Hybrid Biogeography Based Simultaneous Feature Selection and Prediction of N-Myristoylation Substrate Proteins Using Support Vector Machines and Random Forest Classifiers

Shameek Ghosh\textsuperscript{1}, Nayana Ramachandran\textsuperscript{2}, C. Venkateshwari\textsuperscript{3}, and V.K. Jayaraman\textsuperscript{1,*}

\textsuperscript{1} Evolutionary Computing and Image Processing Group, Centre for Development of Advanced Computing (CDAC), Pune, Maharashtra
\textsuperscript{2} Persistent Systems Limited, ‘Aryabhata - Pingala’, 9A / 12, Erandawane, Pune 411004, India
\textsuperscript{3} Satyabhama University, Chennai, India
\{shameekg,jayaramanv\}@cdac.in, nayana_ramachandran@persistent.co.in, venkateswari.ch@gmail.com

Abstract. Majority of proteins undergo important post-translational modifications (PTM) that may alter physical and chemical properties of the protein and mainly their functions. Laboratory processes of determining PTM sites in proteins are laborious and expensive. On the contrary, computational approaches are far swifter and economical; and the models for prediction of PTMs can be quite accurate too. Among the PTMs, Protein N-terminal N-myristoylation by myristoyl-CoA protein N-myristoyltransferase (NMT) is an important lipid anchor modification of eukaryotic and viral proteins; occurring in about 0.5% encoded NMT substrates. Reliable recognition of myristoylation capability from the substrate amino acid sequence is useful for proteomic functional annotation projects as also in building therapeutics targeting the NMT. Using computational techniques, prediction-based models can be developed and new functions of protein substrates can be identified.

In this study, we employ Biogeography based Optimization (BBO) for feature selection along with Support Vector Machines (SVM) and Random Forest for classification of N-myristoylation sequences. The simulations indicate that N-myristoