondria, and chloroplasts. All TMBB proteins are formed by beta-strands spanning the membrane phase and share the same structural motif recalling a barrel [40]. Prokaryotic and eukaryotic ones differ on some peculiar structural characteristics. My research focused on prokaryotic TMBB proteins embedded in the outer membrane of Gram-negative bacteria. This exclude pore-forming toxins, which are secreted to be inserted in the host membrane and have peculiar structural characteristics, different from OMPs. An example of TMBB protein is reported in Figure 8.

All known prokaryotic TMBB proteins [41] have an even number of transmembrane segments per chain, ranging from 4 to 36. When the number of transmembrane segments is at least equal to 8, the

protein chain assumes a closed beta-sheet shape; when the number is 4, more chains interact to form a homomultimeric structure. In all cases, the N- and the C-terminal are localized on the inner (or periplasmatic) side of the outer membrane. All beta-strands along the chain interact in an antiparallel way with their closest neighbor strands, with the obvious exception of the first and last beta-strands, which are in mutual contact in the closed single-chain beta-sheet shape. Transmembrane segments are connected on the periplasmic side mainly by short turns, while longer loops are routinely observed on the extracellular side. Finally, transmembrane segments are characterized by the so-called dyad repeat pattern: alternating hydrophobic and hydrophilic residues, with the first ones facing the membrane and the others facing the interior of the barrel [42].

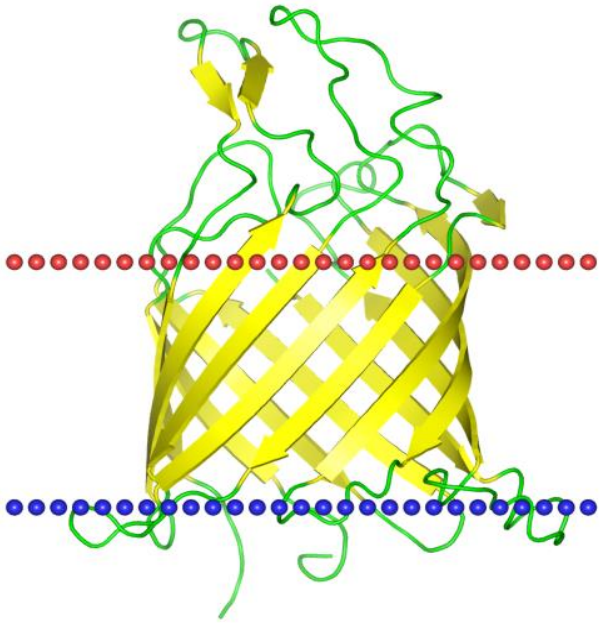

Figure 8. Transmembrane beta-barrel from Escherichia coli (OPM database, PDB ID: 1tly). Beta-strands correspond to yellow arrows, in green, loops and turns. Dotted line represent the periplasmic (blue) and extracellular (red) sides of the bacterial outer membrane.

A broad range of functions has been reported for TMBB proteins (see [42] for a review). Probably their most well-known functions are the general and specific diffusion of molecule and ions, carried out by a class of TMBB proteins known as porins. Anyway, their functions are far away from being limited to these. TMBB proteins act as membrane anchor and cell adhesion proteins, have peptidase

or lipase activity, are involved in signal transduction processes, act as efflux pumps and autotransporters.

TMBB proteins are an important part of the genome, being encoded by some 2-3% of all the genes in Gram-negative bacteria [42]. Nevertheless, TMBB proteins share the same fate as other MPs: even though they represent a large part of the proteome, perform various important functions and attract a great medical interest for the development of new drugs (*e.g.*, antibiotics and vaccines), they still lack an adequate number of resolved structures in PDB [15].

Given the above considerations, it is evident that the annotation of TMBB proteins via dedicated computational methods is a crucial theme. Approaching this problem, one may recognize two main computational tasks: first, the detection (or discrimination) of putative TMBB proteins in large datasets of protein sequences coming from high-throughput Omics studies (e.g., newly sequenced prokaryotic genomes); then, once a TMBB protein is recognized, perform the topology prediction *i.e.*, the identification of the number, the orientation, and the boundaries of transmembrane segments.

1.4.2. State of the Art

During the last twenty years, many computational methods for TMBB proteins annotation have been developed [43-50]. These methods may be divided in two main groups, according to the task they tackle: the first group collects methods devoted to TMBB proteins detection only, while the second one collects methods dealing with both the detection and topology prediction tasks.

In the first group, we list the statistical approach proposed in [43] and the homology-based tool named HHomp [44]. The former method assigns a score to the input protein given amino acid abundances observed in known TMBB structures and taking advantage of the dyad repeat pattern [43]. The latter tool is available via a web server. It recognizes OMPs by building a profile HMM for the input protein, then comparing it to a database of precomputed profile HMMs for families of TMBB proteins. When a hit is found, the protein is assigned to that family [44].

The second group comprises PROFtmb [45], PRED-TMBB [46], BOCTOPUS [47], BetAware [48], PRED-TMBB2 [49] and BOCTOPUS2 [50]. Most methods are based on Hidden Markov Models (HMMs), with the exception of BetAware, which is based on a combination of a neural network for TMBB proteins detection and Grammatical-Restrained Hidden Conditional Random Fields (GRHCRFs) [51] for topology prediction.

PRED-TMBB2 [49] and BOCTOPUS2 [50] are the two most recent methods among those cited above. The first one adopts an HMM divided in three sub-models representing the extracellular, periplasmic, and transmembrane regions. It implements several structural characteristics of prokaryotic TMBB proteins. BOCTOPUS2 is based on two steps. An SVM model discriminates among four per-residue classes: periplasmic, extracellular and, for the transmembrane region, pore-facing and lipid-facing. This per-residue profile is then used in the second step to predict the topology applying an HMM model. Methods developed so far, possibly owing to the reduced availability of training examples, show limited performance, in particular in topology annotation. In this work I apply for the first time in this field deep-learning procedures, testing their ability to extract valuable information even from small training sets.

1.4.3. BetAware-Deep

In the context of my research project, I have developed BetAware-Deep [4], a method designed to tackle both TMBB protein discrimination and topology prediction. BetAware-Deep is a two-step method combining a deep-learning method and a GRHCRF models, already adopted in the previous version of the method, and here extended. The whole methos has been trained on an updated training set counting 58 TMBB proteins with known structure. Moreover, the method introduces a novel formulation of the hydrophobic moment [52], used to model the dyad repeat pattern, which includes the evolutionary information extracted from a sequence profile.

BetAware-Deep was compared, on a novel independent testing set of 15 TMBB proteins, with other recent state-of-the-art methods approaching TMBB protein topology prediction and reported the highest results. In a second benchmark, assessing the performance of the methods in TMBB protein discrimination, BetAware-Deep reported results as high as others available methods. For this test, a large dataset already used to benchmark PRED-TMBB2 was used.

BetAware-Deep is available for the scientific community via an accessible web server with a user-friendly interface at https://busca.biocomp.unibo.it/betaware2.

1.5. Glycine Myristoylation Annotation in Proteins

1.5.1. Biological Background and Motivations

Myristoylation is a type of protein lipidation in which a myristoyl group is covalently attached to a protein residue. Myristoylation in eukaryotes is mostly associated to N-terminal glycines. The reaction is catalyzed by the enzymes N-myristoyltransferases (NMTs), which have been extensively characterized in eukaryotes [53]. Mammals have two NMTs (namely, NMT1 and NMT2) [54], while in lower eukaryotes and plants only one NMT is found. NMT1 and NMT2 have a sequence identity of about 76% [53] but show the same selectivity towards substrates *in vitro* [55].

Glycine myristoylation is mostly a co-translational protein modification, involving the N-terminal glycine exposed upon methionine excision. During this process, the protein chain is still in complex with the ribosome [53]. In fact, NMT presents a basic lysine cluster in its N-terminal region which has been proven to be crucial for the interaction with the ribosome [56]. Anyway, it has been proven that in metazoan glycine myristoylation also occurs post-translationally. In this case, the N-terminal glycine is exposed upon caspase cleavage, mainly during apoptosis [53], but also in other processes involving caspases, *i.e.*: cell differentiation, tumor suppression, neural development, and cell proliferation. Moreover, also NMTs are cleaved by caspases, determining a change in their

subcellular localization [56]. In fact, NMT1, which is mostly associated to membranes, is cleaved by caspase-3 or caspase-8 at Asp72 inducing a relocalization to the cytosol. At the same time, cytosolic NMT2 is cleaved at Asp25 by caspase-3, which relocalize to membranes [56].

Recent studies reported that N-terminal lysine residues may be myristoylated, as suggested by the crystallographic study in [57]. Moreover, myristoylation has been reported on N-terminal cysteine residues in some bacterial proteins, both localized in the inner or outer membrane [58]. Anyway, these studies are still preliminary and this type of myristoylations have very few examples to date.

According to crystallographic studies of NMTs in complex with their substrates [55, 59-62], the enzymes interact with the protein via the first eight residues, the so-called octapeptide (Figure 9). Only the first five residues enter the enzyme pocket, while the last three interact with its surface. Octapeptides are sufficient to have a myristoylation *in vitro* [55]. In a study in which the octapeptide was truncated to six and five residues, it was reported a decrease in the NMT activity and a complete loss, respectively [63].

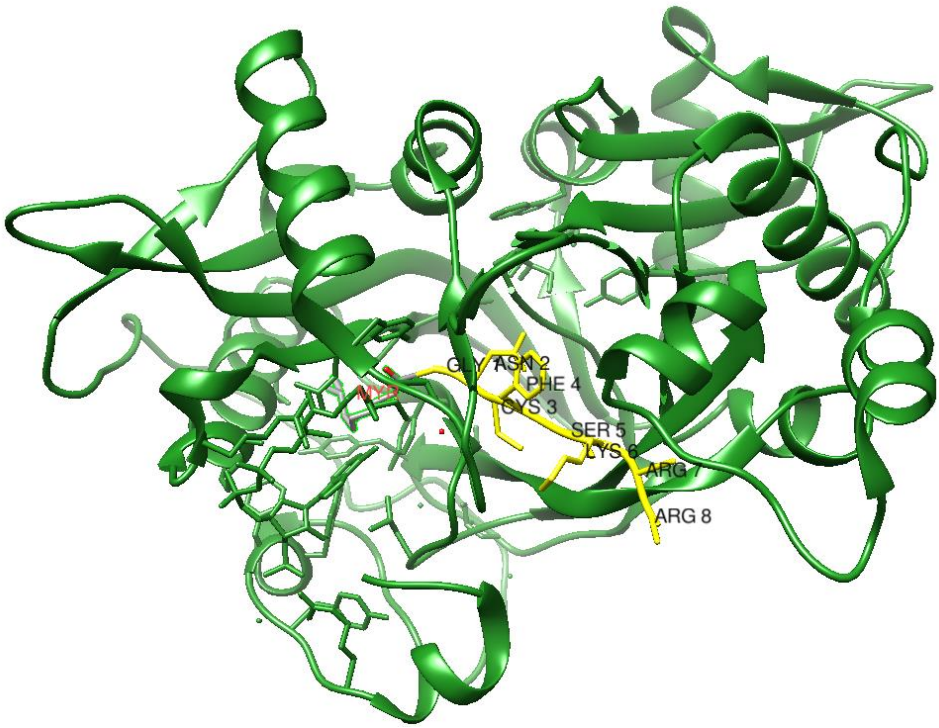

Figure 9. Octapeptide Myr-GNCFSKRR (yellow) in the pocket of the N-myristoyltransferase 1 from H. sapiens (green). PDB ID: 6qrm.

Myristoylated (MYR) proteins represent about the 2% of the genome in eukaryotic organisms [64]. They are principally associated to the plasma membrane and the organelles membrane and are involved mainly in signal transduction, apoptosis and pathological processes mediated by viruses and fungi [53]. Myristoylation of viral proteins is mostly catalyzed by host NMTs, but an open reading frame codifying for NMT was individuated in some viruses, even though not yet functionally characterized [58].

Myristoylation, besides the obvious relevance in organisms, have interesting application in drug design, since it is an effective mean to deliver peptides inside the cell [65]. Moreover, myristoylation may promote binding to albumin, then it is used to improve the stability and bioavailability of polypeptide drug [66].

Myristoylation may be detected thanks to various laboratory techniques. The classical approach relies on radioactive labeling [53]. More recently, bioorthogonal approaches, which allow to induce chemical reactions *in vivo* without interfering with biological processes occurring in the cell, where developed. Such experimental designs use azido or alkyne analogs of the myristic acid., allowing to metabolically incorporate them and to exploit their affinity to fluorophores, biotin, and other probes [67]. Besides these *in vivo* approaches, a high-throughput *in vitro* technique relying on macro-arrays was proposed. In this approach the myristoylation catalyzed by the NMT in presence of octapeptides is coupled to the formation of NADH, monitored by fluorescence [68].

1.5.2. State of the Art

Several computational approaches have been developed to tackle the problem of MYR protein detection. In PROSITE (https://prosite.expasy.org/) [69] it is reported a regular expression describing myristoylation sites (entry: PS00008): G[^DEFHKPRWY]XX[ACGNST][^P] (where X denotes any residues and ^ indicates the exclusion of the listed residues in square brackets). Thanks to information derived from crystallography studies in which the NMT is co-crystallized with its substrate, this

regular expression was refined to G[^DEFRWY]X[^DEKR][ACGST][KR] [56]. Further methods for MYR proteins prediction include NMT predictor [70], Myristoylator [71], and TermiNator3 [72]. NMT predictor is based on a scoring function summarizing the information derived from crystal structures and biochemical analysis of the first 17 residues of MYR proteins. Myristoylator is based on an ensemble of neural networks trained on the NMT predictor training set with the inclusion of a negative training set. Also in this case, the first 17 residues are considered. Conversely, TermiNator3 predicts the myristoylation status of a protein based on pattern scanning.

All the methods developed so far have very poor performance. Moreover, they are designed only for co-translational myristoylation prediction and no method for post-translational myristoylation prediction is available.

1.5.3. SVMyr

SVMyr [5] is an SVM-based method designed for co- and post-translational myristoylation prediction (the latter only in metazoan where it has been experimentally studied). It uses the information contained in octapeptides, which are sufficient for the NMT recognition and catalysis. The octapeptide is codified as a 12-positions array: the first seven positions are compositional per-residue scores (the starting glycine is fixed, then ignored at this point) derived from a Position Specific Scoring Matrix computed starting from octapeptides in the training set and a background distribution; the last five positions include mean physicochemical features for the octapeptide (hydrophobicity, charge, size, secondary structure propensities).

SVMyr searches for internal myristoylation sites (post-translational) via a pattern scanning for caspase cleavage sites exposing a glycine. Once a match is found, the resulting octapeptide is provided to the SVMs.

SVMyr was trained on 232 non-identical co-translationally MYR octapeptides experimentally annotated, as reported in SwissProt, the manually curated part of UniProtKB [73], and 232 non-

identical octapeptides for which it was proven that they do not undergo myristoylation *in vitro* [56]. In a benchmark performed over an independent testing set, having 88 myristoylated and 528 non-myristoylated proteins, SVMyr outperforms other methods. Moreover, it predicts correctly 11 out of 15 post-translational myristoylation sites experimentally validated reported in SwissProt and *in vivo* study [74].

SVMyr is made available via an accessible and usable web server with a user-friendly interface, which may be visited at https://busca.biocomp.unibo.it/lipipred/.

2. BetAware-Deep

2.1. Materials and Methods

BetAware-Deep [4] is a profile-based method for TMBB proteins detection and topology prediction (*i.e.*, given a protein sequence, identify the correct number and orientation of transmembrane segments). It consists of two cascading steps: a deep learning architecture (Bidirectional Long Short-Term Memory, BLSTM) [75] and a probabilistic method for sequence labelling, GRHCRFs [51]. BetAware-Deep also introduces a novel feature, a non-canonical formulation of the hydrophobic moment [49] designed to include evolutionary information in the computation of this measure and to effectively model the dyad repeat pattern observed in transmembrane segments. The main implementation characteristics of BetAware-Deep are summarized in the DOME (Data-Optimization-Model-Evaluation) checklist reported in Appendix 9. A full description of the adopted data and methods follows.

2.1.1. Datasets

The reference annotation of the topology of membrane proteins derives from structural data collected in PDB. Different secondary databases are available, colleting, and cataloguing membrane proteins of different classes. Among them structural data for TMBB are available in MPstruct (https://blanco.biomol.uci.edu/mpstruc/) and OPM [76].

Three datasets were used to train and benchmark BetAware-Deep in the topology prediction task, as summarized in Table 1: a Positive Training Set (PTS), a Negative Training Set (NTS) and a Blind Testing Set (BTS).

PTS and BTS were built starting from the 162 TMBB proteins reported and classified in MPstruc (https://blanco.biomol.uci.edu/mpstruc/), a database collecting MPs for which three-dimensional structures have been determined. From the initial dataset we removed pore-forming toxins since they

have a non-canonical topology, and they are not embedded in the bacterial outer membrane. We reduced redundancy in this initial dataset by clustering sequences with more than 25% of sequence identity at 90% of coverage using the blastclust tool and choosing as representative the longest sequence for each cluster. This procedure resulted in a dataset having 71 TMBB proteins. The dataset was split in PTS, a non-redundant training set counting 58 TMBB proteins, and BTS, a non-redundant blind test set having 13 TMBB proteins. Sequences included in BTS have been selected such that they have, at most, 25% of sequence identity at 50% of coverage among them and with respect to all proteins included in our training set and in those of other methods considered here for the comparative benchmark (BetAware [48], PRED-TMBB2 [49], and BOCTOPUS2 [50]). Other two proteins not reported in MPstruc but present in the OPM (Orientations of Proteins in Membrane) database [76] were added to BTS, since they fulfill redundancy criteria reported above. Finally, BTS counted 15 TMBB proteins.

The choice of the negative dataset for training (NTS) is an issue, due to the abundance of non TMBB proteins known at the structural level and the concomitant need to operate a selection to reduce the example to a number commensurable to the positive dataset. For this reason, we choose a dataset of proteins that possibly are the most difficult to be discriminated from the positive set: prokaryotic globular (non-membrane) proteins, annotated in the all-beta class in SCOPe [77]. This dataset comprises 69 proteins, and it was obtained selecting all the prokaryotic proteins included in this class, then reducing internal redundancy at 25% sequence identity threshold and 50% of coverage, and redundancy against PTS with the same criteria. The introduction of NTS is crucial to allow BetAware-Deep to discriminate between transmembrane and non-transmembrane beta-strands, which may be present in non-barrel regions present before and/or after the barrel itself.

Full length sequences from UniProtKB [73] were retrieved and used for all proteins in the datasets. By this, we consider the real-world case in which the transmembrane barrel region only represents a limited portion of the full sequence.

PTS and NTS were split in 10 cross-validation subsets. Proteins in PTS having 25% sequence identity at 50% of coverage among them were required to be in the same subset, in order to reduce redundancy among subsets.

To test the ability of BetAware-Deep in discriminating TMBB proteins from other protein families in large datasets, and to compare its performances with other methods designed for the same task, we used two datasets: the Positive Discrimination Testing Set (PDTS) and the Negative Discrimination Testing Set (NDTS). These two datasets were already used to test PRED-TMBB2. PDTS contains 1009 TMBB proteins, while NDTS, 7571 non-TMBB proteins (globular and alpha-helical inner MPs).

For details on proteins contained in PTS, NTS and BTS see Appendix 1-3.

Table 1. Datasets for BetAware-Deep training and benchmark.

Dataset	# Proteins	Source
PTS	58	MPstruc
NTS	69	SCOPe
BTS	15	MPstruc and OPM
NDTS	7571	[49]
PDTS	1009	[49]

PTS: positive training set. NTS: negative training set. BTS: blind testing set. NDTS: negative discrimination testing set. PDTS: positive discrimination testing set. The number of proteins contained, and the source database are reported for each dataset.

2.1.2. Topology annotation

The topology of a TMBB protein refers to its organization in the bacterial outer membrane. We can distinguish three distinct compartments: periplasmic, extracellular, and transmembrane region. The topology of prokaryotic TMBB proteins is characterized by: (i) even number of transmembrane segments per chain; (ii) N- and C-terminus in the periplasmic region; (iii) short turns connecting consecutive transmembrane segments on the periplasmic side and long loops on the extracellular side.

The collection of all the segments described above constitutes the barrel region. Before and/or after this region, a non-barrel region may be present, which may contain both alpha-helices or beta-strands.

One major issue is to establish clear rules for localizing transmembrane segments and annotating the localization of the loops with respect to the membrane plan and different approaches have been adopted so far. Basically, the problem arises from the fact that protein structures are obtained in a environment different from the membrane. Although it is routinely easy to recognize the membrane spanning segments of defined secondary structure, it is difficult to infer which are the residues that are in contact with the membrane. To this aim some computational method, like OPM, attempts to model the membrane as a plane strip of a determines thickness where experimental membrane proteins are localized by recognizing highly hydrophobic regions with length compatible with the membrane width. This procedure neglects all variability and dynamicity in the protein-membrane interaction and, although useful to capture the transmembrane segments, it can give very approximate knowledge on the residues interacting with the membrane, in particular in the borders.

Table 2. Labels adopted for BetAware-Deep topology annotation.

Label	Meaning	In Figure 10
n	non-transmembrane region	yellow
i	inner or periplasmic region	red
o	outer or extracellular region	red
T	residues in transmembrane beta-strand and embedded in membrane	light green
E	residues in transmembrane beta-strand and exposed to the solvent	dark green

Topology annotation is obtained merging annotations reported in OPM [76] and the secondary structure computed with DSSP [80] from the PDB file. OPM provides the exact localization of membrane-spanning segment, which is not directly derivable from the structure file deposited in PDB, by simulating the insertion of the protein in a membrane of variable thickness and optimizing the

protein transfer energy from water to the lipid bilayer. The transmembrane segment computed following this procedure often does not cover the entire beta-strand but, as said, boundaries of the membrane-contacting segment may be inaccurate. For these reasons in our approach the annotation was extended to the whole beta-strand, given the DSSP-derived secondary structure.

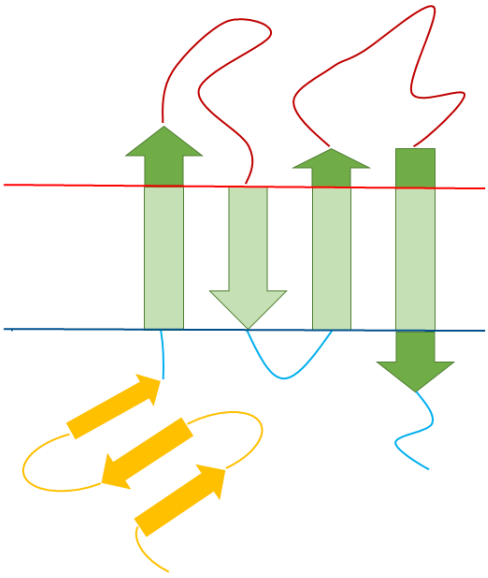

Figure 10. Graphical representation of a TMBB protein topology. Straight lines represent membrane boundaries: in blue, the periplasmic side, and in red, extracellular side. Beta-strands are represented by arrows. In yellow, non-barrel region. In dark red, extracellular loops and in light blue, periplasmic turns. Transmembrane beta-strands are light green if embedded in membrane and dark green in the exposed portion.

Given the information described above, the resulting per-residue annotation along the sequence has five possible labels, graphically depicted in Figure 10, and summarized in Table 2. To indicate the non-barrel region, we used the label *n*. Labels *i* and *o* indicates inner (periplasmic) and outer (extracellular) regions, respectively. Membrane spanning beta-strand, instead, have two possible labels: *T* for residues embedded in membrane and *E* for residues (still in extended conformation) exposed to the solvent.

2.1.3. Sequence Profile

BetAware-Deep exploits the evolutionary information in the form of a sequence profile. Firstly, a Multiple Sequence Alignment is computed aligning the query protein against UniRef90 [70]

(release 2018_03) using PSI-BLAST [79]. The program was run for two iterations with E-value threshold set to 10^{-3}. From the PSI-BLAST output, an MSA was derived stacking all the reported pairwise alignment and eliminating columns having a gap in the query sequence.

Given the MSA, a sequence profile was computed. It reports for each aligned position in the MSA the frequency in which each one of the twenty residues is observed. Then, it is represented by a matrix of *L x 20* dimensions, where *L* is the length of the query protein sequence.

2.1.4. Profile-Weighted Hydrophobic Moment

BetAware-Deep introduces, as a novel feature, the computation of a non-canonical formulation of the hydrophobic moment used to model the dyad repeat patterns observed in transmembrane segments. The residues spanning the membrane in as extended conformation, expose their side chain towards the external and the internal sides of the barrel with an alternate pattern. External side chains take contacts with the lipid phase (or with other transmembrane protein units), while the others face the internal part of the pore that is routinely in contact with the polar solvent.

The alternation of hydrophilic pore-facing residues with hydrophobic lipid-facing residues can be captured by the hydrophobic moment, which measures the amphiphilicity of a short protein segment, being higher when there is a separation between hydrophobic and hydrophilic residues given a specific angle separating sidechains along the backbone. The canonical formula adopted to compute this measure is:

$$\mu(\delta) = \{[\sum_{n=1}^{N} \mathrm{H}[R_n] \sin(\delta n)]^2 + [\sum_{n=1}^{N} \mathrm{H}[R_n] \cos(\delta n)]^2\}^{\frac{1}{2}} \quad (1)$$

were δ is the angle separating two consecutive sidechains and reflects the periodicity to be detected, being δ=100° for alpha-helices and δ=160° or δ=180° for beta-strands; *N* is the length of the window in which the hydrophobic moment is computed; $H[R_n]$ is the hydrophobicity of the residue *R* in position *n*.

We applied a simplified formula in which the window length was fixed to 5 and the angle δ to 180°. This resulted in:

$$\gamma = \left|\sum_{n=1}^{5} \mathrm{H}[R_n](-1)^n\right| \quad (2)$$

In this case the adopted hydrophobicity scale was the White&Wimley scale for the transfer of unfolded peptide chains into octanol [81]. The scale provides an experimental evaluation of the ΔΔG of transfer to an apolar phase of a residue, within a polypeptidic environment. It therefore estimates bilayer partitioning with bulk partitioning.

The hydrophobic moment described above is measured on the protein single sequence. Exploiting the information derived from the comparison of a protein with other members of its family, we introduced a new formulation, named Profile-Weighted Hydrophobic Moment (PWHM). A weighting scheme is applied based on the sequence profile derived from the MSA of the query protein, according to the formula:

$$\gamma = \left|\sum_{n=1}^{5} \sum_{R\in\{A,C,D,\ldots,Y\}} P[R_n]\ \mathrm{H}[R_n](-1)^n\right| \quad (3)$$

where the inner summation takes into consideration all the twenty residues *R* and $P[R_n]$ is the frequency reported in the sequence profile for the residue *R* in position *n.*

Finally, the assigned PWHM for each position is the maximum value reported in the 3-residue window centered on that position. This is done because it is reasonable to think that the hydrophobic moment centered on residues near to the middle of transmembrane segments may be higher with respect to that computed for residues in the edges. This happens because all residues in the window are embedded in membrane and they tend to have hydrophobic residues on the same side (facing the membrane), that is a favorable condition. This possibly results in a higher hydrophobic moment. In the edges, instead, residues exposed to the solvent (which can escape the dyad repeat pattern, since they are not constrained by the membrane) are included in the window. Given that, reporting the

maximum hydrophobic moment in the 3-residue window may amplify the signal provided by this measure.

2.1.5. Workflow

BetAware-Deep combines two predictive steps. The first one is a BLSTM model which takes as input the profile joined with the PWHM. Each position along the sequence is then represented by a 21-dimensional array. BLSTM outputs Per-Residue Probabilities (PRPs), the probability for each residue of being localized in each one of the five compartments defined in the topology annotation phase (Table 2). PRPs are joined to the profile, resulting in a 25-dimensional array. This constitutes the input for the GRHCRF model representing the second step, which provides the topology prediction.

TMBB proteins discrimination is based on the topology prediction: if at least 4 transmembrane segments are predicted, the protein is classified as TMBB protein. This decision has biological basis since all the prokaryotic TMBB proteins observed so far have at least 4 membrane-spanning segments per chain.

2.1.5.1. First step: BLSTM

As a first step in its workflow BetAware-Deep adopts a BLSTM model. It represents an advancement over Long Short-Term Memory (LSTM) [30] models, which allows higher performances when applicable [75].

LSTM are a deep learning method belonging to the class of Recurrent Neural Networks (RNNs), which are devised for the processing of sequential data. In fact, RNNs introduces the concept of memory, which allows to store information contained in previous inputs to generate the next output of the sequence. This is realized and governed by means of a feedback loop. Among all possible RNN schema, the most powerful are gated RNNs, in which connections have weights that may change at each time step. This is done to handle the vanishing gradient issue, the major problem encountered trying to learn long-term dependencies [82,83].

LSTM models are a type of gated RNNs. They introduce for the first time an internal recurrence, that is added to the outer recurrence of RNNs and have a variable weight. In fact, a LSTM cell (Figure 11) have a state cell, that is a regular artificial neuron with a self-loop representing the internal recurrence, and three gating units controlling: the accumulation of the input in the state cell itself (input gate); the self-loop weight (forget gate); the output, which can be eventually shut off (output gate). Thanks to the peculiar gating schema here described, LSTM models can store information over an arbitrary time, delete it once it is already used, and neglect non-relevant positions.

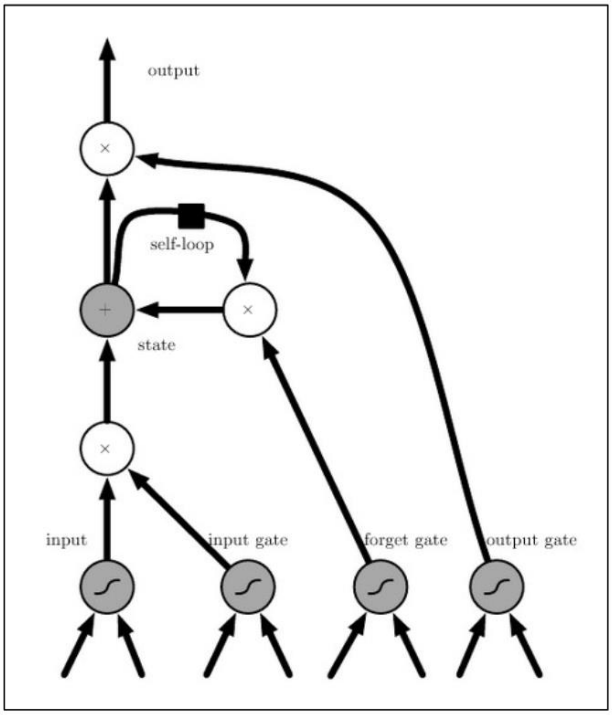

Figure 11. Schematic representation of a LSTM cell. Weights are represented by white circles. Grey circles represent the gate units controlling the weights. The state cell is the grey circle with the self-loop.

A BLSTM model, belonging to the class of bidirectional RNNs [84], consists of two LSTM layers of which one is provided with the sequence and the second one with its reverse copy. The output of the two layers is then merged to obtain the final prediction. This architecture allows the method to include past and future information in the context of each time step.

The BLSTM architecture used by BetAware-Deep is represented in Figure 12. It consists of two recursive LSTM layers whose output is combined. Each one of the two LSTM layers includes 128

cells. A masking layer is applied upstream to reduce the effect of zero-padding, used to have sequences all of the same length in the training set, equal to the maximal length observed. In fact, the application of the masking layer allows to ignore position undergoing padding. The output of the BLSTM is passed to a dropout layer with rate fixed to 0.3, which is used to prevents overfitting. Then, a dense fully connected layer is applied to obtain PRPs from the output provided by the recursive layer.

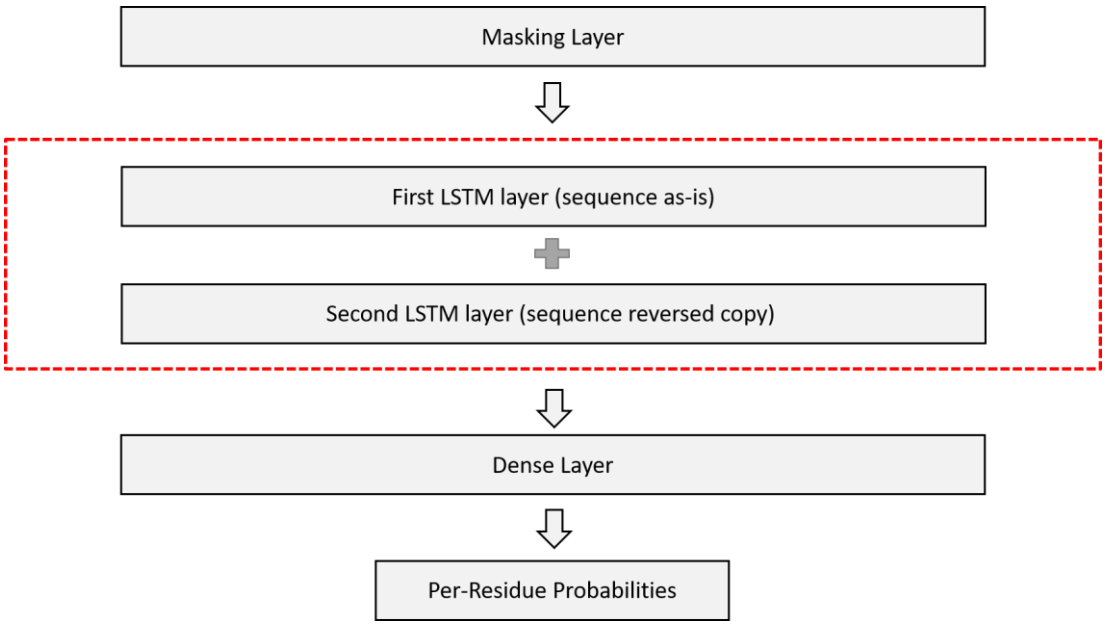

Figure 12. BLSTM architecture adopted by BetAware-Deep. At first, a masking layer to ignore zero-padded positions is applied. Then, the input is passed to the BLSTM (red square), having to LSTM layers scanning the sequence left-to-right and right to left respectively. The two outputs are combined and passed to the dense layer, which produces PRPs.

The training procedure was carried out via gradient descent on the categorical cross-entropy loss function and applying the Adam optimization algorithm [85]. The early stopping technique was used to obtain the best BLSTM model monitoring the validation loss and terminating the training after 20 epochs without any decrease, then the best model was restored. The model has been implemented using the Keras Python library [86].

2.1.5.2. Second step: GRHCRF

The second step adopted by BetAware-Deep is a GRHCRF model, implemented for the first time in the first version of the method, BetAware [49].

Conditional Random Fields (CRFs) [87] are discriminative probabilistic models widely adopted in sequence labeling. In contrast with generative probabilistic models (such as HMMs), which estimates a joint probability, they work modeling a conditional distribution. The definition of a joint probability requires the enumeration of all possible observation sequences. This requirement hampers the modeling of multiple interacting features and long-term dependencies. Moreover, strong independence assumptions are needed to allow these computations. At the contrary, the conditional probability incorporates non-independent attributes of the observations, representing single features or collection of features. Then, the transition probabilities do not depend only on the current observation but also on neighboring observations.

CRFs offers several advantages over generative probabilistic models. Anyway, they lack hidden-states variables, which results in the inability to capture intermediate structures. This limitation is overcome by the introduction of Hidden CRFs (HCRFs) [34], which uses intermediate hidden-state variables.

GRHCRFs add to HCRF a regular grammar defined over a set of constraints known for the problem at hand. The introduction of this grammar ensures that BetAware-Deep outputs only biologically consistent results. In the prediction phase, in fact, GRHCRFs identify the most probable path given the model and the input sequence using a Posterior-Viterbi dynamic-programming algorithm.

GRHCRFs may be represented as a finite-state automaton (Figure 13). The model adopted by BetAware-Deep is based on the 40-states model used by the first version of the method. This model may be divided in three principal sub-models: periplasmic, transmembrane, and extracellular regions. The start and end state are localized in the periplasmic region. To the original GRHCRF model, we added two states with a self-loop (in the periplasmic and in the extracellular side) to model the non-transmembrane region, which may be localized before and/or after the barrel region. Furthermore, we added a label for non-embedded residues in transmembrane beta-strands, which is associated to self-looped states at the two edges of transmembrane segments.

As done for the BLSTM model, the training procedure was carried out using a validation set to find the number of iterations giving the best GRHCRF model.

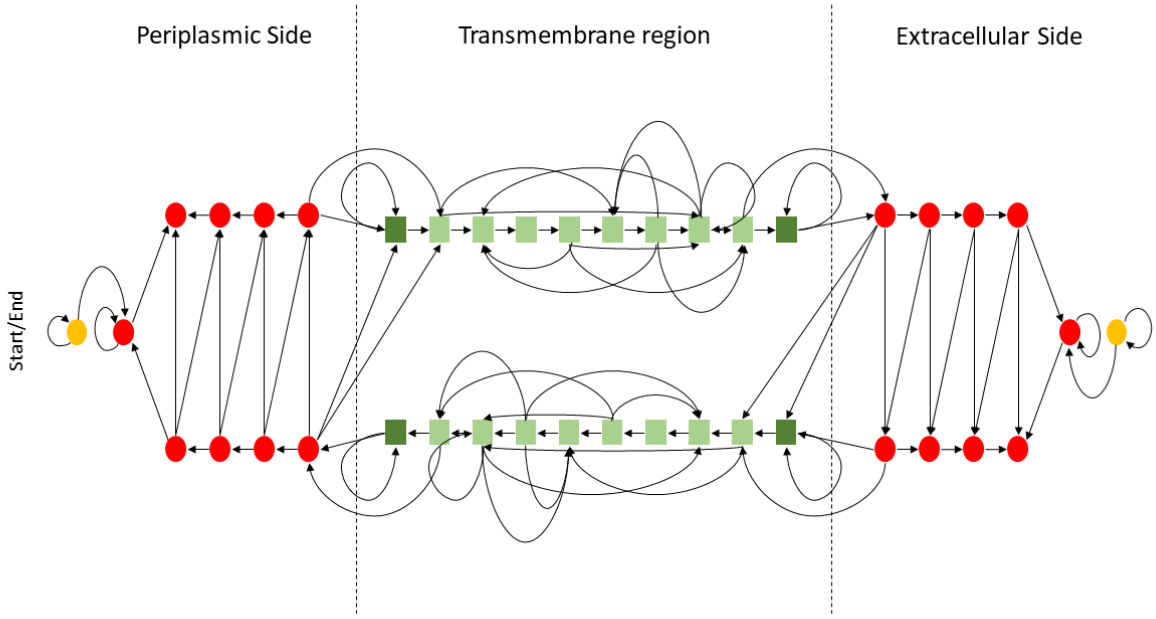

Figure 13. Architecture of the GRHCRF model adopted by BetAware-Deep. Yellow circles model the non-transmembrane region. Red circles model the periplasmic (i) and extracellular (o) region. Transmembrane region is modeled by green squares: in light green, residues embedded in membrane (T), in dark green, non-embedded residues (E).

2.1.6. Evaluation

BetAware-Deep was benchmarked in both the TMBB proteins topology prediction and discrimination tasks.

For topology prediction, the output of BetAware-Deep is reduced to a three-state schema: the five labels are reduced to three by: (i) considering *n*, which indicates non-transmembrane regions, as *i* (periplasmic)*;* (ii) replacing *T* and *E*, the two possible labels for transmembrane segments with *T*, regardless of the actual insertion in the membrane. Following the same logic, the output of BOCTOPUS2, which uses two labels for transmembrane residues, differentiating between pore-

facing (*p*) and lipid-facing (*L)* residues, is reduced considering these two labels as *T*. For PRED-TMBB2 and BetAware, no replacements were needed.

The scoring indexes adopted to evaluate the methods performances include the three-state accuracy (Q_3), Segment Overlap (SOV) [88], Protein Overlap (POV) and the portion of proteins with correct number of predicted transmembrane segment (N_{TM}).

Q_3 is computed as:

$$Q_3 = \frac{\Sigma_i p_i}{N} \quad (4)$$

where p_i is the number of correct positive per-residue predictions for the class *i* and *N* represents the total number of residues.

SOV is computed for the class *T* and defined as:

$$SOV(T) = \frac{1}{N(T)} \sum_{S(T)} \left[\frac{minov(S_1,S_2) + \delta(S_1,S_2)}{maxov(S_1,S_2)} \times len(S_1) \right] \quad (5)$$

where the normalization value *N(T)* is the total length of the observed transmembrane segments; S_1 and S_2 as observed and predicted transmembrane segments, respectively; *minov(S_1,S_2)* is the length of the intersection of the segment pair for the class *T*; *maxov(S_1,S_2)* is the length of the union of the segment pair for the class *T*; *len(S_1)* is the length of the observed segment.

δ(S_1,S_2) in the above definition of SOV is computed as:

$$\delta(S_1,S_2) = min\left\{ maxov(S_1,S_2) - minov(S_1,S_2);\ minov(S_1,S_2); int\left(\frac{len(S_1)}{2}\right);\ int\left(\frac{len(S_2)}{2}\right)\right\} \quad (6)$$

POV is defined as:

$$POV(s) = \begin{cases} 1 & if\ (N_P^S = N_O^S \quad and \quad P_i \cap O_i \geq \theta \quad \forall i \in [1, N_O^S] \quad) \\ 0 & otherwise \end{cases} \quad (7)$$

where N_P^S and N_O^S are the number of predicted and observed transmembrane segments, respectively; P_i and O_i are the *i*-th predicted and observed segments, respectively; θ is equal to the average between the half-lengths of segments P_i and O_i.

The scoring indexes adopted for the evaluation of TMBB proteins detection are sensitivity, specificity, and Matthews Correlation Coefficient (MCC).

Sensitivity is defined as:

$$Sen = \frac{TP}{TP+FN} \quad (8)$$

Specificity is defined as:

$$Spe = \frac{TN}{TN+FP} \quad (9)$$

MCC is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (10)$$

where TP (True Positive) and TN (True Negative) are the numbers of correctly predicted positive and negative proteins, respectively, and FP (False Positive) and FN (False Negative) are the numbers of incorrect positive and negative predictions, respectively.

2.2. Results and Discussion

BetAware-Deep [4] is a web server designed for TMBB proteins detection and topology prediction. The method behind combines a deep learning (BLSTM) and a probabilistic (GRHCRF) method. Moreover, it introduces the computation of a non-canonical formulation of the hydrophobic moment, indicated as PWHM.

BetAware-Deep has been trained on a dataset including 58 TMBB proteins and 69 non-TMBB proteins. To test the method in cross-validation, the training set was split in 10 non-redundant subsets. For each run, eight subsets were used for training, one for validation and one for testing.

For the topology prediction task, the method was benchmarked on a blind test set including 15 TMBB proteins. For the detection task, we took advantage of a large dataset already used to test Pred-TMBB2, which includes 1009 TMBB and 7571 non-TMBB proteins [49].

2.2.1. Hydrophobic Moments

The hydrophobic moment is a measure of the alternance of hydrophobic and hydrophilic residues along a short protein segment. Canonically, it is computed given a single protein sequence. From now on, this version of the hydrophobic moment will be referenced as Unweighted Hydrophobic Moment (UHM).

In BetAware-Deep, the hydrophobic moment is used to model the dyad-repeat pattern observed in the transmembrane segments of TMBB proteins. Instead of the UHM, the formulation adopted by BetAware-Deep includes profile weights in the computation. This allows the inclusion of the evolutionary information. Therefore, this measure is called PWHM.

We tested the ability of the hydrophobic moment only in discriminating transmembrane beta-strands (T or E) from loops and other domains (n, i, and o). We therefore compared the the discriminative power of UHM and PWHM with a ROC curve (Figure 14). By changing the threshold on the computed hydrophobic moment, the ROC curve plots the rate of true positives as a function of the rate of false positives. The area under the curve (AUC) estimates the overall discriminative power, being 0.5 the AUC obtained by a random classifier.

UHM and PWHM report AUC values equal to 0.6 and 0.74, respectively. These results show that our formulation enhances the discriminative power and better captures the signal given by the dyad-repeat

pattern, resulting in a more accurate detection of transmembrane residues. Thus, the inclusion of evolutionary information, in the form of a sequence profile, is beneficial for this task.

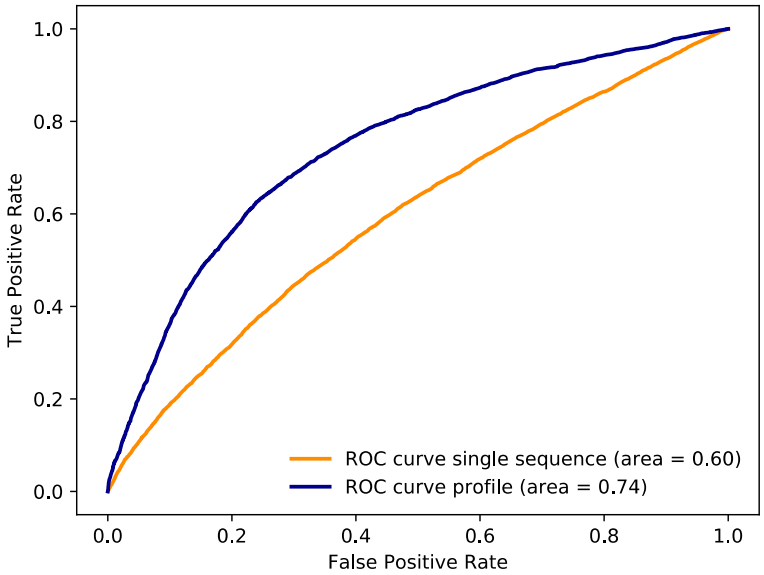

Figure 14. ROC curve describing the ability of unweighted (UHM, orange line) and profile-weighted (PWHM, blue line) hydrophobic moments in discriminating between transmembrane and non-transmembrane residues.

2.2.2. Topology Prediction: Cross-Validation

BetAware-Deep was tested in a 10-fold cross-validation to compare different input encodings for the BLSTM model and assess the best one. In particular, we considered three possible models: (i) a baseline model, which includes only the sequence profile; (ii) a model combining the sequence profile with the UHM; (iii) the combination of the sequence profile and the PWHM. Results are reported in Table 3.

In the cross-validation procedure, both models incorporating the hydrophobic moment outperform the baseline method. Specifically, the model adopting just the profile reported 35 out of 58 correct topologies (POV) and 39 out of 58 proteins with correct number of predicted transmembrane residues (N_{TM}). The inclusion of UHM led to an increase of this metrics to 37 and 40, respectively. Anyway, the highest results were reported by the last method, which have a POV equal to 40 and a N_{TM} equal to 46. Moreover, it reported the highest accuracy (88%) and the highest SOV (95%).

Table 3. Comparison of different BLSTM input encodings obtained with a cross-validation procedure over the positive training set (58 TMBB proteins).

BLSTM input encoding	Q3	SOV	POV	N_{TM}
PROFILE	83%	91%	35	39
PROFILE + UHM	81%	92%	37	40
PROFILE + PWHM	88%	95%	40	46

All inputs encodings are profile-based. The second and third ones also include a hydrophobic moment. UHM: unweighted hydrophobic moment. PWHM: profile-weighted hydrophobic moment. Q3: three-class accuracy. SOV: segment overlap. POV: number of correctly predicted topologies. N_{TM}: proteins with correct number of predicted transmembrane segments.

These results confirm the benefit given by the inclusion of the hydrophobic moment and the predominance of our PWHM over the UHM, as already suggested by the ROC curve in Figure 5. Given these observations, the input encoding combining the profile and the PWHM was selected to be implemented in BetAware-Deep.

2.2.3. Topology Prediction: Blind Test

BetAware-Deep was benchmarked on an independent testing set counting 15 TMBB proteins for the topology prediction task. Indeed, this dataset was designed to allow an unbiased comparison among BetAware-Deep and other state-of-the-art methods for topology prediction, namely: BetAware (first version) [48], Pred-TMBB2 [49], and BOCTOPUS2 [50].

Table 4 reports the results obtained in this comparative analysis. According to these observations, the improvement with respect to the previous version of the method is substantial: BetAware-Deep reports 10 out of 15 correct topologies (POV) and proteins with correct number of predicted transmembrane segments, while BetAware scores 4 and 5 out of 15, respectively. Moreover, between the two, the new version reported the highest accuracy (80% *vs.* 60%) and the highest SOV (94% *vs.* 55%).

Table 4. Comparison among methods for TMBB proteins topology prediction performed over a blind test set (15 TMBB proteins).

Method	Q3	SOV	POV	N_{TM}
BetAware-Deep	80%	94%	10	10
BOCTOPUS2	65%	68%	8	8
Pred-TMBB2	71%	80%	6	11
BetAware	60%	55%	4	5

Q3: three-class accuracy. SOV: segment overlap. POV: number of correctly predicted topologies. N_{TM}: proteins with correct number of predicted transmembrane segments.

Compared with the two recent methods, Pred-TMBB2 and BOCTOPUS2, our method results as the top-performing one for TMBB proteins topology prediction. Indeed, even though Pred-TMBB2 reported a N_{TM} of 11, it has a POV of 6, that is way lower than the one reported by BetAware-Deep. At the same time, it outperforms also BOCTUPUS2, which has both POV and N_{TM} equal to 8. Moreover, BetAware-Deep has the highest accuracy and the highest SOV among all methods. In fact, Pred-TMBB2 reported 71% and 80%, respectively, and BOCTOPUS2 reported 65% and 68%, respectively.

Even though the reduced number of available examples for benchmark analysis limits the comparison among methods, the results we reported highlight that BetAware-Deep at least well-compares with other recent tools for TMBB proteins topology prediction.

2.2.4. Detection of TMBB proteins

Besides TMBB proteins topology prediction, BetAware-Deep is designed also for the detection (or discrimination) of such protein families in large datasets. The same task is also performed by BetAware, Pred-TMBB2, and BOCTOPUS2. All these methods were considered in our comparative analysis. Moreover, HHomp, a method devoted just to the discrimination task, was included. This

benchmark (Table 5) was performed over the 1009 TMBB and the 7571 non-TMBB proteins already used to test Pred-TMBB2 and other available methods in [49].

Table 5. Comparison among methods for TMBB proteins detection performed over a dataset counting 8580 proteins (1009 of which are TMBB proteins) derived from [48].

Method	Sen	Spec	MCC
BetAware-Deep	98.12%	97.53%	0.91
BOCTOPUS2	98.12%	98.81%	0.93
Pred-TMBB2	91.87%	99.14%	0.92
BetAware	67.29%	99.87%	0.8
HHomp	97.73%	99.95%	0.98

Sen: sensitivity, portion of correctly predicted positive examples. Spec: specificity, portion of correctly predicted negative examples; MCC: Matthew's Correlation Coefficient. Results of BetAware-Deep are obtained in this work, while those of the other methods were taken from [49].

In the discriminative benchmark, BetAware-Deep reports high performances, having an MCC of 0.91, sensitivity of 98.12% and specificity of 97.53%. These results are at the level of other state-of-the-art tools. The top-performing method is HHomp. Anyway, it should be noticed that it adopts an approach based on a database of precomputed profile HMMs of putative TMBB proteins. Hence, for an input sequence, this method builds a profile HMM to be compared with those included in the database. This approach is completely different from the machine learning-based approach adopted by the other methods, and it presents a limitation, since it is able to detect only TMBB proteins belonging to previously discovered protein families.

2.2.5. Web Server

BetAware-Deep is made available through an accessible web server provided with a user-friendly interface (https://busca.biocomp.unibo.it/betaware2). In the home page, the user is invited to either paste a sequence in FASTA format or upload an external FASTA file. In both cases, the server accepts in input only a sequence per job.

Once the sequence is submitted, the user is redirected to the page where the results will appear. In Figure 15 the BetAware-Deep results page is shown. These results are obtained with the input protein Adhesin YadA from *Yersinia enterocolitica* (UniProt: A1JUB7).

Figure 15. BetAware-Deep results page. (a) Summary of submitted job, including information about the input and prediction results, i.e.: predicted localization, overall TMBB probability, number and mean length of transmembrane segments. (b) information about transmembrane segments: begin and end position, length, and average of predicted probability scores. (c) interactive feature viewer reporting detailed topology annotations.

In the output page, BetAware-Deep reports information organized in three sections. In the first section (panel in the top left, (a) in Figure 15), it reports general information about the submitted job, *i.e.*: the unique **job ID**, which is internally assigned by BetAware-Deep, the **submission time**, **protein ID** and **protein length**, as extracted from the input FASTA sequence. Moreover, this section reports the **predicted localization**, that is **Outer Membrane TMBB** if BetAware-Deep predicted at least 4 transmembrane segments or **Other: non-TMBB**, otherwise. Other information present are the **overall TMBB probability** (computed as the average probability assigned to predicted membrane-spanning residues by the GRHCRF model), the **number of TM segments** and the **mean TM strand length**.

In the second section (panel on the top right, (b) in Figure 15), it is shown the list of transmembrane segments, each one endowed with the **begin** and **end** position, its **length**, and the **mean score** relative to the label T (transmembrane).

In the last section (panel on the bottom, (c) in Figure 15) there is an interactive feature viewer allowing the user to analyze the whole sequence. In particular, it reports the primary sequence along with two annotation tracks. The first one is the **topology prediction track**, which show the alternance of periplasmic, transmembrane, and extracellular segments. The second one is the **TM probability track**, representing graphically the per-residue transmembrane probabilities. The feature viewer allows to zoom in a particular area of interest. It is also possible to automatically zoom on a specific predicted transmembrane segment by selecting it in the summary table.

Results may be downloaded in a JSON file, storing the complete job results, or in a CSV file, reporting residue level annotation of topology, with associated per-residue probabilities.

The web server has been implemented using the Python Django we framework (https://www.djangoproject.com). For the backend database (storing information about submitted jobs and results) we adopted the PostgreSQL (https://www.postgresql.org) database management system. The web interface has been developed using HTML5, JavaScript and JQuery. In particular, for the web page layout we used the Bootstrap4 toolkit (https://getbootstrap.com/). Tabular data were rendered using the DataTables JQuery plugin (https://datatables.net). For visualizing protein sequences and annotated feature tracks we used the FeatureViewer JavaScript library [89].

3. SVMyr

3.1. Materials and Methods

SVMyr [5] is an SVM-based method designed for the discrimination of co- and post-translationally myristoylated proteins in proteomes. It makes predictions based on octapeptides having a glycine (on which the myristic acid is attached) in the starting position. As a unique feature, SVMyr searches for internal myristoylation sites exposed upon caspase cleavage, implementing a pattern scanning along the protein full sequence. Octapeptides are encoded by means of per-residue scores computed for the seven variable positions and mean physicochemical features (hydrophobicity, charge, size, secondary structure propensities). The main implementation characteristics of SVMyr are summarized in the DOME (Data-Optimization-Model-Evaluation) checklist reported in Appendix 10. A full description of the adopted data and methods follows.

3.1.1. Datasets

A major challenge in this domain is the collection of reliable positive and negative datasets. For the positive class, only 272 proteins are annotated as myristoylated in SwissProt with experimental evidence and can be considered highly reliable. Besides that, different proteome-wide experiments conducted in vivo with techniques based on fluorescence or metabolic labelling provided the myristoylomes of some parasites: *Trypanosoma brucei* [90], *Trypanosoma cruzi* [91], *Leishmania donovani* [92] and *Plasmodium falciparum* [93]. More recently, a study based on protein macroarray assay tested 2048 N-terminal, Gly-starting octapeptides extracted from human and *Arabidopsis thaliana* proteomes, as translated from the corresponding genome sequences, identifying 834 putatively myristoylated proteins. The last experiment also provides a dataset of 1214 octapeptides (1126 of which were mapped to UniProt) that putatively does not undergo myristoylation. Unfortunately, the collection of a reliable negative dataset is quite challenging. To this aim we added to the negative dataset [54], a set of proteins experimentally proven to undergo modifications on the

starting glycine that are incompatible with myristoylation. In particular we collected 64 proteins for which an annotation for the acetylation of the N-terminal glycine s present in SwissProt.

To train and benchmark SVMyr in the co-translational myristoylation prediction task four dataset were used (Table 6): a Positive Training Set (PTS), a Negative Training Set (NTS), a Positive Blind Testing Set (PBTS) and a Negative Blind Testing Set (NBTS).

PTS was built starting from the 272 co-translationally myristoylated proteins with experimental annotation reported in SwissProt [73]. From this initial dataset, we extracted the N-terminal octapeptides and clustered the identical ones, retaining one representative for each cluster. This procedure resulted in a collection of 232 non-identical octapeptides from 37 organisms. The majority of the proteins came from human (133 proteins, 57%), then Arabidopsis (31, 13%) and viruses, which account for 21 species and 25 proteins (11%). Other species include yeasts (12 proteins, 5%), mouse (7, 3%), rat (6, 3%), bovine (4, 2%), and 9 other species (15, 6%).

NTS includes 232 octapeptides for which it was demonstrated that they do not undergo myristoylation *in vitro* in presence of the enzymes NMTs [56], even though they have a glycine in starting position. This study, in fact, provided 1126 non-myristoylated octapeptides, from which we randomly selected our 232 octapeptides, to obtain a balanced training set having positive and negative examples in equal number.

PTS and NTS were split in 10 cross-validation subsets. In both cases, similar octapeptides were required to be in the same subset, adopting Hamming Distance (HD) as measure of similarity. HD is defined as the number of different positions in two strings with the same length. Given that the starting glycine is fixed, the maximal HD is equal to seven. Then, we required octapeptides having HD lower than four to be in the same cross-validation subset, in order to reduce redundancy among them.

PBTS was built starting from the 834 positive examples reported in this work. These examples were from *Arabidopsis thaliana* (483 octapeptides) and *Homo sapiens* (351 octapeptides). Only high and

medium confidence hits (classified as such based on the catalytic efficiency reported in the study) were retained. To these examples, we added the myristoylated proteins identified in proteome-wide experiments conducted on parasites. From this initial dataset, we reduced internal redundancy by clustering octapeptides with HD lower than four and choosing a representative for each cluster. Then, with the same threshold, we reduced redundancy towards the training set of SVMyr and the other methods considered for benchmark, *i.e.*: NMT predictor [70], Myristoylator [71], and TermiNator3 [72]. This resulted in a dataset counting 88 myristoylated proteins.

NBTS was derived from the remaining part of negative *in vitro* examples after the selection of NTS, which includes 232 octapeptides, and the 64 acetylated proteins reported in SwissProt. After having reduced internal redundancy, and redundancy towards the negative training sets of considered methods adopting one, NBTS included 528 non-myristoylated octapeptides, of which 25 were acetylated proteins.

Table 6. Datasets for SVMyr training and benchmark.

Dataset	# Proteins	Source
PTS	232	Swiss-Prot
NTS	232	[55]
PBTS	88	[55,90-93]
NBTS	528	[55], SwissProt
PTBTS	4 11	SwissProt [74]

PTS: positive training set. NTS: negative training set. PBTS: positive blind testing set. NBTS: negative blind testing set. PTBTS: post-translational blind testing set. The number of proteins and the source database are reported for each dataset.

To test SVMyr performances in the post-translational myristoylation detection task, we constructed a further testing set, Post-Translational Blind Testing Set (PTBTS). This dataset included four proteins with a post-translational myristoylation site experimentally annotated in SwissProt and 11 proteins reported in a *in vivo* study in which apoptosis was induced and the internal myristoylation

site was identified [74] and not yet included in UniProt [73]. Then, PTBTS counts 15 examples in total.

For details on protein included in all these datasets, see Appendices 4-8.

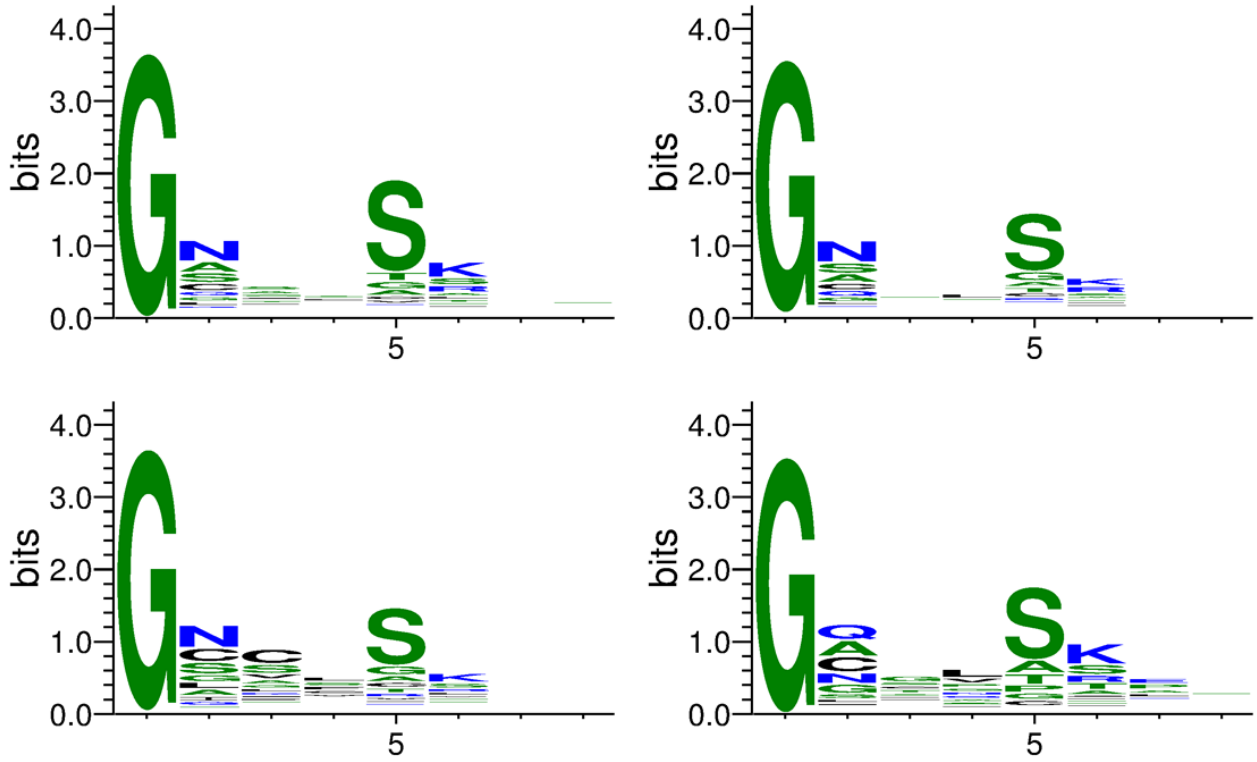

Figure 16. Octapeptides logos. In clockwise order: training octapeptides (232 examples from 37 organisms); Human octapeptides (351, from [55]); octapeptides from parasites (124, from 4 species [90-93]); Arabidopsis octapeptides (483, from [55]). Logos are generated using WebLogo 3.7.4.

A suitable way to graphically represent sequence profiles is the adoption of sequence logos that estimate, position by position, the information conveyed by each residue. Figure 16 shows logos built aligning the octapeptides in PTS (top left corner). It appears that some conserved positions emerge: i) in position 2, asparagine, alanine, serine, and cysteine are the most represented residues; ii) in position 5, serine is highly conserved, followed by threonine, glycine, and alanine; iii) position 6 shows a preference towards positively charged residues (lysine, arginine). These characteristics, which are confirmed in literature (as reviewed also in [54]), may be observed also in octapeptides from human (top right corner) and Arabidopsis (bottom left corner) reported *in vitro* [55]. Notably, a

preference towards cysteines in position 2 and 3 emerges in *Arabidopsis* . This suggests the presence of dual-lipidations since cysteine is known to be palmitoylated [53]. In the octapeptides from the 4 unicellular parasites for which a proteome-wide study was available [90-93], positions 5 and 6 share the same characteristics observed in the other datasets, while, in position 2, glutamine is the most observed residue (rather than asparagine). As for Arabidopsis, this position shows a high presence of cysteines.

Moreover, since SVMyr is designed also for large-scale analysis, we selected eight complete reference proteomes, downloaded from UniProt. These included: *Arabidopsis thaliana, Homo sapiens, Mus musculus, Saccharomyces cerevisiae, Trypanosoma brucei, Trypanosoma cruzi, Leishmania donovani,* and *Plasmodium falciparum*.

3.1.2. Workflow

SVMyr [5] is the first method addressing both co- and post-translational myristoylation.

Firstly, SVMyr extracts the N-terminal octapeptide, where co-translational myristoylation may take place. Then, it searches for putative caspase cleavage sites exposing a glycine along the sequence, via pattern scanning. Once such sites are found, the downstream octapeptides are extracted, since they may undergo post-translational myristoylation.

SVMyr encodes the input octapeptide through a scoring function computed from a novel Position Specific Scoring Matrix (PSSM) and its physicochemical characteristics. PSSM is used to derive seven scores, one for each position of the octapeptide but the glycine in starting position. Physicochemical features (including hydrophobicity, charge, size, and propensity towards alpha-helices or beta-strands formation) are computed as the mean over the octapeptide, providing five additional scores. Then, the input encoding results in a 12-dimensional vector.

The prediction step is implemented by means of an ensemble of ten SVM models, each one trained on a different cross-validation subset. The myristoylation probability is computed as the average

among the probabilities produced by all the SVM models. An input octapeptide is predicted as myristoylated if such probability is at least 50%.

3.1.2.1. PSSM

The PSSM compiled for SVMyr implementation reports in each position a log-odd ratio computed starting from two frequency distributions. The first one is computed stacking all the octapeptides in the training set without gaps, then computing a profile. Therefore, this procedure ends up with a matrix reporting the frequency of the 20 residues in each position of the alignment. The second one is a background distribution computed collecting from SwissProt all the N-terminal octapeptides of eukaryotic proteins with a glycine in starting position, after methionine excision. This resulted in 14,304 non-identical octapeptides, piled-up stating from the glycine residue. As for the first distribution, a profile is computed.

For each position, but the starting glycine, the PSSM value is computed as:

$$PSSM_{R,i} = -\log\frac{f_{R,i}}{b_{R,i}} \quad (11)$$

where $f_{R,i}$ is the frequency observed in myristoylated proteins for the residue R in position i and $b_{R,i}$ is the same frequency observed in the background distribution.

A PSSM was computed for each cross-validation training set, considering only its positive part.

3.1.2.2. Physicochemical features

Together with the scores provided by the PSSM, SVMyr adopts in its input encoding also mean physicochemical features. These include: hydrophobicity, as reported in the Kyte-Doolittle scale [94]; charge, considering its value equal to +1 in presence of arginine or lysine, and -1 for aspartate or glutamate; size, as reported in AAindex [95] (https://www.genome.jp/aaindex); propensity towards secondary structure, both alpha-helix and beta-strand [96].

3.1.2.3. SVM

Due to the paucity of the training set and to the relatively simple feature encoding, we did not apply in this problem deep learning procedures that usually require more data and/or complex encoding to extract generalizable information. Therefore, SVMyr adopts a prediction step based on an ensemble of linear SVMs.

The hyperparameters of the SVM models were optimized adopting a 10-fold cross-validation and a grid search. For each run of cross-validation we used eight subsets for training, one for validation (used to determine the optimal regularization parameter *C*) and one for testing.

To implement SVMyr, we used the Python package scikit-learn (https://scikit-learn.org).

3.1.2.4. Post-translational Myristoylation Prediction

SVMyr address the post-translational myristoylation task by scanning the protein sequence to find caspase cleavage site motifs localized upstream a glycine. In fact, the caspase proteolytic cleavage is necessary to expose the internal myristoylation sites, given the experimental evidence collected so far. Then, this modification takes place during apoptosis and other caspase-mediated processes in metazoan, where caspases are found.

Motifs used for pattern scanning are derived from the Eukaryotic Linear Motifs (ELM) database [97] and summarized in Table 7. In this database four apoptotic caspase cleavage motifs are reported, namely: one validated motif for caspase 3/7 (ELME000321) and three candidate motifs for caspase 2, 6 and 9.

Once a caspase cleavage site is found by this procedure, downstream octapeptides with a glycine in starting position are predicted for myristoylation with the ensemble SVM procedure.

Table 7. Apoptotic caspase cleavage site motifs adopted by SVMyr for the prediction of post-translational myristoylation sites.

Caspase	Motif status	Pattern
Caspase 2	In validation in ELM	[DEIL]X[DEFY]D
Caspase 3-7	Fully-annotated in ELM (ELME000321)	[DSTE][^P][^DEWHFYC]D
Caspase 6	In validation in ELM	[VLIT][EDQ][^DENQRKAPGS]D
Caspase 9	In validation in ELM	[^RK][EDQ]HD

In pattern, ^ indicates the exclusion of residues in square brackets and X indicates any residue.

3.1.3. Evaluation

Metrics used to benchmark SVMyr include sensitivity (8), precision, MCC (10) and F1-score.

Precision is defined as:

$$Pre = \frac{TP}{TP+FP} \quad (12)$$

F1-score is defined as:

$$F1 - score = \frac{2 \times Pre \times Sen}{Pre+Sen} \quad (13)$$

where TP (True Positive) and TN (True Negative) are the numbers of correctly predicted positive and negative proteins, respectively, and FP (False Positive) and FN (False Negative) are the numbers of incorrect positive and negative predictions, respectively.

Moreover, we computed Receiver Operating Characteristic (ROC) curve and the relative Area Under the Curve (AUC), when applicable.

3.2. Results and Discussion

SVMyr [5] is a web server designed for co- and post-translational myristoylation prediction in proteins. The method behind is based on an ensemble of SVM taking as input octapeptides with a glycine in starting position. They are encoded via compositional scores (computed using a PSSM derived in this work) and physicochemical features. Moreover, it scans the sequence searching for putative caspase cleavage sites localized upstream a glycine. This allows the detection of post-translational myristoylation sites, a unique feature of our tool.

SVMyr has been trained on a dataset having 232 co-translationally myristoylated octapeptides with experimental validation in SwissProt and 232 non-myristoylated octapeptides tested *in vitro* [55]. The training set was divided in 10 non-redundant subsets for cross-validation. Moreover, SVMyr was benchmarked on a testing set having 88 high/medium confidence myristoylated octapeptides and 528 non-myristoylated octapeptides.

To test SVMyr in the post-translational myristoylation task, we used a dataset with 4 examples experimentally annotated in SwissProt and 11 examples derived from an *in vivo* study [74].

Finally, SVMyr was used in a proteome-wide analysis involving *Arabidopsis thaliana, Homo sapiens, Mus musculus, Saccharomyces cerevisiae, Trypanosoma brucei, Trypanosoma cruzi, Leishmania donovani,* and *Plasmodium falciparum*.

3.2.1. Co-translational Myristoylation

SVMyr was tested with a 10-fold cross-validation over the training set (Table 8). In this testing procedure, it reported a sensitivity of 65%, precision of 87%, MCC of 0.61 and F1-score of 75%.

In addition, SVMyr was benchmarked adopting the blind test set with other available methods, namely: NMT predictor [70], Myristoylator [71], and TermiNator3 [72]. Moreover, the regular expression proposed in [55] (Regular Motif A) and the PROSITE pattern (Regular Motif B) [69] were

tested. Results are reported in Table 8. In this benchmark, SVMyr reported the highest results among the methods in terms of precision (62%), MCC (0.58) and F1-score (64%). Only TermiNator3 reported a higher sensitivity (81% *vs.* 67%), but this is compensated by a much lower precision (24%), with a MCC and F1-score of 0.27 and 37%, respectively. Similar considerations may be done for Regular Motif B, which have a sensitivity of 72%, but low precision (36%). At the contrary, Regular Motif A had the highest precision (69%), but poor sensitivity (20%). Overall, these results demonstrated that SVMyr outperforms the other method in the co-translational myristoylation prediction task.

Table 8. SVMyr results obtained in cross-validation over the training set and results obtained by all methods over the blind test set.

Method	Dataset	Sen (%)	Pre (%)	MCC	F1 (%)
SVMyr	Cross-validation	65	87	0.61	75
SVMyr	Blind test set	67	62	0.58	64
NMT	Blind test set	44	60	0.46	52
Myristoylator	Blind test set	48	40	0.33	43
TermiNator3	Blind test set	81	24	0.27	37
Regular Motif A	Blind test set	20	69	0.33	32
Regular Motif B	Blind test set	72	36	0.39	48

Regular Motif A: regular expression proposed in [55]. Regular Motif B: regular expression reported in PROSITE [69]. Sen: sensitivity. Pre: precision. MCC: Matthews Correlation Coefficient. F1: F1-score.

These two methods are the only ones providing a score both for positive and negative predictions, allowing the computation of true and false positive rates at different thresholds. NMT provides a score only for the positive class, the other methods perform a binary classification. Thus, their performances are represented as single points in the graph (Figure 17). SVMyr reported the biggest AUC (0.91), while Myristoylator had an AUC of 0.79. Moreover, also the scores reported by the other methods and regular expressions are clearly lower than the one reported by SVMyr.

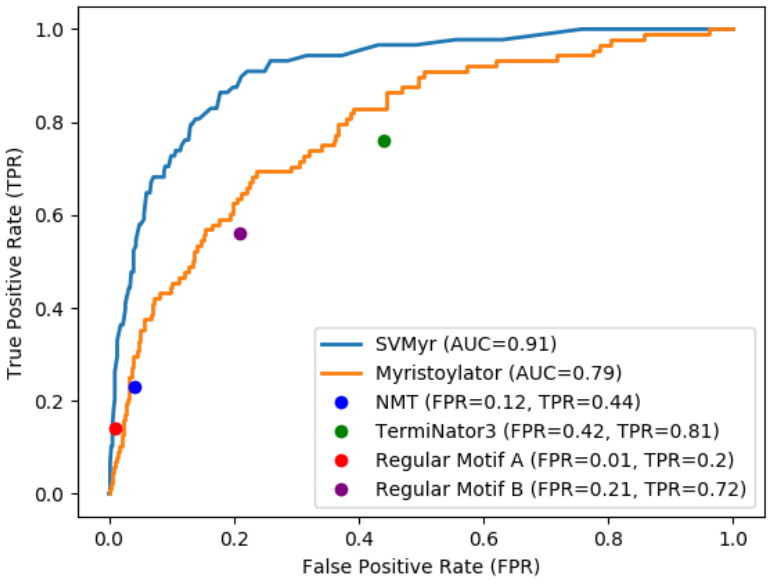

Figure 17. ROC curves computed for SVMyr and Myristoylator. For the other methods it was not possible to compute TPR and FPR at different score thresholds, then they are represented by points.

3.2.2. Post-translational Myristoylation

Among the available methods, SVMyr is the only one addressing the problem of post-translational myristoylation detection. Thus, we tested it on a dataset designed for this task. SVMyr correctly predicted all the four proteins annotated as post-translationally myristoylated in SwissProt, and 7 out of 11 proteins from an *in vivo* study (not included in SwissProt). Overall, 11 out of 15 proteins were correctly classified, showing that SVMyr well performs also in this task.

3.2.3. Proteome Analysis

SVMyr is a fast method designed for large-scale analysis. Therefore, we used it to filter the complete reference proteomes of eight organisms: *Arabidopsis thaliana, Homo sapiens, Mus musculus, Saccharomyces cerevisiae, Trypanosoma brucei, Trypanosoma cruzi, Leishmania donovani,* and *Plasmodium falciparum*. Results of this analysis are reported in Table 9 (for metazoan, *i.e.*, *H. sapiens* and *M. musculus*) and Table 10 (for non-metazoan organisms).

Table 9. Proteome-wide analysis performed with SVMyr on two selected metazoan organisms.

Proteins	***H.sapiens*** ***(#)***	***M.musculus*** ***(#)***
in the Proteome	79038	55341
in the Glyome*	5243	3788
Annotated	401	177
experimentally ^	*373*	*16*
not in training #	*240*	*9*
automatically °	*28*	*161*
predicted CT, annotated	254	152
experimentally ^	*238*	*14*
not in training #	*130*	*9*
automatically °	*16*	*138*
predicted CT	902	719
predicted CT, new targets §	183	272
octapeptides	*158*	*223*
predicted PT	1422	1147
sites	*1487*	*1231*

CT: co-translational. PT: post-translational. *For each proteome, the Glyome size indicates the number of proteins starting with Gly (or MetGly). ^: the number of experimental annotations in UniProt with ECO:0000269 and/or in the reference papers, when present, as quoted among square brackets in the header line. #: the number of proteins not included in the training set of SVMyr.°: the number of proteins with non-experimental annotation for myristoylation in UniProt. §: the number of predicted new MYR protein substrates. We excluded protein isoforms of genes endowed with an isoform previously annotated as myristoylated, either experimentally or computationally.

For co-translational myristoylation, SVMyr predicted 902 proteins in *H. sapiens*, 615 in *A. thaliana*, 719 in *M. musculus*, 39 in *S. cerevisiae*, 119 in *T. brucei*, 194 in *T. cruzi*, 119 in *L. donovani*, and 61 in *P. falciparum*. For all these organisms the portion of co-translational myristoylated proteins over the entire proteome ranges from 1 to 2%, as already observed [61]. In *S. cerevisiae*, SVMyr predicted 39 co-translationally myristoylated proteins, about 0.66% of the proteome, which is lower than 2% as reported in previous analysis [61]. For all these proteomes, SVMyr correctly identifies most (or even all) the experimentally annotated co-translationally myristoylated proteins. In fact, the overall sensitivity is 74% (902 predicted proteins over 1227 annotated ones). Considering only experimental annotations the overall sensitivity is 72% (732 over 1014).

Table 10. Proteome-wide analysis performed with SVMyr on six selected organisms.

Proteins	***A.thaliana* (#)**	***S.cerevisiae* (#)**	***T.brucei* (#)**	***T.cruzi* (#)**	***L.donovani* [+](#)**	***P.falciparum* (#)**
in the Proteome	39334	6050	8587	19242	7960	5384
in the Glyome*	3457	288	463	981	412	232
annotated	506	18	62	16	30	17
experimentally ^	*488*	*12*	*62*	*16*	*30*	*17*
not in training [#]	*457*	*1*	*62*	*16*	*30*	*17*
automatically °	*18*	*6*	*0*	*0*	*0*	*0*
predicted CT, annotated	376	18	54	13	20	15
experimentally ^	*365*	*12*	*54*	*13*	*20*	*15*
not in training [#]	*334*	*1*	*54*	*13*	*20*	*15*
automatically °	*11*	*6*	*0*	*0*	*0*	*0*
predicted CT	615	39	119	194	119	61
predicted CT, new targets [§]	68	21	63	181	99	44
octapeptides	*63*	*21*	*63*	*123*	*97*	*44*

CT: co-translational. PT: post-translational. *For each proteome, the Glyome size indicates the number of proteins starting with Gly (or MetGly). ^: the number of experimental annotations in UniProt with ECO:0000269 and/or in the reference papers, when present, as quoted among square brackets in the header line. [#]: the number of proteins not included in the training set of SVMyr.°: the number of proteins with non-experimental annotation for myristoylation in UniProt. [§]: the number of predicted new MYR protein substrates. We excluded protein isoforms of genes endowed with an isoform previously annotated as myristoylated, either experimentally or computationally.

The percentage of correct predictions in human is 64% (238 out of 373 proteins), in Arabidopsis is 75% (365 out of 488), in mouse is 87% (14 out of 16), and 100% in yeast (12 out of 12). For *T. brucei, T. cruzi, L. donovani,* and *P. falciparum* proteome-wide studies identified a pool of myristoylated proteins [90-93]. These experimental annotations could be used to validate our findings. In *T. brucei*, SVMyr correctly classified 54 out of 62 (87%) experimental examples [90]. In *T. cruzi*, it found 13 out of 16 proteins (81%) [91]. In *L. donovani,* 20 out of 30 proteins (67%) [92]. Finally, in *P. falciparum*, SVMyr identified 15 out of 17 proteins (88%) [93].

These values are slightly lower when considering proteins not included in the training set, with an overall sensitivity of 69% (576 over 832), varying from the 54% (130 out of 240) observed in *H. sapiens* to the 100% observed in mouse and yeast (9 and 1 examples, respectively).

For post-translational myristoylation, due to lack of available experimental data, it was not possible to validate the predictions of SVMyr, which identifies several putative proteins and sites of this type in organism in which this process takes place (metazoan).

It should be noticed that SVMyr is a fast method, which can process the entire human proteome in about ten minutes (nearly a protein in 0.1 seconds on average).

Co-translational myristoylated proteins predicted by SVMyr in the selected proteomes are covered mostly by PFAM domains observed in the training set, including Pkinase, Arf, EF-hand_7, PK_Tyr_Ser-Thr and G-alpha. Figure 18 reports the number of PFAM domains per proteins in the training set and in the new substrates identified by SVMyr in the proteome analysis. In the monodomain proteins included in the training set, the most represented PFAM domains are: Arf (PF00025, 12 proteins), G-alpha (PF00503, 11 proteins), Pkinase (PF00069, 10 proteins), EF-hand_7 (PF13499, 7 proteins), PK_Tyr_Ser-Thr (PF07714, 7 proteins). Remarkably, almost the same domains are found in the new substrates: Pkinase (27 proteins), EF-hand_7 (17 proteins), Arf (14 proteins) , PK_Tyr_Ser-Thr (10 proteins). These protein domains are reported also in literature as domains commonly found in myristoylated proteins [98].

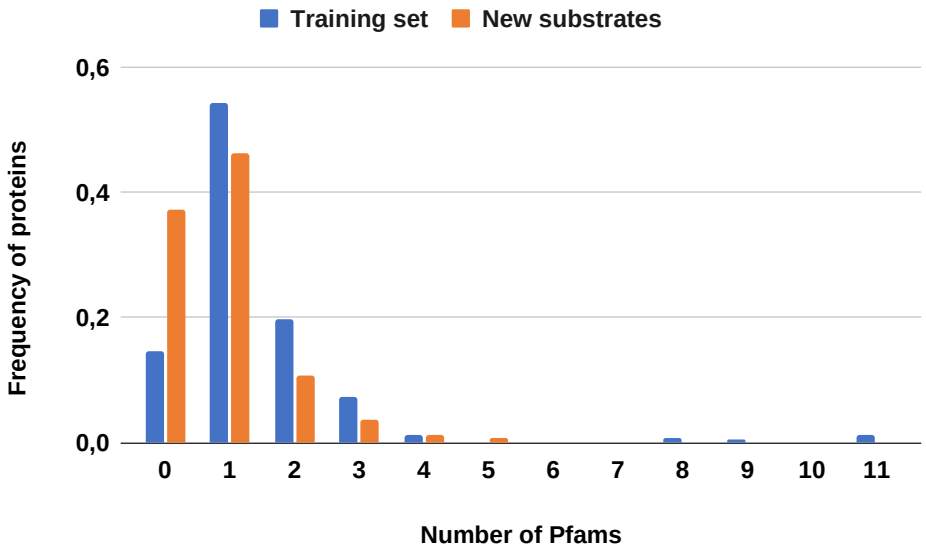

Figure 18. Number of Pfam domains per protein in the positive training set (blue) and the set of MYR substrates predicted in the 8 different proteomes.

3.2.4. Web Server

SVMyr is made available for the scientific community via a web server designed having in mind usability and accessibility (https://busca.biocomp.unibo.it/lipipred).

In the server home page, the user is invited to either paste sequences in FASTA format or upload a FASTA file. SVMyr accepts multiple submissions up to 1,000 proteins per job. For each protein, SVMyr performs both co- and post-translational myristoylation prediction. Upon submission, the user is redirected to a waiting page, then to the output page, in which results are organized in three tables (Figure 18).

The first table reports information about the job, including: **job ID**, **submission time**, and the **number of submitted sequences**.

The second table reports the **co-translational myristoylation prediction** of the input proteins. For each protein it details the **protein accession/ID**, the **prediction** (that is equal to **N-myristoyl glycine** or **not myristoylated**), the **position** in which myristoylation takes place, the **octapeptide** used in the prediction phase, the **probabilistic score,** and the associated classification, *i.e.*: **highly probable** (score ≥ 0.8), **probable** ($0.5 \leq$ score ≤ 0.8) or **improbable** (score ≤ 0.5). The third table lists the post-translational myristoylation sites predicted in the input proteins. This table have one additional column with respect to the second table, reporting the **caspase type** for which a match is found in the pattern scanning procedure. Moreover, the column reporting the octapeptide is modified to include also the sequence containing the caspase cleavage site.

For each entry having a positive prediction in the third table, it is available a detail result page accessible by clicking on the protein accession/ID. In this page, via an interactive feature viewer, the caspase cleavage and the myristoylation site are shown along the primary sequence. Three supplementary tables report detailed information about: (i) the **caspase cleavage sites** found along the sequence, together with the **caspase type** involved, the matching **caspase motif**, **begin** and **end**

positions, the **cleavage site** in the sequence; (ii) the **probable/highly probable myristoylation sites**; (iii) the **improbable myristoylation sites.**

Summary of submitted job

Job ID:	4889e33e-9f52-46a4-967f-4aea79546045
Submission time:	Nov. 24, 2021, 5:02 p.m. (GMT)
Number of submitted sequences:	4

Co-translational myristoylation prediction

Protein Accession/ID	Prediction	Position	Octapeptide	Score	Note
Q06389	N-myristoyl glycine (co-translational)	2	GAKTSKLS	0.9	Highly probable
Q99828	N-myristoyl glycine (co-translational)	2	GGSGSRLS	0.87	Highly probable
P42858	Not myristoylated (co-translational)	-	-	-	No N-terminal Glycine
Q06002	Not myristoylated (co-translational)	-	-	-	No N-terminal Glycine

Post-translational myristoylation prediction

Protein Accession/ID	Prediction	Position	Caspase cleavage | Octapeptide	Score	Caspase type	Note
P42858	N-myristoyl glycine (post-translational)	551	DLND | GTQASSPI	0.9	Cleavage; by Caspase 3-7	Highly probable
Q06002	N-myristoyl glycine (post-translational)	433	DVPD | GGKISKAF	0.74	Cleavage; by Caspase 3-7	Probable
P42858	N-myristoyl glycine (post-translational)	524	SATD | GDEEDILS	0	Cleavage; by Caspase 3-7	Improbable
Q06002	N-myristoyl glycine (post-translational)	533	EKED | GLKEEGGP	0	Cleavage; by Caspase 2	Improbable
Q99828	Not myristoylated (post-translational)	-	-	-	-	No caspase cleavage sites found
Q06389	Not myristoylated (post-translational)	-	-	-	-	No caspase cleavage sites found

Figure 19. The main SVMyr results page. Proteins ID with blue color are embedded with a link to a detailed result page.

4. Conclusions

Biology is now a well-established member of the so-called Big Data Sciences, thanks to the technological advancements in the field of "Omics" sciences, which allow to produce large amounts of data in a constantly reducing time [1,2]. This opens exciting perspectives, as well as new challenges. In fact, big data needs to be effectively stored and analyzed. The latter is one of the main goals in bioinformatics, which is addressed via the development of computational tools for sequence annotation. This is done mainly using machine-learning algorithms, which can derive from training examples rules that can be used to make predictions on new data.

In this context, my PhD research focalized on developing machine leaning-based tools for membrane proteins annotation. This class of proteins is of particular relevance performing a wide range of functions and being a target for about the 60% of the approved drugs [14]. Moreover, they are underrepresented in PDB [15], lacking an adequate number of resolved structures, mainly because of technical issues encountered in the crystallization process [16,17]. In particular, I developed two methods, which are made available through web server: BetAware-Deep [4] and SVMyr [5].

BetAware-Deep is designed for prokaryotic TMBB proteins detection and topology prediction. It combines two predictive steps: a BLSTM model, a deep-learning method designed to effectively handle sequential data, and GRHCRFs model, a probabilistic graphical model introducing a regular grammar ensuring biologically relevant predictions. Moreover, BetAware-Deep adopts the PWHM to model the dyad repeat pattern observed in transmembrane segments. The PWHM uses the evolutionary information contained on an MSA and proved to be more effective than the canonical formulation of the hydrophobic moment.

BetAware-Deep outperforms other available state-of-the-art methods for topology prediction in an independent benchmark designed in our study. In addition, it reported results at the level of other predictors in the detection task performed over a large dataset.

BetAware-Deep is made available via a web server at https://busca.biocomp.unibo.it/betaware2, where the user can analyze a protein sequence thanks to a user-friendly interface reporting tabular results and a graphical feature viewer showing predictions along the sequence.

SVMyr is a method designed for co- and post-translational myristoylation prediction in eukaryotic proteins. It is based on SVM and uses as inputs Gly-starting octapeptides, encoded via a per-residue compositional score and mean physicochemical features. The post-translational myristoylation prediction is a unique feature of this method and it is performed via a pattern search for caspase cleavage motifs exposing a glycine.

SVMyr outperforms, in an independent benchmark, other method and patterns available for performing the co-translational myristoylation prediction task. Moreover, it reports good sensitivity in the post-translational myristoylation task.

SVMyr is a fast method designed to analyze large-scale proteomic datasets. Tested on diverse reference proteomes derived from UniProt, it confirms many (if not all) the experimentally annotated myristoylation sites reported by SwissProt in each organism. This analysis may be performed rapidly by SVMyr: the whole human proteome was processed in just 10 minutes.

The method is accessible at https://busca.biocomp.unibo.it/lipipred/. This web server is free, easy to use, and accepts large submission (up to 1000 sequences in FASTA format). It reports prediction in tabular form and allows the user to visualize them with a graphical feature viewer.

The methods here described represent valuable tools that can be used to annotate membrane proteins of great interest: TMBB proteins and myristoylated proteins. They are made available to the entire scientific community and free to use, developed having in mind good practices for the development of machine learning-based tools. Given that, I think that they provide an important contribution in the field, helping researchers to fill the gap between protein sequences and structural/functional data available.

References

1. Pal,S. et al. (2020) Big data in biology: The hope and present-day challenges in it. Gene Reports, 21, 100869.
2. Stephens,Z.D. et al. (2015) Big Data: Astronomical or Genomical? PLoS Biol, 13, e1002195.
3. Altaf-Ul-Amin,Md. et al. (2014) Systems Biology in the Context of Big Data and Networks. BioMed Research International, 2014, 1–11.
4. Madeo,G. et al. (2021) BetAware-Deep: An Accurate Web Server for Discrimination and Topology Prediction of Prokaryotic Transmembrane β-barrel Proteins. Journal of Molecular Biology, 433, 166729.
5. Madeo,G. et al. (2022) SVMyr: a web server detecting co- and post-translational myristoylation in proteins. Journal of Molecular Biology, 167605.
6. Watson,H. (2015) Biological Membranes. Essay in Biochemistry, 59, 43-69.
7. Cooper,G.M. (2002) The Cell: a molecular approach. Sinauer Associates. Structure of the plasma membrane.
8. Alberts,B. et al. (2002) Membrane Proteins in Biology of the Cell. 4th edition. New York: Garland Science.
9. Gromiha,M.M. and Ou,Y.-Y. (2013) Bioinformatics approaches for functional annotation of membrane proteins. Briefings in Bioinformatics, 15, 155–168.
10. White,S.H. and Wimley,W.C. (1999) MEMBRANE PROTEIN FOLDING AND STABILITY: Physical Principles. Annu. Rev. Biophys. Biomol. Struct., 28, 319–365.
11. Ferguson,M.A.J. (1991) Lipid anchors on membrane proteins. Current Opinion in Structural Biology, 4, 522-529.
12. Silhavy,T.J. et al. (2010) The bacterial cell envelope. Cold Spring Hard Perspect Biology, 2(5)a000414.
13. Roumia,A.F. (2020) Landscape of Eukaryotic Transmembrane Beta Barrel Proteins. Journal of Proteome Research, 19(3), 1209-1221.
14. Yıldırım,M.A. et al. (2007) Drug—target network. Nat Biotechnol, 25, 1119–1126.
15. Berman,H.M. (2000) The Protein Data Bank. Nucleic Acids Research, 28, 235–242.
16. Grisshammer,R. and Tateu,C.G. (1995) Overexpression of integral membrane proteins for structural studies. Quart. Rev. Biophys., 28, 315–422.
17. Martin,J. and Sawyer,A. (2019) Elucidating the structure of membrane proteins. BioTechniques, 66, 167–170.
18. Zhang,S. et al. (2018) QTY code enables design of detergent-free chemokine receptors that retain ligand-binding activities. Proc Natl Acad Sci USA, 115, E8652–E8659.
19. Maric,S. et al. (2014) Stealth carriers for low-resolution structure determination of membrane proteins in solution. Acta Cryst D Biol Crystallogr, 70, 317–328.

20. Josts,I. et al. (2018) Conformational States of ABC Transporter MsbA in a Lipid Environment Investigated by Small-Angle Scattering Using Stealth Carrier Nanodiscs. Structure, 26, 1072-1079.e4.
21. Baker,L.A. et al. (2018) Combined 1H-Detected Solid-State NMR Spectroscopy and Electron Cryotomography to Study Membrane Proteins across Resolutions in Native Environments. Structure, 26, 161-170.e3.
22. Larrañaga,P. et al. (2006) Machine learning in bioinformatics. Briefings in Bioinformatics, 7, 86–112.
23. Duda,R.O. and Hart,P. (1973) Pattern Classification and Scene Analysis. Jon Wiley and Sons.
24. Kleinbaum,D.G. et al. (1982) Logistic regression analysis of epidemiologic data: theory and practice. Communications in Statistics - Theory and Methods, 11, 485–547.
25. Breiman,L. et al. (1993) Classification and Regression Trees. Chapman and Hall.
26. Fix,E. and Hodges,J.L.(1989) Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. International Statistical Review / Revue Internationale de Statistique, 57, 238.
27. McCulloch,W.S. and Pitts,W. (1990). A logical calculus of the ideas immanent in nervous activity. 1943. Bulletin of mathematical biology, 52(1-2), 99–97.
28. Cherkassky,V. (1997) The Nature of Statistical Learning Theory. IEEE Trans. Neural Netw., 8, 1564–1564.
29. Baldi, P. (2021) Deep Learning in Science. Cambridge University Press.
30. Hochreiter,S. and Schmidhuber,J. (1997) Long Short-Term Memory. Neural Computation, 9, 1735–1780.
31. Cho,K. et al. (2014) Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).
32. Rumelhart,D.E. et al. (1986) Learning representations by back-propagating errors. Nature, 323, 533–536.
33. Krogh,A. et al. (1994) Hidden Markov Models in Computational Biology. Journal of Molecular Biology, 235, 1501–1531.
34. Suzek,B.E. et al. (2014) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 31, 926–932.
35. Jumper,J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.
36. Walsh,I. et al. (2021) DOME: recommendations for supervised machine learning validation in biology. Nature Methods, 18(10), 1122-1127.
37. Bernal-Llinares, M. et al. (2020) Identifiers.org: Compact Identifier services in the cloud. Bioinformatics, 37(12), 1781-1782.
38. Gray, A.J.G, Goble, C.A. and Jimenez, R., 2017. Bioschemas: From Potato Salad to Protein Annotation. In International Semantic Web Conference (Posters, Demos & Industry Tracks)."
39. Wilkinson,M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

40. Schulz,G.E. (2000) β-Barrel membrane proteins. Current Opinion in Structural Biology, 10, 443–447.
41. Wimley,W.C. (2003) The versatile β-barrel membrane protein. Current Opinion in Structural Biology, 13, 404–411.
42. Galdiero,S. et al. (2007) β-Barrel Membrane Bacterial Proteins: Structure, Function, Assembly and Interaction with Lipids. CPPS, 8, 63–82.
43. Freeman,T.C.,Jr. and Wimley,W.C. (2010) A highly accurate statistical approach for the prediction of transmembrane β-barrels. Bioinformatics, 26, 1965–1974.
44. Remmert,M. et al. (2009) HHomp—prediction and classification of outer membrane proteins. Nucleic Acids Research, 37, W446–W451.
45. Bigelow,H.R. (2004) Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Research, 32, 2566–2577.
46. Bagos,P.G. et al. (2004) PRED-TMBB: a web server for predicting the topology of -barrel outer membrane proteins. Nucleic Acids Research, 32, W400–W404.
47. Hayat,S. and Elofsson,A. (2012) BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. Bioinformatics, 28, 516–522.
48. Savojardo,C. et al. (2013) BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. Bioinformatics, 29, 504–505.
49. Tsirigos,K.D. et al. (2016) PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. Bioinformatics, 32, i665–i671.
50. Hayat,S. et al. (2016) Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β-barrel proteins. Bioinformatics, 32, 1571–1573.
51. Fariselli,P. et al. (2009) Grammatical-Restrained Hidden Conditional Random Fields for Bioinformatics applications. Algorithms Mol Biol, 4.
52. Eisenberg,D. et al. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. Proceedings of the National Academy of Sciences, 81, 140–144.
53. Jiang,H. et al. (2018) Protein Lipidation: Occurrence, Mechanisms, Biological Functions, and Enabling Technologies. Chem. Rev., 118, 919–988.
54. Giang,D.K. and Cravatt,B.F. (1998) A Second Mammalian N-Myristoyltransferase. Journal of Biological Chemistry, 273, 6595–6598.
55. Castrec,B. et al. (2018) Structural and genomic decoding of human and plant myristoylomes reveals a definitive recognition pattern. Nat Chem Biol, 14, 671–679.
56. Perinpanayagam,M.A. et al. (2012) Regulation of co- and post-translational myristoylation of proteins during apoptosis: interplay of N -myristoyltransferases and caspases. FASEB j., 27, 811–821.
57. Dian,C. et al. (2020) High-resolution snapshots of human N-myristoyltransferase in action illuminate a mechanism promoting N-terminal Lys and Gly myristoylation. Nat Commun, 11.

58. Giglione,C. and Meinnel,T. (2022) Mapping the myristoylome through a complete understanding of protein myristoylation biochemistry. Progress in Lipid Research, 85, 101139.
59. Weston,S.A. et al. (1998) Crystal structure of the anti-fungal target N-myristoyl transferase. Nat Struct Mol Biol, 5, 213–221.
60. Farazi,T.A. et al. (2001) Structures of Saccharomyces cerevisiae N-myristoyltransferase with Bound MyristoylCoA and Peptide Provide Insights about Substrate Recognition and Catalysis. Biochemistry, 40, 6335–6343.
61. Bhatnagar,R.S. et al. (1998) Structure of N-myristoyltransferase with bound myristoylCoA and peptide substrate analogs. Nat Struct Mol Biol, 5, 1091–1097.
62. Maurer-Stroh,S. et al. (2002) N-terminal N -myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences 1 1Edited by J. Thornton. Journal of Molecular Biology, 317, 523–540.
63. McWherter,C.A. et al. (1997) Scanning Alanine Mutagenesis and De-peptidization of a Candida albicans Myristoyl-CoA:ProteinN-Myristoyltransferase Octapeptide Substrate Reveals Three Elements Critical for Molecular Recognition. Journal of Biological Chemistry, 272, 11874–11880.
64. Meinnel,T. et al. (2020) Myristoylation, an Ancient Protein Modification Mirroring Eukaryogenesis and Evolution. Trends in Biochemical Sciences, 45, 619–632.
65. Nelson,A.R. et al. (2007) Myristoyl-Based Transport of Peptides into Living Cells. Biochemistry, 46, 14771–14781.
66. Hackett,M.J. et al. (2013) Fatty acids as therapeutic auxiliaries for oral and parenteral formulations. Advanced Drug Delivery Reviews, 65, 1331–1339.
67. Hang,H.C. et al. (2011) Bioorthogonal Chemical Reporters for Analyzing Protein Lipidation and Lipid Trafficking. Acc. Chem. Res., 44, 699–708.
68. Traverso,J.A. et al. (2013) High-throughput profiling of N-myristoylation substrate specificity across species including pathogens. Proteomics, 13, 25–36.
69. Hulo,N. (2006) The PROSITE database. Nucleic Acids Research, 34, D227–D230.
70. Maurer-Stroh,S. et al. (2002) N-terminal N -myristoylation of proteins: prediction of substrate proteins from amino acid sequence 1 1Edited by J. Thornton. Journal of Molecular Biology, 317, 541–557.
71. Bologna,G. et al. (2004) N-Terminal myristoylation predictions by ensembles of neural networks. PROTEOMICS, 4, 1626–1632.
72. Martinez,A. et al. (2008) Extent of N-terminal modifications in cytosolic proteins from eukaryotes. Proteomics, 8, 2809–2831.
73. The UniProt consortium (2020) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49, D480–D489.
74. Thinon,E. et al. (2014) Global profiling of co- and post-translationally N-myristoylated proteomes in human cells. Nat Commun, 5.

75. Graves,A. and Schmidhuber,J. (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18, 602–610.
76. Lomize,M.A. et al. (2011) OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Research, 40, D370–D376.
77. Fox,N.K. et al. (2013) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucl. Acids Res., 42, D304–D309.
78. Suzek,B.E. et al. (2014) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 31, 926–932.
79. Altschul,S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25, 3389–3402.
80. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22, 2577–2637.
81. White,S.H. and Wimley,W.C. (1998) Hydrophobic interactions of peptides with membrane interfaces. Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes, 1376, 339–352.
82. Bengio,Y. et al. (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw., 5, 157–166.
83. Hochreiter,S. et al. (2001) Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. A Field Guide to Dynamical Recurrent Networks. IEEE Press.
84. Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. IEEE Trans. Signal Process., 45, 2673–2681.
85. Kingma,D.P. and Ba,J.L. (2015) Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.
86. Chollet,F. (2015) Keras. GitHub; https://github.com/fchollet/keras.
87. Lafferty,J. et al. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, 282-289.
88. Zemla,A. et al. (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins, 34, 220–223.
89. Paladin,L. et al. (2020) The Feature-Viewer: a visualization tool for positional annotations on a sequence, Bioinformatics, 36 (10), 3244–3245.
90. Wright,M.H. et al. (2016) Global Profiling and Inhibition of Protein Lipidation in Vector and Host Stages of the Sleeping Sickness Parasite Trypanosoma brucei. ACS Infectious Diseases, 2(6), 427-441.
91. Roberts,A.J. and Fairlamb,A.H. (2016) The N-myristoylome of Trypanosoma cruzi. Scientific Reports, 6, 31078.

92. Wright,M.H. et al. (2015) Global Analysis of Protein N-Myristoylation and Exploration of N-Myristoyltransferase as a Drug Target in the Neglected Human Pathogen Leishmania donovani. Chemistry and Biology. 22(3), 342-354.
93. Wright,M.H. et al. (2014) Validation of N-myristoyltransferase as an antimalarial drug target using an integrated chemical biology approach. Nat. Chem. 6, 112–121.
94. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology, 157, 105–132.
95. Kawashima,S. (2000) AAindex: Amino Acid index database. Nucleic Acids Research, 28, 374–374.
96. Chou,P.Y. and Fasman,G.D. (1974) Prediction of protein conformation. Biochemistry, 13, 222–245.
97. Kumar,M. et al. (2019) ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Research.
98. Yuan,M. et al. (2020) N-myristoylation: from cell biology to translational medicine. Acta Pharmacol Sin., 41, 1005–1015.

Appendix 1 – Positive Training Set of BetAware-Deep (58 proteins)

UniProt ID	Organism	PDB ID	Chain	Experimental Method	Resolution (Å)
P22340	*Salmonella typhimurium*	1a0s	P	X-RAY DIFFRACTION	2.4
P05825	*Escherichia coli*	1fep	A	X-RAY DIFFRACTION	2.4
P0A921	*Escherichia coli*	1fw2	A	X-RAY DIFFRACTION	2.6
P39767	*Rhodobacter blasticus*	1h6s	1	X-RAY DIFFRACTION	3
Q51227	*Neisseria meningitidis*	1k24	A	X-RAY DIFFRACTION	2.03
P13036	*Escherichia coli*	1kmo	A	X-RAY DIFFRACTION	2
P37001	*Escherichia coli*	1mm4	A	SOLUTION NMR	
P26466	*Salmonella typhimurium*	1mpr	A	X-RAY DIFFRACTION	2.8
Q9RP17	*Neisseria meningitidis*	1p4t	A	X-RAY DIFFRACTION	2.55
P10384	*Escherichia coli*	1t16	A	X-RAY DIFFRACTION	2.6
P0A927	*Escherichia coli*	1tlw	A	X-RAY DIFFRACTION	3.1
P02930	*Escherichia coli*	1tqq	A	X-RAY DIFFRACTION	2.75
Q8GKS5	*Neisseria meningitidis*	1uyn	X	X-RAY DIFFRACTION	2.6
Q51487	*Pseudomonas aeruginosa*	1wp1	A	X-RAY DIFFRACTION	2.56
P42512	*Pseudomonas aeruginosa*	1xkw	A	X-RAY DIFFRACTION	2
Q9HVD1	*Pseudomonas aeruginosa*	2erv	A	X-RAY DIFFRACTION	2
P76045	*Escherichia coli*	2f1c	X	X-RAY DIFFRACTION	2.3
P0A910	*Escherichia coli*	2ge4	A	SOLUTION NMR	
P17315	*Escherichia coli*	2hdi	A	X-RAY DIFFRACTION	2.5
P48632	*Pseudomonas aeruginosa*	2iah	A	X-RAY DIFFRACTION	2.73
P06996	*Escherichia coli*	2j1n	C	X-RAY DIFFRACTION	2
P24017	*Klebsiella pneumoniae*	2k0l	A	SOLUTION NMR	
Q51486	*Pseudomonas aeruginosa*	2lhf	A	SOLUTION NMR	
P05695	*Pseudomonas aeruginosa*	2o4v	A	X-RAY DIFFRACTION	1.94
P69856	*Escherichia coli*	2wjq	A	X-RAY DIFFRACTION	2
Q9HWW1	*Pseudomonas aeruginosa*	2x27	X	X-RAY DIFFRACTION	2.4
P17811	*Yersinia pestis*	2x4m	A	X-RAY DIFFRACTION	2.55
Q8GNN6	*Escherichia coli*	2ynk	A	X-RAY DIFFRACTION	2.64
P06129	*Escherichia coli*	2ysu	A	X-RAY DIFFRACTION	3.5
Q9RBW8	*Ralstonia pickettii*	3bry	A	X-RAY DIFFRACTION	3.2
Q79AD2	*Serratia marcescens*	3csl	A	X-RAY DIFFRACTION	2.7
Q9HVJ6	*Pseudomonas aeruginosa*	3dwo	X	X-RAY DIFFRACTION	2.2
Q72JD8	*Thermus thermophilus*	3dzm	A	X-RAY DIFFRACTION	2.801
Q48152	*Haemophilus influenzae*	3emo	A	X-RAY DIFFRACTION	3
P72412	*Shigella dysenteriae*	3fhh	A	X-RAY DIFFRACTION	2.6
Q8ZPT3	*Salmonella typhimurium*	3fid	A	X-RAY DIFFRACTION	1.9
P02931	*Escherichia coli*	3hw9	A	X-RAY DIFFRACTION	2.61
O33407	*Pseudomonas aeruginosa*	3kvn	X	X-RAY DIFFRACTION	2.499
P35077	*Bordetella pertussis*	3njt	A	X-RAY DIFFRACTION	3.5
C5I2D9	*Pseudomonas fluorescens*	3qlb	A	X-RAY DIFFRACTION	3.26
Q45340	*Bordetella pertussis*	3qq2	A	X-RAY DIFFRACTION	3
Q8D0Z7	*Yersinia pestis*	3qra	A	X-RAY DIFFRACTION	1.801
P18895	*Pseudomonas aeruginosa*	3rbh	A	X-RAY DIFFRACTION	2.301
P30130	*Escherichia coli*	3rfz	B	X-RAY DIFFRACTION	2.8
Q7BSW5	*Escherichia coli*	3slj	A	X-RAY DIFFRACTION	2.481

Q9HVS0	*Pseudomonas aeruginosa*	3syb	A	X-RAY DIFFRACTION	2.7
Q9K0U9	*Neisseria meningitidis*	3v89	A	X-RAY DIFFRACTION	3.1
Q5RT80	*Neisseria meningitidis*	3vzt	X	X-RAY DIFFRACTION	2.3
P0ADE4	*Escherichia coli*	4c00	A	X-RAY DIFFRACTION	2.25
P11922	*Yersinia pseudotuberculosis*	4e1t	A	X-RAY DIFFRACTION	2.263
P46359	*Yersinia pestis*	4epa	A	X-RAY DIFFRACTION	3.2
A5VZA8	*Pseudomonas putida*	4gey	A	X-RAY DIFFRACTION	2.7
Q934G3	*Dickeya dadantii*	4pr7	A	X-RAY DIFFRACTION	2.1
Q83SQ0	*Shigella flexneri*	4q35	A	X-RAY DIFFRACTION	2.393
A6XB80	*Acinetobacter baumannii*	4rl9	A	X-RAY DIFFRACTION	2.7
Q9I5U2	*Pseudomonas aeruginosa*	5iva	A	X-RAY DIFFRACTION	2.988
P0C6Q6	*Vibrio cholerae*	5onu	A	X-RAY DIFFRACTION	2.22
P0A942	*Escherichia coli*	6fsu	A	X-RAY DIFFRACTION	2.6

Appendix 2 – Negative Training Set of BetAware-Deep (69 proteins)

UniProt ID	Organism	PDB ID	Chain	Experimental Method	Resolution (Å)
P0ABD8	*Escherichia coli*	1a6x	A	SOLUTION NMR	
P07103	*Dickeya dadantii*	1aiw	A	SOLUTION NMR	
Q53654	*Staphylococcus aureus*	1amx	A	X-RAY DIFFRACTION	2
P00778	*Lysobacter enzymogenes*	1boq	A	X-RAY DIFFRACTION	2.1
P01092	*Streptomyces tendae*	1bvn	T	X-RAY DIFFRACTION	2.5
P04766	*Geobacillus stearothermophilus*	1d1n	A	SOLUTION NMR	
Q46079	*Pedobacter heparinus*	1dbg	A	X-RAY DIFFRACTION	1.7
P00646	*Escherichia coli*	1e44	B	X-RAY DIFFRACTION	2.4
P56930	*Thermus thermophilus*	1feu	A	X-RAY DIFFRACTION	2.3
P0CL66	*Borrelia burgdorferi*	1fj1	E	X-RAY DIFFRACTION	2.68
Q9RCK8	*Streptomyces tendae*	1g6e	A	SOLUTION NMR	
Q05128	*Zaire ebolavirus*	1h2d	A	X-RAY DIFFRACTION	2.6
Q820S6	*Nitrosomonas europaea*	1iby	A	X-RAY DIFFRACTION	1.65
Q2FV99	*Staphylococcus aureus*	1ija	A	SOLUTION NMR	
P0A7C2	*Escherichia coli*	1jhc	A	X-RAY DIFFRACTION	2
Q9RP27	*Haemophilus influenzae*	1jov	A	X-RAY DIFFRACTION	1.57
P39805	*Bacillus subtilis*	1l1c	A	SOLUTION NMR	
P14930	*Neisseria gonorrhoeae*	1l1d	A	X-RAY DIFFRACTION	1.85
P12293	*Paracoccus denitrificans*	1lrw	A	X-RAY DIFFRACTION	2.5
Q55418	*Synechocystis sp.*	1mi8	A	X-RAY DIFFRACTION	2
P9WNF5	*Mycobacterium tuberculosis*	1nyo	A	SOLUTION NMR	
O85094	*Pseudomonas savastanoi*	1o9y	A	X-RAY DIFFRACTION	2.29
O32079	*Bacillus subtilis*	1oru	A	X-RAY DIFFRACTION	1.8
P45206	*Haemophilus influenzae*	1ou9	A	X-RAY DIFFRACTION	1.8
O66640	*Aquifex aeolicus*	1p6v	A	X-RAY DIFFRACTION	3.2
Q57221	*Yersinia pseudotuberculosis*	1pm4	A	X-RAY DIFFRACTION	1.755
P0A790	*Escherichia coli*	1ppy	A	X-RAY DIFFRACTION	1.95
P0A6Y8	*Escherichia coli*	1q5l	A	SOLUTION NMR	
Q7A189	*Staphylococcus aureus*	1qwx	A	X-RAY DIFFRACTION	1.5
P76344	*Escherichia coli*	1s7d	A	X-RAY DIFFRACTION	2.17
Q92RG6	*Rhizobium meliloti*	1so9	A	SOLUTION NMR	
O06522	*Haemophilus ducreyi*	1sr4	A	X-RAY DIFFRACTION	2
P9WHW3	*Mycobacterium tuberculosis*	1w74	A	X-RAY DIFFRACTION	2.6
P04450	*Geobacillus stearothermophilus*	1whi	A	X-RAY DIFFRACTION	1.5
Q5SM30	*Thermus thermophilus*	1wk2	A	X-RAY DIFFRACTION	2.5
P96142	*Thermus thermophilus*	1wka	A	X-RAY DIFFRACTION	1.7
P0A1J1	*Salmonella typhimurium*	1wlg	A	X-RAY DIFFRACTION	1.8
P28248	*Escherichia coli*	1xs1	A	X-RAY DIFFRACTION	1.8
P08874	*Bacillus subtilis*	1yfb	A	SOLUTION NMR	
P74334	*Synechocystis sp.*	2biw	A	X-RAY DIFFRACTION	2.39
P77667	*Escherichia coli*	2d2a	A	X-RAY DIFFRACTION	2.7
Q8YSC3	*Nostoc sp.*	2ii7	A	X-RAY DIFFRACTION	2.8
Q7X4S4	*Bacillus licheniformis*	2jem	A	X-RAY DIFFRACTION	1.78
Q84B82	*Aliivibrio fischeri*	2kmt	A	SOLUTION NMR	
Q892G2	*Clostridium tetani*	2qyz	A	X-RAY DIFFRACTION	2.04

B4EH87	*Burkholderia cenocepacia*	2vnv	A	X-RAY DIFFRACTION	1.7
Q9I0M4	*Pseudomonas aeruginosa*	2w7q	A	X-RAY DIFFRACTION	1.88
O80297	*Escherichia phage If1*	2x9a	A	X-RAY DIFFRACTION	2.47
B1JSA0	*Yersinia pseudotuberculosis*	2y6t	A	X-RAY DIFFRACTION	2.74
Q2MDE2	*Microcystis aeruginosa*	2yhh	A	SOLUTION NMR	
O84671	*Chlamydia trachomatis*	3gqs	A	X-RAY DIFFRACTION	2.2
P18429	*Bacillus subtilis*	3hd8	B	X-RAY DIFFRACTION	2.39
A1S3D0	*Shewanella amazonensis*	3hsa	A	X-RAY DIFFRACTION	1.99
Q7WTN6	*Rhodothermus marinus*	3jxs	A	X-RAY DIFFRACTION	1.6
E5Q9D7	*Mycolicibacterium thermoresistibile*	3nfw	A	X-RAY DIFFRACTION	1.6
P19478	*Treponema pallidum*	3pjl	A	X-RAY DIFFRACTION	1.7
Q6VAL9	*Pseudorhizobium banfieldiae*	4aay	B	X-RAY DIFFRACTION	2.7
E8T502	*Thermovibrio ammonificans*	4c3t	A	X-RAY DIFFRACTION	1.69
Q8XXK6	*Ralstonia solanacearum*	4csd	A	X-RAY DIFFRACTION	1.35
Q2SWY6	*Burkholderia thailandensis*	4eqy	A	X-RAY DIFFRACTION	1.8
Q8DKB4	*Thermosynechococcus elongatus*	4n8f	A	X-RAY DIFFRACTION	2
Q2YMM2	*Brucella abortus*	4q14	A	X-RAY DIFFRACTION	1.7
P44602	*Haemophilus influenzae*	4rt6	A	X-RAY DIFFRACTION	2.8
P02976	*Staphylococcus aureus*	4y4y	A	X-RAY DIFFRACTION	3
E6UBR9	*Ruminococcus albus*	5ayd	A	X-RAY DIFFRACTION	2.3
A6TD90	*Klebsiella pneumoniae*	5cai	A	X-RAY DIFFRACTION	2.3
A4ISU9	*Geobacillus thermodenitrificans*	5dmb	D	X-RAY DIFFRACTION	2.301
O34714	*Bacillus subtilis*	5hi0	A	X-RAY DIFFRACTION	2.602
Q9PBB0	*Xylella fastidiosa*	5j7n	A	X-RAY DIFFRACTION	2.9

Appendix 3 – Blind Test Set of BetAware-Deep (15 proteins)

UniProt ID	Organism	PDB ID	Chain	Experimental Method	Resolution (Å)
A1JUB7	*Yersinia enterocolitica*	2lme	A	SOLID-STATE NMR	
A0A0F6C2F5	*Escherichia coli*	4q79	A	X-RAY DIFFRACTION	3.1
A5W3Z9	*Pseudomonas putida*	4rl8	A	X-RAY DIFFRACTION	2.3
Q48391	*Klebsiella oxytoca*	4v3g	A	X-RAY DIFFRACTION	2.513
Q6D8U4	*Pectobacterium atrosepticum*	4zgv	A	X-RAY DIFFRACTION	3.2
Q8A5H5	*Bacteroides thetaiotaomicron*	5fq6	D	X-RAY DIFFRACTION	2.8
Q659I5	*Campylobacter jejuni*	5ldv	A	X-RAY DIFFRACTION	2.1
O86021	*Vibrio cholerae*	5oyk	A	X-RAY DIFFRACTION	3.2
P45758	*Escherichia coli*	5wq7	A	ELECTRON MICROSCOPY	3.04
A0A062F4L9	*Acinetobacter baumannii*	6gie	A	X-RAY DIFFRACTION	2.1
Q5I6C7	*Flavobacterium johnsoniae*	6h3i	A	ELECTRON MICROSCOPY	3.5
A5FJM7	*Flavobacterium johnsoniae*	6h3i	F	ELECTRON MICROSCOPY	3.5
Q76HJ9	*Acinetobacter baumannii*	6hcp	A	X-RAY DIFFRACTION	1.83
Q00595	*Pseudomonas oleovorans*	6qam	A	SOLUTION NMR	
P37650	*Escherichia coli*	6tzk	A	X-RAY DIFFRACTION	1.852

Appendix 4 – Positive Training Set of SVMyr (232 proteins)

UniProt ID	Organism	Octapeptide	Annotation Source
Q99828	*Homo sapiens*	GGSGSRLS	SwissProt
P42325	*Drosophila melanogaster*	GKQNSKLK	SwissProt
Q06389	*Saccharomyces cerevisiae*	GAKTSKLS	SwissProt
Q9NUM4	*Homo sapiens*	GKSLSHLP	SwissProt
Q9FIL1	*Arabidopsis thaliana*	GPRCSKLS	SwissProt
O81223	*Arabidopsis thaliana*	GCSVSKKK	SwissProt
Q8IZE3	*Homo sapiens*	GSENSALK	SwissProt
P34164	*Saccharomyces cerevisiae*	GTTTSHPA	SwissProt
Q8W4L3	*Arabidopsis thaliana*	GGQCSSLS	SwissProt
Q9D6Y7-2	*Mus musculus*	GDSASKVI	SwissProt
Q05175	*Rattus norvegicus*	GSKLSKKK	SwissProt
P53701	*Homo sapiens*	GLSPSAPA	SwissProt
Q8MMZ7	*Toxoplasma gondii*	GACISKNS	SwissProt
Q8WU20	*Homo sapiens*	GSCCSCPD	SwissProt
Q9M324	*Arabidopsis thaliana*	GARCSKFS	SwissProt
Q9FHD7	*Arabidopsis thaliana*	GCEVSKLS	SwissProt
Q22663	*Caenorhabditis elegans*	GSSTSTPA	SwissProt
P11076	*Saccharomyces cerevisiae*	GLFASKLF	SwissProt
Q99653	*Homo sapiens*	GSRASTLL	SwissProt
P18085	*Homo sapiens*	GLTISSLF	SwissProt
P0CM16	*Cryptococcus neoformans*	GLSVSKLL	SwissProt
P12931	*Homo sapiens*	GSNKSKPK	SwissProt
P62166	*Homo sapiens*	GKSNSKLK	SwissProt
Q99618	*Homo sapiens*	GSAKSVPV	SwissProt
P39968	*Saccharomyces cerevisiae*	GSCCSCLK	SwissProt
Q65200	*African swine fever virus*	GGSTSKNS	SwissProt
Q9H4G4	*Homo sapiens*	GKSASKQF	SwissProt
Q09711	*Schizosaccharomyces pombe*	GKSQSKLS	SwissProt
Q9VLL3	*Drosophila melanogaster*	GKAQSKRS	SwissProt
P34727	*Ajellomyces capsulatus*	GMAFSKLF	SwissProt
P62330	*Homo sapiens*	GKVLSKIF	SwissProt
P25020	*Rous sarcoma virus*	GSSKSKPK	SwissProt
P80723	*Homo sapiens*	GGKLSKKK	SwissProt
P37235	*Homo sapiens*	GKQNSKLR	SwissProt
P05961	*Human immunodeficiency virus type 1*	GARASVLS	SwissProt
Q5T2Q4	*Homo sapiens*	GNILTCCV	SwissProt
Q9Y478	*Homo sapiens*	GNTSSERA	SwissProt
Q8N7R7	*Homo sapiens*	GNTLTCCV	SwissProt
P07612	*Vaccinia virus*	GAAASIQT	SwissProt
O43741	*Homo sapiens*	GNTTSDRV	SwissProt
Q13286	*Homo sapiens*	GGCAGSRR	SwissProt
Q9EPI6	*Rattus norvegicus*	GAAASRRR	SwissProt
Q8ND76	*Homo sapiens*	GNTTSCCV	SwissProt
P68446	*Vaccinia virus*	GTAATIQT	SwissProt

Q9ULE6	*Homo sapiens*	GTTASTAQ	SwissProt
Q9CRB9	*Mus musculus*	GGTASTRR	SwissProt
Q9NX63	*Homo sapiens*	GGTTSTRR	SwissProt
P00387	*Homo sapiens*	GAQLSTLG	SwissProt
P16710	*Vaccinia virus*	GAAVTLNR	SwissProt
P22219	*Saccharomyces cerevisiae*	GAQLSLVV	SwissProt
Q66282	*Coxsackievirus B3*	GAQVSTQK	SwissProt
P03093	*Simian virus 40*	GAALTLLG	SwissProt
P03096	*Murine polyomavirus*	GAALTILV	SwissProt
P03300	*Poliovirus type 1*	GAQVSSQK	SwissProt
P10823	*Saccharomyces cerevisiae*	GLCASSEK	SwissProt
P49006	*Homo sapiens*	GSQSSKAP	SwissProt
P07947	*Homo sapiens*	GCIKSKEN	SwissProt
F4I7Y4	*Arabidopsis thaliana*	GCCQSSFL	SwissProt
O75896	*Homo sapiens*	GASGSKAR	SwissProt
P29966	*Homo sapiens*	GAQFSKTA	SwissProt
P16051	*Dictyostelium discoideum*	GICASSME	SwissProt
Q9LS26	*Arabidopsis thaliana*	GCLHSKTA	SwissProt
Q944A7	*Arabidopsis thaliana*	GCCQSLFS	SwissProt
Q0D715	*Oryza sativa subsp. japonica*	GLCSSSSA	SwissProt
O81445	*Arabidopsis thaliana*	GCFHSKAA	SwissProt
Q9UPV7	*Homo sapiens*	GVLMSKRQ	SwissProt
Q9NS86	*Homo sapiens*	GETMSKRL	SwissProt
P11283	*Mouse mammary tumor virus*	GVSGSKGQ	SwissProt
Q5VT66	*Homo sapiens*	GAAGSSAL	SwissProt
P53139	*Saccharomyces cerevisiae*	GLCGSKTQ	SwissProt
Q84ME1	*Arabidopsis thaliana*	GISLSKRR	SwissProt
Q9LTB8	*Arabidopsis thaliana*	GCFHSTAA	SwissProt
Q7XJT7	*Arabidopsis thaliana*	GCCYSLSS	SwissProt
Q6IAA8	*Homo sapiens*	GCCYSSEN	SwissProt
Q9LYB4	*Arabidopsis thaliana*	GASSSSSV	SwissProt
O95843	*Homo sapiens*	GNGKSIAG	SwissProt
P21457	*Bos taurus*	GNSKSGAL	SwissProt
P04289	*Human herpesvirus 1*	GLSFSGAR	SwissProt
Q9BVX2	*Homo sapiens*	GSQHSAAA	SwissProt
Q7XJR9	*Arabidopsis thaliana*	GLCFSSAA	SwissProt
Q86XR7	*Homo sapiens*	GIGKSKIN	SwissProt
O75695	*Homo sapiens*	GCFFSKRR	SwissProt
Q969Z3	*Homo sapiens*	GASSSSAL	SwissProt
Q8R4L0	*Mus musculus*	GSLSSRGK	SwissProt
Q96BS2	*Homo sapiens*	GAAHSASE	SwissProt
Q717R9	*Homo sapiens*	GSGSSRSS	SwissProt
P08754	*Homo sapiens*	GCTLSAED	SwissProt
A8MTJ3	*Homo sapiens*	GSGISSES	SwissProt
P08631	*Homo sapiens*	GGRSSCED	SwissProt
Q8R4T1	*Mus musculus*	GSGSSRSG	SwissProt
Q9H6Q3	*Homo sapiens*	GSLPSRRK	SwissProt

Q8CFC9	*Rattus norvegicus*	GSVSSLIS	SwissProt
P19087	*Homo sapiens*	GSGASAED	SwissProt
Q9CB01	*Arabidopsis thaliana*	GCASSLPD	SwissProt
Q9H8Y8	*Homo sapiens*	GSSQSVEI	SwissProt
P0C6Y6	*Porcine reproductive and respiratory syndrome virus*	GSLWSKIS	SwissProt
Q7Z494	*Homo sapiens*	GTASSLVS	SwissProt
P11488	*Homo sapiens*	GAGASAEE	SwissProt
Q969J3	*Homo sapiens*	GSEQSSEA	SwissProt
P11078	*Reovirus type 3*	GNASSIVQ	SwissProt
Q9MB58	*Arabidopsis thaliana*	GSGASKNT	SwissProt
Q9NRX5	*Homo sapiens*	GSVLGLCS	SwissProt
Q96TA1	*Homo sapiens*	GDVLSTHL	SwissProt
P04899	*Homo sapiens*	GCTVSAED	SwissProt
P09471	*Homo sapiens*	GCTLSAEE	SwissProt
Q96EQ8	*Homo sapiens*	GSVLSTDS	SwissProt
Q96SL1	*Homo sapiens*	GSRWSSEE	SwissProt
Q7RTS9	*Homo sapiens*	GSNSSRIG	SwissProt
A4GNA8	*Arabidopsis thaliana*	GNGNSTET	SwissProt
P18541	*Lymphocytic choriomeningitis virus*	GQGKSREE	SwissProt
O60291	*Homo sapiens*	GSILSRRI	SwissProt
P07611	*Vaccinia virus*	GGGVSVEL	SwissProt
Q02952	*Homo sapiens*	GAGSSTEQ	SwissProt
Q91DM1	*Equine arteritis virus*	GLVWSLIS	SwissProt
Q96MG8	*Homo sapiens*	GGAVSAGE	SwissProt
Q9BRQ8	*Homo sapiens*	GSQVSVES	SwissProt
O15121	*Homo sapiens*	GSRVSRED	SwissProt
Q08358	*African swine fever virus*	GNRGSSTS	SwissProt
O43149	*Homo sapiens*	GNAPSHSS	SwissProt
P21137-8	*Caenorhabditis elegans*	GNAASGGS	SwissProt
Q96PY5	*Homo sapiens*	GNAGSMDS	SwissProt
P84077	*Homo sapiens*	GNIFANLF	SwissProt
P61204	*Homo sapiens*	GNIFGNLL	SwissProt
Q923S6	*Mus musculus*	GNNFSSVS	SwissProt
P38116	*Saccharomyces cerevisiae*	GNIFSSMF	SwissProt
Q96KC2	*Homo sapiens*	GLIFAKLW	SwissProt
Q9S752	*Arabidopsis thaliana*	GNISSSGG	SwissProt
Q60898	*Mus musculus*	GNSMKSTS	SwissProt
P40616	*Homo sapiens*	GGFFSSIF	SwissProt
G5EBX9	*Caenorhabditis elegans*	GCGPSSGR	SwissProt
P40327	*Saccharomyces cerevisiae*	GQGVSSGQ	SwissProt
P06239	*Homo sapiens*	GCGCSSHP	SwissProt
Q96LZ3	*Homo sapiens*	GNEASYPA	SwissProt
O88778	*Rattus norvegicus*	GNEASLEG	SwissProt
P63098	*Homo sapiens*	GNEASYPL	SwissProt
Q8IVF7	*Homo sapiens*	GNLESAEG	SwissProt
Q13237	*Homo sapiens*	GNGSVKPK	SwissProt
Q8W4I7	*Arabidopsis thaliana*	GNCCRSPA	SwissProt

Q9Y3C5	*Homo sapiens*	GNCLKSPT	SwissProt
Q9NPB3	*Homo sapiens*	GNCAKRPW	SwissProt
Q9NUQ9	*Homo sapiens*	GNLLKVLT	SwissProt
P17568	*Homo sapiens*	GAHLVRRY	SwissProt
Q02368	*Bos taurus*	GAHLARRY	SwissProt
P18064	*Arabidopsis thaliana*	GLLCSRSR	SwissProt
Q9C0E8	*Homo sapiens*	GGLFSRWR	SwissProt
Q8IV36	*Homo sapiens*	GSTDSKLN	SwissProt
O15355	*Homo sapiens*	GAYLSQPN	SwissProt
P63252	*Homo sapiens*	GSVRTNRY	SwissProt
Q06850	*Arabidopsis thaliana*	GNTCVGPS	SwissProt
Q38870	*Arabidopsis thaliana*	GNACVGPN	SwissProt
Q8IWE4	*Homo sapiens*	GQCVTKCK	SwissProt
P26313	*Junin mammarenavirus*	GQFISFMQ	SwissProt
Q9LU05	*Arabidopsis thaliana*	GYWKSKVV	SwissProt
Q96262	*Arabidopsis thaliana*	GYWNSKVV	SwissProt
Q04272	*Saccharomyces cerevisiae*	GQKSSKVH	SwissProt
P13200	*Human cytomegalovirus*	GAELCKRI	SwissProt
P27216	*Homo sapiens*	GNRHAKAS	SwissProt
Q8N9N7	*Homo sapiens*	GNSALRAH	SwissProt
O73557	*Lassa virus*	GNKQAKAP	SwissProt
P42526	*Dictyostelium discoideum*	GNRAFKAH	SwissProt
P13231	*Dictyostelium discoideum*	GNRAFKSH	SwissProt
Q564G3	*Rattus norvegicus*	GNSAARSD	SwissProt
P62241	*Homo sapiens*	GISRDNWH	SwissProt
Q14254	*Homo sapiens*	GNCHTVGP	SwissProt
P03355	*Moloney murine leukemia virus*	GQTVTTPL	SwissProt
Q9SG12	*Arabidopsis thaliana*	GHCYSRNI	SwissProt
Q6P9B6	*Homo sapiens*	GNSRSRVG	SwissProt
P38405	*Homo sapiens*	GCLGGNSK	SwissProt
Q38868	*Arabidopsis thaliana*	GNCFAKNH	SwissProt
P06241	*Homo sapiens*	GCVQCKDK	SwissProt
O75410-5	*Homo sapiens*	GGSHSQTP	SwissProt
Q14699	*Homo sapiens*	GCGLNKLE	SwissProt
P63092	*Homo sapiens*	GCLGNSKT	SwissProt
Q8NHG7	*Homo sapiens*	GLCFPCPG	SwissProt
O43687	*Homo sapiens*	GQLCCFPF	SwissProt
P68181	*Mus musculus*	GNTAIAKK	SwissProt
P17612	*Homo sapiens*	GNAAAAKK	SwissProt
P22694	*Homo sapiens*	GNAATAKK	SwissProt
P79880	*Gallus gallus*	GNMDGKAV	SwissProt
Q86XD5-2	*Homo sapiens*	GCIGSRTV	SwissProt
Q920K5	*Rattus norvegicus*	GCGGSRAD	SwissProt
P62191	*Homo sapiens*	GQSQSGGH	SwissProt
P46065	*Bos taurus*	GNIMDGKS	SwissProt
P09108	*Tetronarce californica*	GQDQTKQQ	SwissProt
Q8NHG8	*Homo sapiens*	GAKQSGPA	SwissProt

Q38872	*Arabidopsis thaliana*	GNSCRGSF	SwissProt
Q99570	*Homo sapiens*	GNQLAGIA	SwissProt
P25296	*Saccharomyces cerevisiae*	GAAPSKIV	SwissProt
O00408	*Homo sapiens*	GQACGHSI	SwissProt
Q9UJ68-5	*Homo sapiens*	GNSASNIV	SwissProt
P29728	*Homo sapiens*	GNGESQLS	SwissProt
Q8WU08	*Homo sapiens*	GANTSRKP	SwissProt
Q9P203	*Homo sapiens*	GANASNYP	SwissProt
Q96PX1	*Homo sapiens*	GALTSRQH	SwissProt
P11801	*Homo sapiens*	GCGTSKVL	SwissProt
Q7FZF1	*Arabidopsis thaliana*	GCVCSKQL	SwissProt
Q9H1R2-3	*Homo sapiens*	GNGMTKVL	SwissProt
Q9NRW4	*Homo sapiens*	GNGMNKIL	SwissProt
Q494U1	*Homo sapiens*	GNSHCVPQ	SwissProt
Q9P206-3	*Homo sapiens*	GNSHHKRK	SwissProt
Q8TB92	*Homo sapiens*	GNVPSAVK	SwissProt
Q5M775-3	*Homo sapiens*	GNHSGRPE	SwissProt
P00519-2	*Homo sapiens*	GQQPGKVL	SwissProt
P03145	*Duck hepatitis B virus*	GQHPAKSM	SwissProt
P08473	*Homo sapiens*	GKSESQMD	SwissProt
Q9SCY5	*Arabidopsis thaliana*	GNVNAREE	SwissProt
P16452	*Homo sapiens*	GQALGIKS	SwissProt
O75688	*Homo sapiens*	GAFLDKPK	SwissProt
P29473	*Bos taurus*	GNLKSVGQ	SwissProt
O00461	*Homo sapiens*	GNGMCSRK	SwissProt
O60936	*Homo sapiens*	GNAQERPS	SwissProt
Q8WVD5	*Homo sapiens*	GQQISDQT	SwissProt
Q7LDG7-2	*Homo sapiens*	GTQRLCGR	SwissProt
P03362	*Human T-cell leukemia virus 1*	GQIFSRSA	SwissProt
Q9NZU7	*Homo sapiens*	GGGDGAAF	SwissProt
O75716	*Homo sapiens*	GHALCVCS	SwissProt
Q5JWF2	*Homo sapiens*	GVRNCLYG	SwissProt
Q9P2G1	*Homo sapiens*	GNTTTKFR	SwissProt
Q9NR22	*Homo sapiens*	GMKHSSRC	SwissProt
O95466	*Homo sapiens*	GNAAGSAE	SwissProt
Q84VQ1	*Arabidopsis thaliana*	GNANGKDE	SwissProt
Q7L9B9	*Homo sapiens*	GSTLGCHR	SwissProt
P61313	*Homo sapiens*	GAYKYIQE	SwissProt
Q9Y689	*Homo sapiens*	GILFTRIW	SwissProt
Q9LYW5	*Arabidopsis thaliana*	GNLISLIF	SwissProt
Q91J24	*Beet curly top virus*	GNLISTSC	SwissProt
Q13319	*Homo sapiens*	GTVLSLSP	SwissProt
Q9P032	*Homo sapiens*	GALVIRGI	SwissProt
Q155Q3-2	*Homo sapiens*	GGTQVKCL	SwissProt
Q7L014	*Homo sapiens*	GRESRHYR	SwissProt
Q96FZ7	*Homo sapiens*	GNLFGRKK	SwissProt
O75167	*Homo sapiens*	GQTSVSTL	SwissProt

Appendix 5 – Negative Training Set of SVMyr (232 proteins)

UniProt ID	Organism	Octapeptide	Annotation Source
F4JG06	*Arabidopsis thaliana*	GGEDDKDK	[55]
Q9M073	*Arabidopsis thaliana*	GTETVSFK	[55]
Q9ZW76	*Arabidopsis thaliana*	GTETVVHD	[55]
Q9FI46	*Arabidopsis thaliana*	GTESGSDP	[55]
Q8N4P3	*Homo sapiens*	GSEAAQLL	[55]
O22960	*Arabidopsis thaliana*	GFKLNSLF	[55]
Q9HCP6	*Homo sapiens*	GIKTALPA	[55]
Q9FS16	*Arabidopsis thaliana*	GSPMASLV	[55]
Q8L725	*Arabidopsis thaliana*	GIIRFQIL	[55]
Q8L608	*Arabidopsis thaliana*	GFNVVVFL	[55]
P22607	*Homo sapiens*	GAPACALA	[55]
Q38864	*Arabidopsis thaliana*	GVLVISLL	[55]
Q8VY22	*Arabidopsis thaliana*	GFKLISLL	[55]
P15289	*Homo sapiens*	GAPRSLLL	[55]
C0LGI2	*Arabidopsis thaliana*	GLCLAQLA	[55]
O64810	*Arabidopsis thaliana*	GSTLKHLL	[55]
Q02817	*Homo sapiens*	GLPLARLA	[55]
Q9FJA2	*Arabidopsis thaliana*	GKRATTSV	[55]
Q9LFQ7	*Arabidopsis thaliana*	GLHPVSEA	[55]
Q9Y606	*Homo sapiens*	GLQLRALL	[55]
P30181	*Arabidopsis thaliana*	GTETVSKP	[55]
Q9H944	*Homo sapiens*	GVTCVSQM	[55]
Q9SB68	*Arabidopsis thaliana*	GSAARQPL	[55]
B3H455	*Arabidopsis thaliana*	GKNNVRLQ	[55]
Q5SRH9	*Homo sapiens*	GQKGHKDS	[55]
Q8L8Y3	*Arabidopsis thaliana*	GMSNRSVS	[55]
O64632	*Arabidopsis thaliana*	GGKGKKRR	[55]
Q6NLH7	*Arabidopsis thaliana*	GKKNKRSQ	[55]
Q9LZ65	*Arabidopsis thaliana*	GAKAKKAL	[55]
Q96DA0	*Homo sapiens*	GAQGAQES	[55]
Q8TCT0	*Homo sapiens*	GATGAAEP	[55]
Q8VYF0	*Arabidopsis thaliana*	GKREKKPN	[55]
Q945N1	*Arabidopsis thaliana*	GKKTKKPG	[55]
P83916	*Homo sapiens*	GKKQNKKK	[55]
Q9FIQ3	*Arabidopsis thaliana*	GGSNKNLI	[55]
Q9NPI1	*Homo sapiens*	GKKHKKHK	[55]
O22768	*Arabidopsis thaliana*	GGLGGSGP	[55]
O23512	*Arabidopsis thaliana*	GGEGGAEP	[55]
Q9LVY1	*Arabidopsis thaliana*	GENGAKRW	[55]
Q9LIL5	*Arabidopsis thaliana*	GGKNKIEI	[55]
Q9LVS6	*Arabidopsis thaliana*	GEKGLKRS	[55]
Q9MAC6	*Arabidopsis thaliana*	GSKNKKQR	[55]
P45973	*Homo sapiens*	GKKTKRTA	[55]
Q1G3M8	*Arabidopsis thaliana*	GGKNRSHS	[55]
Q9BYN0	*Homo sapiens*	GLRAGGTL	[55]

O48679	*Arabidopsis thaliana*	GLLPCSCP	[55]
Q9LXD0	*Arabidopsis thaliana*	GDTALSLK	[55]
Q9NWV4	*Homo sapiens*	GKIALQLK	[55]
P82251	*Homo sapiens*	GDTGLRKR	[55]
F4HZI6	*Arabidopsis thaliana*	GDTALEKT	[55]
Q93VR4	*Arabidopsis thaliana*	GLSGVLHV	[55]
Q96EG1	*Homo sapiens*	GWLFLKVL	[55]
Q9LIS1	*Arabidopsis thaliana*	GLVQEEGS	[55]
Q9SD62	*Arabidopsis thaliana*	GVPCIVMR	[55]
A0A1P8B9R9	*Arabidopsis thaliana*	GLFDCRVY	[55]
Q8N468	*Homo sapiens*	GCDGRVSG	[55]
F4JBR5	*Arabidopsis thaliana*	GILDKGKQ	[55]
F4HZZ0	*Arabidopsis thaliana*	GFEDGPRC	[55]
B3H6U7	*Arabidopsis thaliana*	GSLAAREG	[55]
Q9FK13	*Arabidopsis thaliana*	GEMEIEEI	[55]
O48707	*Arabidopsis thaliana*	GEAAKDQT	[55]
F4KEM2	*Arabidopsis thaliana*	GCIGSSQA	[55]
Q9C501	*Arabidopsis thaliana*	GTRVTQFS	[55]
P46777	*Homo sapiens*	GFVKVVKN	[55]
Q9Y5F8	*Homo sapiens*	GGSCAQRR	[55]
Q8RWZ3	*Arabidopsis thaliana*	GSSTGDLV	[55]
Q86VD7	*Homo sapiens*	GNGVKEGP	[55]
Q9ASW3	*Arabidopsis thaliana*	GKKGSGGW	[55]
A7REE8	*Arabidopsis thaliana*	GIQTCSVL	[55]
Q6NQI8	*Arabidopsis thaliana*	GLLESVKS	[55]
Q9LDF2	*Arabidopsis thaliana*	GLFESVKS	[55]
P25874	*Homo sapiens*	GGLTASDV	[55]
Q9UFC0	*Homo sapiens*	GPLSARLL	[55]
Q9SRT1	*Arabidopsis thaliana*	GFSRAKRV	[55]
Q9SX77	*Arabidopsis thaliana*	GTLGRAIH	[55]
Q9C8X2	*Arabidopsis thaliana*	GALEAERA	[55]
O82253	*Arabidopsis thaliana*	GVVTVPES	[55]
P36404	*Homo sapiens*	GLLTILKK	[55]
Q9SN94	*Arabidopsis thaliana*	GDNSGRSR	[55]
Q9SIF3	*Arabidopsis thaliana*	GLPSSLES	[55]
Q9FNC2	*Arabidopsis thaliana*	GCLLGCFG	[55]
Q8L4A7	*Arabidopsis thaliana*	GDSQYSFS	[55]
Q96L21	*Homo sapiens*	GRRPARCY	[55]
Q9LSW5	*Arabidopsis thaliana*	GFFSFLGR	[55]
Q8W108	*Arabidopsis thaliana*	GEVVKDGR	[55]
O00255	*Homo sapiens*	GLKAAQKT	[55]
B3H6G6	*Arabidopsis thaliana*	GLHKHKRA	[55]
Q9SA65	*Arabidopsis thaliana*	GSSKFKRA	[55]
P08588	*Homo sapiens*	GAGVLVLG	[55]
Q96LL3	*Homo sapiens*	GAGVGVAG	[55]
Q66GS4	*Arabidopsis thaliana*	GTLVGHIL	[55]
Q9ZV87	*Arabidopsis thaliana*	GKIMEWAA	[55]

Q9C9Z0	*Arabidopsis thaliana*	GSDSTLSL	[55]
Q67XL4	*Arabidopsis thaliana*	GFLTAAIR	[55]
Q7RTY7	*Homo sapiens*	GLLASAGL	[55]
Q96P11	*Homo sapiens*	GLYAAAAG	[55]
Q9SAH5	*Arabidopsis thaliana*	GLLAAIGV	[55]
Q8IXM3	*Homo sapiens*	GVLAAAAR	[55]
Q8IWF2	*Homo sapiens*	GLSAAAPL	[55]
P59120	*Arabidopsis thaliana*	GLSKTIPL	[55]
Q13686	*Homo sapiens*	GKMAAAVG	[55]
Q9SX83	*Arabidopsis thaliana*	GKDKTLPL	[55]
Q6ZRI0	*Homo sapiens*	GVLASALC	[55]
Q8NEB5	*Homo sapiens*	GKAAAAVA	[55]
Q84JF8	*Arabidopsis thaliana*	GLFKFIFL	[55]
Q9C9F0	*Arabidopsis thaliana*	GFDFSTSK	[55]
C0LGG8	*Arabidopsis thaliana*	GFIFSTEK	[55]
Q93V51	*Arabidopsis thaliana*	GLLSNRID	[55]
Q9T065	*Arabidopsis thaliana*	GLLSKKAS	[55]
Q9M872	*Arabidopsis thaliana*	GFLSNKIS	[55]
F4I526	*Arabidopsis thaliana*	GLDSKEAD	[55]
Q84MB1	*Arabidopsis thaliana*	GVLSMKGG	[55]
Q9SMZ9	*Arabidopsis thaliana*	GLLKKKDS	[55]
Q9SV05	*Arabidopsis thaliana*	GYLSCKAG	[55]
Q9FFE7	*Arabidopsis thaliana*	GLFSHKIS	[55]
Q9LVW3	*Arabidopsis thaliana*	GVFGSNES	[55]
Q2V4J5	*Arabidopsis thaliana*	GITKTSVT	[55]
Q96BD6	*Homo sapiens*	GQKVTGGI	[55]
O15260	*Homo sapiens*	GQNDLMGT	[55]
A8MZ25	*Homo sapiens*	GQKKTMGT	[55]
Q1PEX3	*Arabidopsis thaliana*	GTYKAEDD	[55]
Q9LVC9	*Arabidopsis thaliana*	GVFSFVCK	[55]
Q8LDM2	*Arabidopsis thaliana*	GSIDAAVL	[55]
Q9C0A0	*Homo sapiens*	GSVTGAVL	[55]
P69891	*Homo sapiens*	GHFTEEDK	[55]
Q96KT6	*Homo sapiens*	GQSLQEGR	[55]
Q9LYP2	*Arabidopsis thaliana*	GFFGRLFG	[55]
B3H6B2	*Arabidopsis thaliana*	GFKGRLNV	[55]
Q8W4A6	*Arabidopsis thaliana*	GSRSRNDN	[55]
Q94A02	*Arabidopsis thaliana*	GQIPRFLS	[55]
Q38909	*Arabidopsis thaliana*	GFITRFLV	[55]
Q9LIE8	*Arabidopsis thaliana*	GFRTRNLS	[55]
Q9LIE9	*Arabidopsis thaliana*	GSRSQNLS	[55]
Q9CAJ8	*Arabidopsis thaliana*	GFPVGYSE	[55]
F4HW79	*Arabidopsis thaliana*	GSSQGSTL	[55]
Q9MA50	*Arabidopsis thaliana*	GLLPVVGI	[55]
Q8N8D7	*Homo sapiens*	GCCTGRCS	[55]
Q9BYP8	*Homo sapiens*	GCCPGDCF	[55]
Q9LZ61	*Arabidopsis thaliana*	GGVKRKIS	[55]

Q9M1G8	*Arabidopsis thaliana*	GSMAQKSV	[55]
Q7X9I0	*Arabidopsis thaliana*	GRVKLKIK	[55]
Q9C633	*Arabidopsis thaliana*	GGVKRKIA	[55]
Q9XIF2	*Arabidopsis thaliana*	GSVKRKSV	[55]
Q9S7Q7	*Arabidopsis thaliana*	GRKKLEIK	[55]
Q38837	*Arabidopsis thaliana*	GRGKVEVK	[55]
P29384	*Arabidopsis thaliana*	GRGRVELK	[55]
Q9LK30	*Arabidopsis thaliana*	GSVSLKIG	[55]
F4JM75	*Arabidopsis thaliana*	GEEKRRIS	[55]
Q9FG70	*Arabidopsis thaliana*	GLVDSLLG	[55]
Q8VY57	*Arabidopsis thaliana*	GLLEAFLN	[55]
O14773	*Homo sapiens*	GLQACLLG	[55]
Q93Y31	*Arabidopsis thaliana*	GLWDALLN	[55]
Q9SJY5	*Arabidopsis thaliana*	GLKGFAEG	[55]
Q9C998	*Arabidopsis thaliana*	GSAFDPLV	[55]
Q8LGG8	*Arabidopsis thaliana*	GSEPTKVM	[55]
Q2QAV0	*Arabidopsis thaliana*	GVEDYHVI	[55]
Q9CAB6	*Arabidopsis thaliana*	GSEEEKVV	[55]
P57764	*Homo sapiens*	GSAFERVV	[55]
Q9C8M3	*Arabidopsis thaliana*	GKENSKVV	[55]
Q9M2W3	*Arabidopsis thaliana*	GEEDTATV	[55]
F4K4Y5	*Arabidopsis thaliana*	GEEDTKVI	[55]
Q9FVQ1	*Arabidopsis thaliana*	GKSKSATK	[55]
P11717	*Homo sapiens*	GAAAGRSP	[55]
Q9NYF8	*Homo sapiens*	GRSNSRSH	[55]
Q6QPM2	*Arabidopsis thaliana*	GFSMFFSP	[55]
Q9NYS0	*Homo sapiens*	GKGCKVVV	[55]
Q9SRT7	*Arabidopsis thaliana*	GFFLCSSS	[55]
Q9ZV75	*Arabidopsis thaliana*	GAAIMRNG	[55]
Q9M1U3	*Arabidopsis thaliana*	GINSKHVV	[55]
Q9SPG6	*Arabidopsis thaliana*	GKSSSSEE	[55]
Q9LVX6	*Arabidopsis thaliana*	GISKKSQV	[55]
Q9SW40	*Arabidopsis thaliana*	GFGTSSSS	[55]
Q9FLF7	*Arabidopsis thaliana*	GSSADTET	[55]
Q1PF14	*Arabidopsis thaliana*	GSIEGQET	[55]
Q3EC50	*Arabidopsis thaliana*	GFVGSLIR	[55]
Q8LFX7	*Arabidopsis thaliana*	GFSRSLNR	[55]
Q9FX59	*Arabidopsis thaliana*	GEGKASTL	[55]
Q9SLF1	*Arabidopsis thaliana*	GFQRSISS	[55]
Q13239	*Homo sapiens*	GNSMKSTP	[55]
Q7XA06	*Arabidopsis thaliana*	GFFTSVLG	[55]
P17927	*Homo sapiens*	GASSPRSP	[55]
O82637	*Arabidopsis thaliana*	GFCFCLSS	[55]
Q9C9U3	*Arabidopsis thaliana*	GFSPSSSW	[55]
Q1G3Y4	*Arabidopsis thaliana*	GFSGKTYH	[55]
F4IAF5	*Arabidopsis thaliana*	GVMSRRVL	[55]
Q9C9A3	*Arabidopsis thaliana*	GALDSLSE	[55]

Q3E6T0	*Arabidopsis thaliana*	GKGGSLSE	[55]
Q9SKJ7	*Arabidopsis thaliana*	GKGRALSD	[55]
Q6R0A6	*Arabidopsis thaliana*	GKGRAPCC	[55]
Q8LPS4	*Arabidopsis thaliana*	GAPEKSQS	[55]
P04839	*Homo sapiens*	GNWAVNEG	[55]
Q94K91	*Arabidopsis thaliana*	GVKQALRS	[55]
Q9Y5S8	*Homo sapiens*	GNWVVNHW	[55]
P62942	*Homo sapiens*	GVQVETIS	[55]
Q9Y6V0	*Homo sapiens*	GNEASLEG	[55]
O82388	*Arabidopsis thaliana*	GVKVASSS	[55]
Q9LNM6	*Arabidopsis thaliana*	GNEAALRS	[55]
Q9ZNX9	*Arabidopsis thaliana*	GVVSISSS	[55]
Q8L731	*Arabidopsis thaliana*	GDAESTKD	[55]
F4K495	*Arabidopsis thaliana*	GGFLVLNS	[55]
Q3E8H4	*Arabidopsis thaliana*	GMSGSSGL	[55]
Q94AW9	*Arabidopsis thaliana*	GTEVSTSP	[55]
F4JCR2	*Arabidopsis thaliana*	GQKKKTSA	[55]
Q9SV13	*Arabidopsis thaliana*	GTEDYTFP	[55]
P21917	*Homo sapiens*	GNRSTADA	[55]
Q9FYR2	*Arabidopsis thaliana*	GSETMTNL	[55]
O95359	*Homo sapiens*	GNENSTSD	[55]
P93048	*Arabidopsis thaliana*	GNETKTNG	[55]
O64818	*Arabidopsis thaliana*	GGGNAQKS	[55]
A0A1P8BDG3	*Arabidopsis thaliana*	GGGEKRKS	[55]
Q9SF37	*Arabidopsis thaliana*	GFYGGGSM	[55]
Q9LW85	*Arabidopsis thaliana*	GFLIGGSC	[55]
Q9LSD2	*Arabidopsis thaliana*	GNHQADKK	[55]
Q8W4F0	*Arabidopsis thaliana*	GWLTKILK	[55]
O64586	*Arabidopsis thaliana*	GLVTKALK	[55]
Q9SJB4	*Arabidopsis thaliana*	GHLKSLFT	[55]
F4ITP1	*Arabidopsis thaliana*	GALRRRNV	[55]
Q9LQK0	*Arabidopsis thaliana*	GIIERIKE	[55]
Q9CAI1	*Arabidopsis thaliana*	GIVERIKE	[55]
A0A1P8BG44	*Arabidopsis thaliana*	GLPHTASN	[55]
Q9FKG5	*Arabidopsis thaliana*	GDGALIVA	[55]
Q0WNP8	*Arabidopsis thaliana*	GSGNLIKA	[55]
F4I699	*Arabidopsis thaliana*	GIFPGLIN	[55]
Q93YU5	*Arabidopsis thaliana*	GIFNGLPV	[55]
Q9LDR9	*Arabidopsis thaliana*	GHLGFLVM	[55]
Q93WK6	*Arabidopsis thaliana*	GSLMSGWD	[55]
Q9LHE8	*Arabidopsis thaliana*	GSLSGIIQ	[55]
Q39202	*Arabidopsis thaliana*	GSLSCSII	[55]
O65567	*Arabidopsis thaliana*	GSLRFSIP	[55]
Q7X6T3	*Arabidopsis thaliana*	GSLKLSTV	[55]
Q9ZQC6	*Arabidopsis thaliana*	GSLERSKK	[55]
F4JHZ4	*Arabidopsis thaliana*	GSLESGIP	[55]

Appendix 6 – Positive Blind Test Set of SVMyr (88 proteins)

UniProt ID	Organism	Octapeptide	Annotation Source
E9BEM4	*Leishmania donovani*	GAVPSREC	[92]
Q9LZW2	*Arabidopsis thaliana*	GNNYRFKL	[55]
Q4DPJ1	*Trypanosoma cruzi*	GAWISQLK	[91]
E9B870	*Leishmania donovani*	GAAVARVV	[92]
E9BT99	*Leishmania donovani*	GQVGGTAT	[92]
Q9LTV4	*Arabidopsis thaliana*	GNRRAPCC	[55]
Q38AS5	*Trypanosoma brucei*	GSTSSACR	[90]
F4HXI5	*Arabidopsis thaliana*	GTTLGKPF	[55]
Q38BC4	*Trypanosoma brucei*	GSCQAVCG	[90]
Q8IJW0	*Plasmodium falciparum*	GNTPGGMN	[93]
Q57ZG4	*Trypanosoma brucei*	GHCCATQT	[90]
Q96A22	*Homo sapiens*	GNRVCCGG	[55]
Q8GXV2	*Arabidopsis thaliana*	GNHCTRIP	[55]
Q8N7L0	*Homo sapiens*	GQNWKRQQ	[55]
Q940H2	*Arabidopsis thaliana*	GLVGCVGK	[55]
Q8GYJ4	*Arabidopsis thaliana*	GQAQSDEN	[55]
Q84JS7	*Arabidopsis thaliana*	GGVFVLRK	[55]
Q9FMA5	*Arabidopsis thaliana*	GAMMVMMG	[55]
Q9SW55	*Arabidopsis thaliana*	GLMRSMLP	[55]
Q585N1	*Trypanosoma brucei*	GVMLPKPV	[90]
A8MQ27	*Homo sapiens*	GNTVHRTL	[55]
E9B7A4	*Leishmania donovani*	GAALRKEA	[92]
E9BEM8	*Leishmania donovani*	GQSAPTPT	[92]
P49703	*Homo sapiens*	GNHLTEMA	[55]
Q9FLZ5	*Arabidopsis thaliana*	GLKLSRGP	[55]
Q581X5	*Trypanosoma brucei*	GCGGSAPA	[90]
E9BCZ7	*Leishmania donovani*	GQAKTKLN	[92]
O49715	*Arabidopsis thaliana*	GNCICVTE	[55]
Q9Y512	*Homo sapiens*	GTVHARSL	[55]
C0H4R4	*Plasmodium falciparum*	GNVLNRII	[93]
Q38EI8	*Trypanosoma brucei*	GNVLSWFE	[90]
O65688	*Arabidopsis thaliana*	GSLSTILR	[55]
Q8L7K7	*Arabidopsis thaliana*	GSVCCVAV	[55]
Q9FVS0	*Arabidopsis thaliana*	GCWLKQPQ	[55]
Q582H5	*Trypanosoma brucei*	GCNLSSST	[90]
Q8IY42	*Homo sapiens*	GCRCCKII	[55]
Q9S810	*Arabidopsis thaliana*	GAYRAEDD	[55]
Q4DXG4	*Trypanosoma cruzi*	GQSNGAKH	[91]
Q582S6	*Trypanosoma brucei*	GGAVVKNY	[90]
Q57U43	*Trypanosoma brucei*	GCFCCCCC	[90]
Q9H6R6-2	*Homo sapiens*	GTFCSVIK	[55]
Q9FHM7	*Arabidopsis thaliana*	GNTYCILG	[55]
Q8GXG1	*Arabidopsis thaliana*	GGWAIAVH	[55]
Q8IKM6	*Plasmodium falciparum*	GNLCCSNN	[93]
Q9ZWJ3	*Arabidopsis thaliana*	GSHVAQKQ	[55]

O75838-2	*Homo sapiens*	GNKQTIFT	[55]
Q9XIQ4	*Arabidopsis thaliana*	GQKIHAFM	[55]
C6S3C8	*Plasmodium falciparum*	GAGQTKEI	[93]
Q9SF49	*Arabidopsis thaliana*	GNLHGIHR	[55]
E9BAI0	*Leishmania donovani*	GLLNTKPC	[92]
Q99487	*Homo sapiens*	GVNQSVGF	[55]
Q9UJT9	*Homo sapiens*	GANNGKQY	[55]
Q8GWT2	*Arabidopsis thaliana*	GGVQCYHL	[55]
Q38BS1	*Trypanosoma brucei*	GGCVASLI	[90]
Q9STT7	*Arabidopsis thaliana*	GNHVPAGF	[55]
F4JLT3	*Arabidopsis thaliana*	GNCIHTLL	[55]
Q38BV2	*Trypanosoma brucei*	GNCLCCRD	[90]
Q386D8	*Trypanosoma brucei*	GQAGGKEQ	[90]
Q8IVV8	*Homo sapiens*	GSCSGRCA	[55]
Q9LIF6	*Arabidopsis thaliana*	GQQLRRAV	[55]
F4KH94	*Arabidopsis thaliana*	GNVQDIMK	[55]
Q9ZV66	*Arabidopsis thaliana*	GAFCKLID	[55]
Q9S7U2	*Arabidopsis thaliana*	GIVTTKTK	[55]
E9BG73	*Leishmania donovani*	GQPNTKDS	[92]
Q9SJ61	*Arabidopsis thaliana*	GNVCVHMV	[55]
Q38D65	*Trypanosoma brucei*	GQWLASAF	[90]
F4JS23	*Arabidopsis thaliana*	GSSMGFLG	[55]
Q4CZT4	*Trypanosoma cruzi*	GCTNTKEK	[91]
Q38EE5	*Trypanosoma brucei*	GSDLSIVL	[90]
Q9C9I9	*Arabidopsis thaliana*	GCICATAR	[55]
Q4D708	*Trypanosoma cruzi*	GQLLSFNA	[91]
O04331	*Arabidopsis thaliana*	GSQQAAVS	[55]
C0H4A5	*Plasmodium falciparum*	GNNCCAGR	[93]
Q9NS25	*Homo sapiens*	GQQSSVRR	[55]
Q4DZM9	*Trypanosoma cruzi*	GNLVARLR	[91]
Q4GY77	*Trypanosoma brucei*	GGVVGKIP	[90]
Q38DK7	*Trypanosoma brucei*	GASEAKGE	[90]
Q38EM1	*Trypanosoma brucei*	GQLISGLW	[90]
E9BAH9	*Leishmania donovani*	GSNASHTE	[92]
Q94C32	*Arabidopsis thaliana*	GSSSKEET	[55]
Q384A3	*Trypanosoma brucei*	GCQQSGVR	[90]
E9BIF0	*Leishmania donovani*	GAGGVSPQ	[92]
Q3ECI5	*Arabidopsis thaliana*	GNCMERWM	[55]
A0PJX0	*Homo sapiens*	GQCLRYQM	[55]
P0CG00	*Homo sapiens*	GQCRNWKW	[55]
Q9FKK9	*Arabidopsis thaliana*	GSINSVAE	[55]
E9BBH3	*Leishmania donovani*	GQNMPKPP	[92]
P0C7M6	*Homo sapiens*	GSKCCKGG	[55]

Appendix 7 – Negative Blind Testing Set of SVMyr (528 proteins)

UniProt ID	Organism	Octapeptide	Annotation Source
A0A0G2JMR5	*Homo sapiens*	GLSLPKEK	[55]
A0A1I9LT31	*Arabidopsis thaliana*	GEHESWAA	[55]
A0A1P8APA8	*Arabidopsis thaliana*	GAEEFPSV	[55]
A0A1P8ATH4	*Arabidopsis thaliana*	GIVQIGHC	[55]
A0A1P8AYW1	*Arabidopsis thaliana*	GSERASNN	[55]
A0A1P8B4T0	*Arabidopsis thaliana*	GGDTFKDD	[55]
A0A1P8B7F8	*Arabidopsis thaliana*	GDVILFID	[55]
A0A1P8BHQ5	*Arabidopsis thaliana*	GEMTARSS	[55]
A0A3Q5AD24	*Homo sapiens*	GEAFYTVK	[55]
A0JJX5	*Arabidopsis thaliana*	GFLFGLFI	[55]
A4D2G3	*Homo sapiens*	GGNQTSIT	[55]
A6NMZ2	*Homo sapiens*	GGCMHSTQ	[55]
A8MR40	*Arabidopsis thaliana*	GVDYYKVL	[55]
A8MRI5	*Arabidopsis thaliana*	GDSFIRPH	[55]
A8MSF6	*Arabidopsis thaliana*	GVKRAPNM	[55]
A8MU10	*Homo sapiens*	GSIPSKPC	[55]
B3H4C4	*Arabidopsis thaliana*	GGGSVPPP	[55]
B3H4F0	*Arabidopsis thaliana*	GVFRGLMG	[55]
B3H4H8	*Arabidopsis thaliana*	GGMCMSAC	[55]
B3H5J9	*Arabidopsis thaliana*	GFDLCPQR	[55]
B3H6A6	*Arabidopsis thaliana*	GKNHHPLG	[55]
B9DGF6	*Arabidopsis thaliana*	GSSCLACF	[55]
C0SUT9	*Arabidopsis thaliana*	GTELMRIC	[55]
F4HQM5	*Arabidopsis thaliana*	GLLQLKSY	[55]
F4HS31	*Arabidopsis thaliana*	GSVNVPAG	[55]
F4HUM4	*Arabidopsis thaliana*	GRGKNQPT	[55]
F4HVS0	*Arabidopsis thaliana*	GGGNLHSL	[55]
F4HW02	*Arabidopsis thaliana*	GGEERSGD	[55]
F4I2G0	*Arabidopsis thaliana*	GFTFTKIY	[55]
F4I2J8	*Arabidopsis thaliana*	GSHGKGKR	[55]
F4I421	*Arabidopsis thaliana*	GRKEPSSR	[55]
F4I679	*Arabidopsis thaliana*	GVFPGFGS	[55]
F4IDB2	*Arabidopsis thaliana*	GVSFKISK	[55]
F4IEY4	*Arabidopsis thaliana*	GNQKLKWT	[55]
F4IFM9	*Arabidopsis thaliana*	GMINPYVQ	[55]
F4IHS9	*Arabidopsis thaliana*	GEMKSMQM	[55]
F4II36	*Arabidopsis thaliana*	GRRKQSKP	[55]
F4II93	*Arabidopsis thaliana*	GVDGKLKS	[55]
F4IIZ5	*Arabidopsis thaliana*	GINEFSSF	[55]
F4IK44	*Arabidopsis thaliana*	GSEERMMI	[55]
F4IMS7	*Arabidopsis thaliana*	GNGSLYLW	[55]
F4INY4	*Arabidopsis thaliana*	GNKRFRSD	[55]
F4IPY7	*Arabidopsis thaliana*	GPFGMETT	[55]
F4IRU6	*Arabidopsis thaliana*	GSEVVNPT	[55]
F4ITQ2	*Arabidopsis thaliana*	GALQLMEN	[55]

F4IUT0	*Arabidopsis thaliana*	GIADESKC	[55]
F4IUU9	*Arabidopsis thaliana*	GACNASQI	[55]
F4IVV8	*Arabidopsis thaliana*	GTIDFRAA	[55]
F4IXX4	*Arabidopsis thaliana*	GIIEEGTI	[55]
F4IZP3	*Arabidopsis thaliana*	GTQNGLSD	[55]
F4J027	*Arabidopsis thaliana*	GSCLACFD	[55]
F4J061	*Arabidopsis thaliana*	GASGRWIK	[55]
F4J394	*Arabidopsis thaliana*	GIGEDQMQ	[55]
F4J420	*Arabidopsis thaliana*	GTLWFGDF	[55]
F4J7Y0	*Arabidopsis thaliana*	GFYSKSIR	[55]
F4JBG1	*Arabidopsis thaliana*	GNYRFKDP	[55]
F4JBM4	*Arabidopsis thaliana*	GGLDVKKV	[55]
F4JDF8	*Arabidopsis thaliana*	GSRGNRVG	[55]
F4JG55	*Arabidopsis thaliana*	GGTRHCYG	[55]
F4JGJ7	*Arabidopsis thaliana*	GVNVSGAV	[55]
F4JL85	*Arabidopsis thaliana*	GLVMRFDL	[55]
F4JTL3	*Arabidopsis thaliana*	GLALFSSD	[55]
F4JWE4	*Arabidopsis thaliana*	GDTHDFTD	[55]
F4JZA9	*Arabidopsis thaliana*	GKSMVRFA	[55]
F4K2E9	*Arabidopsis thaliana*	GVDPFKTT	[55]
F4K5T1	*Arabidopsis thaliana*	GAARGYKV	[55]
F4K753	*Arabidopsis thaliana*	GGNCRGPS	[55]
F4K8P3	*Arabidopsis thaliana*	GANRSIWN	[55]
F4KCH7	*Arabidopsis thaliana*	GTKQPRNY	[55]
F4KD38	*Arabidopsis thaliana*	GLDQEDLD	[55]
F4KG57	*Arabidopsis thaliana*	GVAQAMEA	[55]
F4KHB6	*Arabidopsis thaliana*	GEYCNEDL	[55]
F4KJ98	*Arabidopsis thaliana*	GKHLFRSS	[55]
O04087	*Arabidopsis thaliana*	GTPRSPAT	[55]
O04551	*Arabidopsis thaliana*	GSLDLPYA	[55]
O14668	*Homo sapiens*	GRVFLTGE	[55]
O14949	*Homo sapiens*	GREFGNLT	[55]
O14972	*Homo sapiens*	GTALDIKI	[55]
O15427	*Homo sapiens*	GGAVVDEG	[55]
O23010	*Arabidopsis thaliana*	GTKARRPN	[55]
O23203	*Arabidopsis thaliana*	GHARTRTG	[55]
O23487	*Arabidopsis thaliana*	GQKFWENQ	[55]
O23515	*Arabidopsis thaliana*	GAYKYVSE	[55]
O23550	*Arabidopsis thaliana*	GFAPVTPA	[55]
O23661	*Arabidopsis thaliana*	GGLIDLNV	[55]
O49389	*Arabidopsis thaliana*	GVAVLNPQ	[55]
O60423	*Homo sapiens*	GTGPAQTP	[55]
O60674	*Homo sapiens*	GMACLTMT	[55]
O64760	*Arabidopsis thaliana*	GAQKKGGA	[55]
O65251	*Arabidopsis thaliana*	GIKGLTKL	[55]
O65555	*Arabidopsis thaliana*	GLSNDRIK	[55]
O65583	*Arabidopsis thaliana*	GSKSVVDM	[55]

O65607	*Arabidopsis thaliana*	GKQKQQTI	[55]
O65657	*Arabidopsis thaliana*	GVGGTLEY	[55]
O75593	*Homo sapiens*	GPCSGSRL	[55]
O75844	*Homo sapiens*	GMWASLDA	[55]
O80437	*Arabidopsis thaliana*	GAQEKRRR	[55]
O80738	*Arabidopsis thaliana*	GSKSFGNL	[55]
O80774	*Arabidopsis thaliana*	GGGFRVLH	[55]
O80845	*Arabidopsis thaliana*	GTTLDVSR	[55]
O80924	*Arabidopsis thaliana*	GIYGVMTG	[55]
O81024	*Arabidopsis thaliana*	GAAKNIWA	[55]
O81270	*Arabidopsis thaliana*	GSKTEMME	SwissProt (acetyl-Gly)
O82286	*Arabidopsis thaliana*	GLDSSFVN	[55]
O82393	*Arabidopsis thaliana*	GKFDAKDV	[55]
O82785	*Arabidopsis thaliana*	GSDQCFSR	[55]
O95136	*Homo sapiens*	GSLYSEYL	[55]
O95159	*Homo sapiens*	GLCKCPKR	[55]
O95810	*Homo sapiens*	GEDAAQAE	SwissProt (acetyl-Gly)
P00017	*Aptenodytes patagonicus*	GDIEKGKK	SwissProt (acetyl-Gly)
P02643	*Oryctolagus cuniculus*	GDEEKRNR	SwissProt (acetyl-Gly)
P04175	*Sus scrofa*	GDSNVDTG	SwissProt (acetyl-Gly)
P05023	*Homo sapiens*	GKGVGRDK	[55]
P05161	*Homo sapiens*	GWDLTVKM	[55]
P06485	*Human herpesvirus 1*	GVVVVNVM	SwissProt (acetyl-Gly)
P08708	*Homo sapiens*	GRVRTKTV	[55]
P0C227	*Nerita albicilla*	GDVDVLKS	SwissProt (acetyl-Gly)
P0C883	*Arabidopsis thaliana*	GSYSAGFP	[55]
P11574	*Arabidopsis thaliana*	GTNDLDIE	[55]
P12235	*Homo sapiens*	GDHAWSFL	SwissProt (acetyl-Gly)
P25405	*Saara hardwickii*	GTAGKVIK	SwissProt (acetyl-Gly)
P26583	*Homo sapiens*	GKGDPNKP	[55]
P27701	*Homo sapiens*	GSACIKVT	[55]
P28329	*Homo sapiens*	GLRTAKKR	[55]
P30825	*Homo sapiens*	GCKVLLNI	[55]
P31006	*Sus scrofa*	GSPRPVVL	SwissProt (acetyl-Gly)
P35658	*Homo sapiens*	GDEMDAMI	[55]
P38405	*Homo sapiens*	GCLGGNSK	[55]
P42776	*Arabidopsis thaliana*	GNSSEEPK	[55]
P42791	*Arabidopsis thaliana*	GIDLIAGG	[55]
P43116	*Homo sapiens*	GNASNDSQ	[55]
P46059	*Homo sapiens*	GMSKSHSF	[55]
P46093	*Homo sapiens*	GNHTWEGC	[55]
P46313	*Arabidopsis thaliana*	GAGGRMPV	[55]
P46604	*Arabidopsis thaliana*	GLDDSCNT	[55]
P48523	*Arabidopsis thaliana*	GSVEAGEK	[55]
P49689	*Arabidopsis thaliana*	GKVHGSLA	[55]
P50570	*Homo sapiens*	GNRGMEEL	[55]
P50651	*Arabidopsis thaliana*	GSLKEGQG	[55]

P50993	*Homo sapiens*	GRGAGREY	[55]
P54577	*Homo sapiens*	GDAPSPEE	SwissProt (acetyl-Gly)
P56749	*Homo sapiens*	GCRDVHAA	[55]
P56774	*Arabidopsis thaliana*	GVTKKPDL	[55]
P56798	*Arabidopsis thaliana*	GQKINPLG	[55]
P56801	*Arabidopsis thaliana*	GKDTIADI	[55]
P59223	*Arabidopsis thaliana*	GRMHSRGK	[55]
P59817	*Homo sapiens*	GDIFLCKK	[55]
P61353	*Homo sapiens*	GKFMKPGK	[55]
P62266	*Homo sapiens*	GKCRGLRT	[55]
P62491	*Homo sapiens*	GTRDDEYD	SwissProt (acetyl-Gly)
P63092	*Homo sapiens*	GCLGNSKT	[55]
P69891	*Homo sapiens*	GHFTEEDK	SwissProt (acetyl-Gly)
P80017	*Molpadia arenicola*	GATQSFQS	SwissProt (acetyl-Gly)
P80018	*Molpadia arenicola*	GGTLAIQA	SwissProt (acetyl-Gly)
P81536	*Byssochlamys spectabilis*	GTTPNSEG	SwissProt (acetyl-Gly)
P92518	*Arabidopsis thaliana*	GLSTHCQL	[55]
P93834	*Arabidopsis thaliana*	GKVAVATT	[55]
Q02972	*Arabidopsis thaliana*	GKVLQKEA	[55]
Q04917	*Homo sapiens*	GDREQLLQ	SwissProt (acetyl-Gly)
Q058K9	*Arabidopsis thaliana*	GMEEGIKD	[55]
Q08211	*Homo sapiens*	GDVKNFLY	[55]
Q0WL56	*Arabidopsis thaliana*	GKEKFHIN	[55]
Q0WML0	*Arabidopsis thaliana*	GNKKLLTG	[55]
Q0WPZ7	*Arabidopsis thaliana*	GIVLEPPC	[55]
Q0WQY3	*Arabidopsis thaliana*	GFTLVFTG	[55]
Q0WRB2	*Arabidopsis thaliana*	GHFSSMFN	[55]
Q13427	*Homo sapiens*	GIKVQRPR	[55]
Q14108	*Homo sapiens*	GRCCFYTA	[55]
Q14439	*Homo sapiens*	GHNGSWIS	[55]
Q14683	*Homo sapiens*	GFLKLIEI	[55]
Q15743	*Homo sapiens*	GNITADNS	[55]
Q15907	*Homo sapiens*	GTRDDEYD	SwissProt (acetyl-Gly)
Q15910	*Homo sapiens*	GQTGKKSE	[55]
Q16678	*Homo sapiens*	GTSLSPND	[55]
Q16881	*Homo sapiens*	GCAEGKAV	[55]
Q1HDT3	*Arabidopsis thaliana*	GTLVNGTI	[55]
Q1PFN9	*Arabidopsis thaliana*	GFGGFNGD	[55]
Q2HIW3	*Arabidopsis thaliana*	GPMMMRAE	[55]
Q2TAA8	*Homo sapiens*	GGHLSPWP	[55]
Q2V323	*Arabidopsis thaliana*	GWFIKERR	[55]
Q2V3B2	*Arabidopsis thaliana*	GEPKDSLA	[55]
Q38967	*Arabidopsis thaliana*	GETAAANN	[55]
Q39216	*Arabidopsis thaliana*	GTNEVTRI	[55]
Q39232	*Arabidopsis thaliana*	GAYETEKP	[55]
Q3B7T1	*Homo sapiens*	GDAKEAGA	[55]
Q3E7U8	*Arabidopsis thaliana*	GARRSSHH	[55]

Q3E8U4	*Arabidopsis thaliana*	GKDGQDWA	[55]
Q3E8X7	*Arabidopsis thaliana*	GRVHAECD	[55]
Q3ECR5	*Arabidopsis thaliana*	GVANLRVM	[55]
Q3ED65	*Arabidopsis thaliana*	GLDFSSEQ	[55]
Q43383	*Arabidopsis thaliana*	GHDSFCYL	[55]
Q4PSL7	*Arabidopsis thaliana*	GRVIRAQR	[55]
Q4V3E2	*Arabidopsis thaliana*	GSPNAAAE	[55]
Q501D5	*Arabidopsis thaliana*	GGPAYDCL	[55]
Q52LD8	*Homo sapiens*	GCGLRKLE	[55]
Q56W59	*Arabidopsis thaliana*	GSRDFISS	[55]
Q56XX3	*Arabidopsis thaliana*	GARVQVQH	[55]
Q56YU8	*Arabidopsis thaliana*	GEQSPSQP	[55]
Q58FY9	*Arabidopsis thaliana*	GDDLPDWR	[55]
Q5BJF2	*Homo sapiens*	GAPATRRC	[55]
Q5BPZ5	*Arabidopsis thaliana*	GVTETSTY	[55]
Q5EAI9	*Arabidopsis thaliana*	GQRNRNVD	[55]
Q5JWF2	*Homo sapiens*	GVRNCLYG	[55]
Q5PNY6	*Arabidopsis thaliana*	GFLWRTRS	[55]
Q5PP38	*Arabidopsis thaliana*	GKQGPCYH	[55]
Q5XF36	*Arabidopsis thaliana*	GSAGVASS	[55]
Q5XKR9	*Homo sapiens*	GGCPVRKR	[55]
Q5XV54	*Arabidopsis thaliana*	GHSILEKM	[55]
Q5XVI1	*Arabidopsis thaliana*	GLNLNPIL	[55]
Q66GK1	*Arabidopsis thaliana*	GSSFNAQI	[55]
Q67XC4	*Arabidopsis thaliana*	GLCFQLNL	[55]
Q67XT3	*Arabidopsis thaliana*	GNASENFD	[55]
Q67Z75	*Arabidopsis thaliana*	GSYTVWSC	[55]
Q67ZB6	*Arabidopsis thaliana*	GIQIIGQI	[55]
Q67ZW1	*Arabidopsis thaliana*	GFRDICYR	[55]
Q67ZZ1	*Arabidopsis thaliana*	GEELQYQQ	[55]
Q680P8	*Arabidopsis thaliana*	GHSNVWNS	[55]
Q682H0	*Arabidopsis thaliana*	GFIIAIAK	[55]
Q6DCA0	*Homo sapiens*	GKRRCVPP	[55]
Q6DR24	*Arabidopsis thaliana*	GFSFTATM	[55]
Q6GKW1	*Arabidopsis thaliana*	GTRQVYEE	[55]
Q6I9Y2	*Homo sapiens*	GAVTDDEV	SwissProt (acetyl-Gly)
Q6NMR8	*Arabidopsis thaliana*	GTVVYQQG	[55]
Q6NQN5	*Arabidopsis thaliana*	GDKLRLSI	[55]
Q6NVV3	*Homo sapiens*	GAQVRLPP	[55]
Q6WQI6	*Homo sapiens*	GNWGLGIA	[55]
Q6XR72	*Homo sapiens*	GRYSGKTC	[55]
Q6ZMN7	*Homo sapiens*	GFALERFA	[55]
Q6ZRP0	*Homo sapiens*	GSRPCSPS	[55]
Q76G19	*Homo sapiens*	GCNMCVVQ	[55]
Q7RTT9	*Homo sapiens*	GSVGSQRL	[55]
Q7X9H2	*Arabidopsis thaliana*	GMKKVKLS	[55]
Q7XJJ7	*Arabidopsis thaliana*	GKYQVMKR	[55]

Q7Y227	*Arabidopsis thaliana*	GEIQERLS	[55]
Q7Y229	*Arabidopsis thaliana*	GGDLKSQL	[55]
Q7Z7L8	*Homo sapiens*	GNKQPQKV	[55]
Q84M24	*Arabidopsis thaliana*	GSSKRQFK	[55]
Q84RJ7	*Arabidopsis thaliana*	GGVEGNQW	[55]
Q84VV1	*Arabidopsis thaliana*	GKRGPKKL	[55]
Q84WW3	*Arabidopsis thaliana*	GVEEGAGV	[55]
Q86UQ4	*Homo sapiens*	GHAGCQFK	[55]
Q86VF5	*Homo sapiens*	GVATTLQP	[55]
Q86YM7	*Homo sapiens*	GEQPIFST	SwissProt (acetyl-Gly)
Q8GUQ8	*Arabidopsis thaliana*	GSLKKDGE	[55]
Q8GWT5	*Arabidopsis thaliana*	GEVWTWII	[55]
Q8GWV0	*Arabidopsis thaliana*	GCLISPVM	[55]
Q8GX45	*Arabidopsis thaliana*	GSEGRSIA	[55]
Q8GXG9	*Arabidopsis thaliana*	GFTKDQLL	[55]
Q8GXI1	*Arabidopsis thaliana*	GSFHRRTF	[55]
Q8GXX0	*Arabidopsis thaliana*	GEKPWQPL	[55]
Q8GYJ3	*Arabidopsis thaliana*	GKYIRKSK	[55]
Q8GYP8	*Arabidopsis thaliana*	GFALVLIF	[55]
Q8IY57	*Homo sapiens*	GDKKSPTR	[55]
Q8L706	*Arabidopsis thaliana*	GFIVGVVI	[55]
Q8L765	*Arabidopsis thaliana*	GTTRVCSE	[55]
Q8L783	*Arabidopsis thaliana*	GCCKVPAL	[55]
Q8L7H2	*Arabidopsis thaliana*	GQNFPNGL	[55]
Q8L8M9	*Arabidopsis thaliana*	GVTGGLVR	[55]
Q8L9Y4	*Arabidopsis thaliana*	GSRGIIND	[55]
Q8LBH2	*Arabidopsis thaliana*	GTLQSWRK	[55]
Q8LBW2	*Arabidopsis thaliana*	GYEPDPDA	[55]
Q8LBW3	*Arabidopsis thaliana*	GGPGSSPC	[55]
Q8LDQ4	*Arabidopsis thaliana*	GGLAMEEM	[55]
Q8LDZ6	*Arabidopsis thaliana*	GISTNHTT	[55]
Q8LFH7	*Arabidopsis thaliana*	GKGTGSFG	[55]
Q8LFU8	*Arabidopsis thaliana*	GAIEKEGY	[55]
Q8LFZ9	*Arabidopsis thaliana*	GSGRDRDD	[55]
Q8LGG0	*Arabidopsis thaliana*	GVEKQVIR	[55]
Q8LGJ5	*Arabidopsis thaliana*	GTDTVMSG	[55]
Q8LPT3	*Arabidopsis thaliana*	GSAAELTE	[55]
Q8N7H1	*Homo sapiens*	GGKSAVRH	[55]
Q8N813	*Homo sapiens*	GTGASEKQ	[55]
Q8NB46	*Homo sapiens*	GILSITDQ	[55]
Q8NGB4	*Homo sapiens*	GAKNNVTE	[55]
Q8NGC5	*Homo sapiens*	GNWTAAVT	[55]
Q8NGV0	*Homo sapiens*	GSFNTSFE	[55]
Q8NGY1	*Homo sapiens*	GQTNVTSW	[55]
Q8NH50	*Homo sapiens*	GQHNLTVL	[55]
Q8RW97	*Arabidopsis thaliana*	GSEPRFEP	[55]
Q8RWY1	*Arabidopsis thaliana*	GWPWADHW	[55]

Q8RX22	*Arabidopsis thaliana*	GCTVREKH	[55]
Q8RY74	*Arabidopsis thaliana*	GISQVHYC	[55]
Q8S8Q9	*Arabidopsis thaliana*	GEEEENPN	[55]
Q8VY23	*Arabidopsis thaliana*	GVSLLKQQ	[55]
Q8VZ42	*Arabidopsis thaliana*	GTPEFPDL	[55]
Q8VZE9	*Arabidopsis thaliana*	GTPVEVSK	[55]
Q8VZM1	*Arabidopsis thaliana*	GASLPPKE	[55]
Q8VZT9	*Arabidopsis thaliana*	GGGDHGHG	[55]
Q8VZW3	*Arabidopsis thaliana*	GTEMVMVH	[55]
Q8W1D5	*Arabidopsis thaliana*	GSKLKLYP	[55]
Q8W4Q5	*Arabidopsis thaliana*	GITYLHIS	[55]
Q8WTV0	*Homo sapiens*	GCSAKARW	[55]
Q8WTW4	*Homo sapiens*	GSGCRIEC	[55]
Q8WUN7	*Homo sapiens*	GGCVGAQH	[55]
Q8WXX5	*Homo sapiens*	GLLDLCEE	[55]
Q8WY22	*Homo sapiens*	GARASGGP	[55]
Q92620	*Homo sapiens*	GDTSEDAS	SwissProt (acetyl-Gly)
Q92625	*Homo sapiens*	GKEQELLE	SwissProt (acetyl-Gly)
Q92989	*Homo sapiens*	GEEANDDK	[55]
Q93V61	*Arabidopsis thaliana*	GWIPCPCW	[55]
Q93V70	*Arabidopsis thaliana*	GPMIRTEE	[55]
Q93VR3	*Arabidopsis thaliana*	GTTNGTDY	[55]
Q93YN1	*Arabidopsis thaliana*	GCSWLSCH	[55]
Q93ZR2	*Arabidopsis thaliana*	GTMHRSGA	[55]
Q940B8	*Arabidopsis thaliana*	GGQMQQNN	[55]
Q941D7	*Arabidopsis thaliana*	GSETFLEI	[55]
Q945M9	*Arabidopsis thaliana*	GIEDMHSK	[55]
Q94AC1	*Arabidopsis thaliana*	GTSSCGDH	[55]
Q94AK0	*Arabidopsis thaliana*	GIKRAKAS	[55]
Q94AX9	*Arabidopsis thaliana*	GTVVGTVE	[55]
Q94BY4	*Arabidopsis thaliana*	GHHHDGGD	[55]
Q94CG0	*Arabidopsis thaliana*	GLFGTKKI	[55]
Q94CK9	*Arabidopsis thaliana*	GMVGLRDV	[55]
Q94F30	*Arabidopsis thaliana*	GAVAINRK	[55]
Q94F37	*Arabidopsis thaliana*	GIFSRSSI	[55]
Q94KL5	*Arabidopsis thaliana*	GLATTTSS	[55]
Q96AA3	*Homo sapiens*	GSQEVLGH	[55]
Q96AX9	*Homo sapiens*	GWKPSEAR	[55]
Q96BI3	*Homo sapiens*	GAAVFFGC	[55]
Q96CE8	*Homo sapiens*	GSRKCGGC	[55]
Q96EH8	*Homo sapiens*	GAQLCFEA	[55]
Q96LK8	*Homo sapiens*	GVTGAHGF	[55]
Q96MF4	*Homo sapiens*	GDECSNPD	[55]
Q96P15	*Homo sapiens*	GSLSTANV	[55]
Q9BQY9	*Homo sapiens*	GAGNFLTA	[55]
Q9BRQ6	*Homo sapiens*	GSTESSEG	[55]
Q9BSY9	*Homo sapiens*	GANQLVVL	[55]

Q9BTY7	*Homo sapiens*	GEAGAGAG	SwissProt (acetyl-Gly)
Q9BY42	*Homo sapiens*	GCDGGTIP	[55]
Q9C0I3	*Homo sapiens*	GDSGSRRS	[55]
Q9C4Z8	*Arabidopsis thaliana*	GKDDHHEQ	[55]
Q9C566	*Arabidopsis thaliana*	GRSKCFMD	[55]
Q9C590	*Arabidopsis thaliana*	GGLRCWLQ	[55]
Q9C5G6	*Arabidopsis thaliana*	GSEGPKAI	[55]
Q9C5Q2	*Arabidopsis thaliana*	GSIRGNIE	[55]
Q9C5T3	*Arabidopsis thaliana*	GSFDRQRA	[55]
Q9C760	*Arabidopsis thaliana*	GSDIVADG	[55]
Q9C7G0	*Arabidopsis thaliana*	GSLQTPIE	[55]
Q9C829	*Arabidopsis thaliana*	GDSENVQQ	[55]
Q9C8H1	*Arabidopsis thaliana*	GFEALNWY	[55]
Q9C969	*Arabidopsis thaliana*	GSFGMLSR	[55]
Q9C9N8	*Arabidopsis thaliana*	GNPSVNDL	[55]
Q9C9P3	*Arabidopsis thaliana*	GSSMEEKV	[55]
Q9C9Z7	*Arabidopsis thaliana*	GIISDNAQ	[55]
Q9CA75	*Arabidopsis thaliana*	GQDGSPAH	[55]
Q9CAL6	*Arabidopsis thaliana*	GNPGSDTE	[55]
Q9CAL7	*Arabidopsis thaliana*	GDQPQEFQ	[55]
Q9CAS6	*Arabidopsis thaliana*	GLMNRSKN	[55]
Q9FE29	*Arabidopsis thaliana*	GYETKSTL	[55]
Q9FE70	*Arabidopsis thaliana*	GSFLEVLC	[55]
Q9FFY4	*Arabidopsis thaliana*	GEMMYKLF	[55]
Q9FG59	*Arabidopsis thaliana*	GIEVCVKA	[55]
Q9FGB0	*Arabidopsis thaliana*	GQDRGFGF	[55]
Q9FGC6	*Arabidopsis thaliana*	GSLHLNSN	[55]
Q9FGF4	*Arabidopsis thaliana*	GSKKRSND	[55]
Q9FGJ3	*Arabidopsis thaliana*	GDSDRDSG	[55]
Q9FGJ9	*Arabidopsis thaliana*	GILGCDAH	[55]
Q9FIB6	*Arabidopsis thaliana*	GDSGKLEA	[55]
Q9FIF3	*Arabidopsis thaliana*	GISRDSIH	[55]
Q9FIK6	*Arabidopsis thaliana*	GIVSEEAI	[55]
Q9FIK8	*Arabidopsis thaliana*	GEPLGLLQ	[55]
Q9FIV0	*Arabidopsis thaliana*	GGGRAMAT	[55]
Q9FIZ7	*Arabidopsis thaliana*	GTVIEGKL	[55]
Q9FK15	*Arabidopsis thaliana*	GSSKDSAS	[55]
Q9FKM2	*Arabidopsis thaliana*	GSSPAPFA	[55]
Q9FKR9	*Arabidopsis thaliana*	GHQSSWMK	[55]
Q9FLE4	*Arabidopsis thaliana*	GSMYRASK	[55]
Q9FLH8	*Arabidopsis thaliana*	GEDAISGN	[55]
Q9FLM0	*Arabidopsis thaliana*	GPTYRALP	[55]
Q9FLN5	*Arabidopsis thaliana*	GFLITTLI	[55]
Q9FLT9	*Arabidopsis thaliana*	GKDGEGDK	[55]
Q9FN26	*Arabidopsis thaliana*	GSKFHAFM	[55]
Q9FNG0	*Arabidopsis thaliana*	GRGSLRKL	[55]
Q9FNK2	*Arabidopsis thaliana*	GPYLGPMR	[55]

Q9FNN9	*Arabidopsis thaliana*	GPLRQFVQ	[55]
Q9FPD5	*Arabidopsis thaliana*	GDKNKDDS	[55]
Q9FPS2	*Arabidopsis thaliana*	GFKLQMSW	[55]
Q9FQ04	*Arabidopsis thaliana*	GVPAFYRW	[55]
Q9FT92	*Arabidopsis thaliana*	GDITWVEE	[55]
Q9FUG4	*Arabidopsis thaliana*	GNDERKRP	[55]
Q9FXB0	*Arabidopsis thaliana*	GLVTDEVR	[55]
Q9FY94	*Arabidopsis thaliana*	GVMINHHF	[55]
Q9FYM0	*Arabidopsis thaliana*	GLEITVTS	[55]
Q9FZ45	*Arabidopsis thaliana*	GSRYPSHQ	[55]
Q9FZ93	*Arabidopsis thaliana*	GHDNITKL	[55]
Q9GZU0	*Homo sapiens*	GDPNSRKK	[55]
Q9H295	*Homo sapiens*	GIWTSGTD	[55]
Q9H340	*Homo sapiens*	GLNKSAST	[55]
Q9HCS5	*Homo sapiens*	GCFCAVPE	[55]
Q9LDD4	*Arabidopsis thaliana*	GSFNDTSC	[55]
Q9LDQ1	*Arabidopsis thaliana*	GMDIADKE	[55]
Q9LE63	*Arabidopsis thaliana*	GRSPCCDK	[55]
Q9LF22	*Arabidopsis thaliana*	GKARGVNS	[55]
Q9LF59	*Arabidopsis thaliana*	GNDQHNHS	[55]
Q9LFA2	*Arabidopsis thaliana*	GSFAGACE	[55]
Q9LFL3	*Arabidopsis thaliana*	GDNLMDKV	[55]
Q9LFM5	*Arabidopsis thaliana*	GTCRESEP	[55]
Q9LFP7	*Arabidopsis thaliana*	GLDAVKAK	[55]
Q9LFR9	*Arabidopsis thaliana*	GSYVEQAR	[55]
Q9LJ47	*Arabidopsis thaliana*	GRWVRPEV	[55]
Q9LJZ5	*Arabidopsis thaliana*	GLMMGADP	[55]
Q9LK23	*Arabidopsis thaliana*	GSGQWHME	[55]
Q9LMF1	*Arabidopsis thaliana*	GSRFVSNE	[55]
Q9LMG9	*Arabidopsis thaliana*	GFKRTFDA	[55]
Q9LMM2	*Arabidopsis thaliana*	GVGEMNKE	[55]
Q9LMZ9	*Arabidopsis thaliana*	GSTDEPGS	[55]
Q9LND0	*Arabidopsis thaliana*	GGGGMFEE	[55]
Q9LNJ7	*Arabidopsis thaliana*	GAAEARAL	[55]
Q9LPC2	*Arabidopsis thaliana*	GSLVKAYY	[55]
Q9LPC4	*Arabidopsis thaliana*	GIYSCSAV	[55]
Q9LPH1	*Arabidopsis thaliana*	GKEKDKNR	[55]
Q9LPV9	*Arabidopsis thaliana*	GTSSDPIQ	[55]
Q9LS45	*Arabidopsis thaliana*	GRPVGQTN	[55]
Q9LSL8	*Arabidopsis thaliana*	GRYELHYG	[55]
Q9LT23	*Arabidopsis thaliana*	GIRENGIM	[55]
Q9LTA6	*Arabidopsis thaliana*	GTGWRRAF	[55]
Q9LTX1	*Arabidopsis thaliana*	GSADLVDD	[55]
Q9LU40	*Arabidopsis thaliana*	GVDLRQVV	[55]
Q9LUA9	*Arabidopsis thaliana*	GYMCDFCG	[55]
Q9LUD4	*Arabidopsis thaliana*	GGFRFHQY	[55]
Q9LUK5	*Arabidopsis thaliana*	GFFRAATH	[55]

Q9LUM0	*Arabidopsis thaliana*	GTRDSNNR	[55]
Q9LUT0	*Arabidopsis thaliana*	GCFGCCGG	[55]
Q9LUY6	*Arabidopsis thaliana*	GSENGSLM	[55]
Q9LV59	*Arabidopsis thaliana*	GDSEDETG	[55]
Q9LV76	*Arabidopsis thaliana*	GVMEKKLR	[55]
Q9LVD5	*Arabidopsis thaliana*	GDSTFLDR	[55]
Q9LW86	*Arabidopsis thaliana*	GHGTNRVE	[55]
Q9LW88	*Arabidopsis thaliana*	GDSDNAIP	[55]
Q9LYG3	*Arabidopsis thaliana*	GSTPTDLP	[55]
Q9LYW6	*Arabidopsis thaliana*	GIKILKLN	[55]
Q9LZA4	*Arabidopsis thaliana*	GGGYVLFG	[55]
Q9LZF1	*Arabidopsis thaliana*	GEKKEETA	[55]
Q9M129	*Arabidopsis thaliana*	GFIDGKWA	[55]
Q9M1B5	*Arabidopsis thaliana*	GRPLFYDI	[55]
Q9M1H3	*Arabidopsis thaliana*	GKKKSDES	[55]
Q9M1Z4	*Arabidopsis thaliana*	GKNQAYKA	[55]
Q9M2I0	*Arabidopsis thaliana*	GKQLAKKI	[55]
Q9M2J0	*Arabidopsis thaliana*	GKQINNTF	[55]
Q9M2J5	*Arabidopsis thaliana*	GNLVDNKF	[55]
Q9M2S7	*Arabidopsis thaliana*	GDAIDLSG	[55]
Q9M2U3	*Arabidopsis thaliana*	GPIKTIKK	[55]
Q9M308	*Arabidopsis thaliana*	GQYATVWD	[55]
Q9M7Q2	*Arabidopsis thaliana*	GTHINFNN	[55]
Q9M8S6	*Arabidopsis thaliana*	GGSSGGGV	[55]
Q9M9G0	*Arabidopsis thaliana*	GHVQLLTP	[55]
Q9M9K1	*Arabidopsis thaliana*	GSSGDVNW	[55]
Q9MAS5	*Arabidopsis thaliana*	GQQSLIYS	[55]
Q9NQ55	*Homo sapiens*	GQSGRSRH	[55]
Q9NQA5	*Homo sapiens*	GGFLPKAE	[55]
Q9NRD0	*Homo sapiens*	GQGLWRVV	[55]
Q9NVL8	*Homo sapiens*	GLSHSKTH	[55]
Q9NWC5	*Homo sapiens*	GNFRGHAL	[55]
Q9NZD8	*Homo sapiens*	GEIKVSPD	[55]
Q9NZP6	*Homo sapiens*	GNLLSKFR	[55]
Q9S721	*Arabidopsis thaliana*	GLDWGPVL	[55]
Q9S757	*Arabidopsis thaliana*	GISLAFMA	[55]
Q9S7L7	*Arabidopsis thaliana*	GGADWGPV	[55]
Q9S7V4	*Arabidopsis thaliana*	GSGAGNFL	[55]
Q9S7X6	*Arabidopsis thaliana*	GEAVEVMF	[55]
Q9SCW5	*Arabidopsis thaliana*	GTVCESVA	[55]
Q9SCX5	*Arabidopsis thaliana*	GIEKRKKM	[55]
Q9SD44	*Arabidopsis thaliana*	GFGAIRSI	[55]
Q9SF13	*Arabidopsis thaliana*	GTVDIFNG	[55]
Q9SFC4	*Arabidopsis thaliana*	GTWKNKNS	[55]
Q9SFU0	*Arabidopsis thaliana*	GTENQGYP	[55]
Q9SFW6	*Arabidopsis thaliana*	GGSDENRH	[55]
Q9SG63	*Arabidopsis thaliana*	GRTTWFDV	[55]

Q9SHG0	*Arabidopsis thaliana*	GYDNVCGE	[55]
Q9SHM1	*Arabidopsis thaliana*	GPFHQQSR	[55]
Q9SIB6	*Arabidopsis thaliana*	GCFGRTPK	[55]
Q9SIE8	*Arabidopsis thaliana*	GHYLVPIH	[55]
Q9SII8	*Arabidopsis thaliana*	GGEGDSSQ	[55]
Q9SIM4	*Arabidopsis thaliana*	GFKRFVEI	[55]
Q9SIN2	*Arabidopsis thaliana*	GFTSRGNP	[55]
Q9SJW5	*Arabidopsis thaliana*	GGLGSPCG	[55]
Q9SK71	*Arabidopsis thaliana*	GSGNHVDI	[55]
Q9SKA6	*Arabidopsis thaliana*	GRDQEGSP	[55]
Q9SKD4	*Arabidopsis thaliana*	GWTRPPHG	[55]
Q9SKH6	*Arabidopsis thaliana*	GRRRRSQQ	[55]
Q9SL28	*Arabidopsis thaliana*	GSGKTNRP	[55]
Q9SLF3	*Arabidopsis thaliana*	GDGTEFVV	[55]
Q9SLI0	*Arabidopsis thaliana*	GDQQKIHP	[55]
Q9SP35	*Arabidopsis thaliana*	GTPETSRE	[55]
Q9SQR5	*Arabidopsis thaliana*	GRNLGSAF	[55]
Q9SRB0	*Arabidopsis thaliana*	GHHSCCNQ	[55]
Q9SRE5	*Arabidopsis thaliana*	GSEQDQRK	[55]
Q9SRH7	*Arabidopsis thaliana*	GFDSVKVM	[55]
Q9SRN1	*Arabidopsis thaliana*	GSKQPYLN	[55]
Q9SRT8	*Arabidopsis thaliana*	GSRQGPPK	[55]
Q9SSK1	*Arabidopsis thaliana*	GDEIVPPA	[55]
Q9SSM4	*Arabidopsis thaliana*	GNTDKLMN	[55]
Q9STM8	*Arabidopsis thaliana*	GVIRTSRT	[55]
Q9SUQ7	*Arabidopsis thaliana*	GTNGTTCP	[55]
Q9SUS4	*Arabidopsis thaliana*	GLPEDFIT	[55]
Q9SV91	*Arabidopsis thaliana*	GCIGVVNV	[55]
Q9SVC9	*Arabidopsis thaliana*	GLTPTATL	[55]
Q9SVL6	*Arabidopsis thaliana*	GRMDYLAM	[55]
Q9SX25	*Arabidopsis thaliana*	GGTKLTHV	[55]
Q9SX28	*Arabidopsis thaliana*	GSLLQGFT	[55]
Q9T081	*Arabidopsis thaliana*	GGLKFHVL	[55]
Q9T095	*Arabidopsis thaliana*	GEIATEFT	[55]
Q9UBF8	*Homo sapiens*	GDTVVEPA	SwissProt (acetyl-Gly)
Q9UET6	*Homo sapiens*	GRTSKDKR	[55]
Q9UL36	*Homo sapiens*	GLCGLLER	[55]
Q9UMX6	*Homo sapiens*	GQEFSWEE	[55]
Q9UNX4	*Homo sapiens*	GLTKQYLR	[55]
Q9UNX9	*Homo sapiens*	GLARALRR	[55]
Q9UP83	*Homo sapiens*	GWVGGRRR	[55]
Q9UQR0	*Homo sapiens*	GQTVNEDS	[55]
Q9XI22	*Arabidopsis thaliana*	GAPLVCHG	[55]
Q9XIB3	*Arabidopsis thaliana*	GTFLGHFV	[55]
Q9XIF8	*Arabidopsis thaliana*	GEKEEVKL	[55]
Q9Y580	*Homo sapiens*	GAAAAEAD	SwissProt (acetyl-Gly)
Q9Y6F6	*Homo sapiens*	GMDLTCPF	[55]

Q9Y6Z5	*Homo sapiens*	GAAGSDGR	[55]
Q9ZPH4	*Arabidopsis thaliana*	GKGGREKI	[55]
Q9ZPS0	*Arabidopsis thaliana*	GMTTDSMK	[55]
Q9ZPU0	*Arabidopsis thaliana*	GWCITVVH	[55]
Q9ZPY1	*Arabidopsis thaliana*	GTHVAPWK	[55]
Q9ZS51	*Arabidopsis thaliana*	GSSPPKKT	[55]
Q9ZU00	*Arabidopsis thaliana*	GLINQWFP	[55]
Q9ZUE1	*Arabidopsis thaliana*	GDQGVQQM	[55]
Q9ZUI4	*Arabidopsis thaliana*	GIPDAAQD	[55]
Q9ZUI8	*Arabidopsis thaliana*	GLMDTRWE	[55]
Q9ZUW8	*Arabidopsis thaliana*	GKPTTQNN	[55]
Q9ZV27	*Arabidopsis thaliana*	GFSDAGIY	[55]
Q9ZVT0	*Arabidopsis thaliana*	GFGSVYRS	[55]

Appendix 8 – Post-translational Blind Testing Set (15 proteins)

Uniprot ID	*Organism*	Myristoylation Site	Caspase Site/Octapeptide	Annotation Source
P42858	*Homo sapiens*	551	DLND/GTQASSPI	SwissProt
Q06002	*Bos taurus*	433	DVPD/GGKISKAF	SwissProt
Q13177	*Homo sapiens*	213	SHVD/GAAKSLDK	SwissProt
Q12934	*Homo sapiens*	434	DVPD/GGQISKGF	SwissProt
O60503	*Homo sapiens*	596	EVID/GSQVSSGP	[74]
Q8IVF2	*Homo sapiens*	2847	VEAD/GSFPSMQG	[74]
Q9BVC5	*Homo sapiens*	106	IVFD/GSSTSTSI	[74]
O75122	*Homo sapiens*	17	ESVD/GNRPSSAA	[74]
Q13620	*Homo sapiens*	44	SATD/GNTSTTPP	[74]
P06396	*Homo sapiens*	404	DQTD/GLGLSYLS	[74]
Q12906	*Homo sapiens*	440	VEVD/GNSFEASG	[74]
O60664	*Homo sapiens*	10	AEAD/GSTQVTVE	[74]
Q96T37	*Homo sapiens*	751	DRSD/GSAPSTST	[74]
O94875	*Homo sapiens*	46	QSLD/GTTSSSIP	[74]
Q96FJ0	*Homo sapiens*	208	EQID/GSALSCFS	[74]

Appendix 9 – BetAware-Deep DOME card

DOME	Version	1.0
Data	Provenance	Training set and blind set for topology prediction: 142 proteins from the Protein Data Bank (PDB). (Minimum resolution: 1.5 Å) Blind set for discrimination: 8580 proteins from PRED-TMBB2 [49]
	Dataset splits	Training set: 58 TMBB and 69 non-TMBB proteins. Balancing: 46% positive and 54% negative. TM residues: 11,579; non-TM residues: 39,022. 10-fold cross-validation split Blind test set (topology): 15 positive examples Blind test set (discrimination): 1009 positive examples, 7571 negative examples
	Redundancy between data splits	Maximum sequence identity 25% at 50% coverage between training and blind test sets, and among cross-validation splits.
	Availability of data	Yes. URL: https://busca.biocomp.unibo.it/betaware2/datasets/
Optimization	Algorithm	Long Short Time Memory Network + Grammatical-restrained hidden conditional random fields
	Meta-predictions	No
	Data encoding	Sequence profiles, Profile-weighted hydrophobic moment
	Parameters	418,053 parameters for BLSTM; 7,472 parameters for GRHCRFs
	Features	21 features per residue for BLSTM; 25 features per residue for GRHCRF
	Fitting	For BLSTM, parameters are about 10 times the number of training examples. Overfitting is limited with regularization (dropout). For GRHCRF, the number of training examples is about 6 times the parameters, suggesting neither over- nor under-fitting.
	Regularization	Dropout used in all BLSTM layers with high rate (50%). Gaussian regularization adopted in GRHRCRFs.
	Availability of configuration	No
Model	Interpretability	Black box, as correlation between input and output is masked. No attempt was made to make the model transparent.
	Output	Classification at the protein level (TMBB or not TMBB). Labelling of the sequence
	Execution time	about 12 seconds per protein
	Availability of software	Web server. URL: https://busca.biocomp.unibo.it/betaware2/
Evaluation	Evaluation method	Independent dataset
	Performance Measures	For protein classification: Sensitivity, Specificity, Matthews Correlation Coefficient. Labelling: Accuracy, Segment Overlap Value, Number of correct topologies
	Comparison	BetAware, BOCTOPUS2, PRED-TMBB2, HHomp
	Confidence	Non estimated
	Availability of evaluation	No

Appendix 10 – SVMyr DOME card

DOME	Version	1.0
Data	Provenance	Datasets for co-translational myristoylation: 257 octapeptides from SwissProt; 552 from [55]; 18 from [90]; 5 from [91]; 11 from [92]; 5 from [93]. Dataset for post-translational myristoylation: 4 proteins from SwissProt and 11 from [74]
	Dataset splits	Training set: 232 positive octapeptides and 232 negative octapeptides. Balancing 50%-50%. 10-fold cross-validation split Testing set: 88 positive octapeptides and 528 negative octapeptides. Balancing: 14% - 86%.
	Redundancy between data splits	Maximum Hamming Distance equal to 4 between training and testing sets and among cross-validation subsets
	Availability of data	Yes. URL: https://busca.biocomp.unibo.it/lipipred/datasets/
Optimization	Algorithm	Ensemble of Support Vector Machines
	Meta-predictions	No
	Data encoding	Position Specific Scoring Matrix, Physicochemical features (hydrophobicity, size, charge, secondary structure propensities)
	Parameters	121 support vectors (average over SVMs)
	Features	12 features per octapeptide
	Fitting	The number of examples is about 5 times parameters, suggesting neither over- nor under-fitting.
	Regularization	L2 regularization
	Availability of configuration	No
Model	Interpretability	Black box, as correlation between input and output is masked. No attempt was made to make the model transparent.
	Output	Classification of the protein as co-translationally myristoylated or not. Annotation of putative post-translational myristoylation sites.
	Execution time	0.1 seconds per protein
	Availability of software	Web server. URL: https://busca.biocomp.unibo.it/lipipred/datasets/
Evaluation	Evaluation method	Independent dataset
	Performance Measures	Sensitivity; Precision; Matthews Correlation Coefficient (MCC); F1-score; Receiver Operating Characteristic curve and relative Area Under the Curve.
	Comparison	NMT predictor; Myristoylator; TermiNator3; available patterns: PROSITE and [53].
	Confidence	Non estimated
	Availability of evaluation	No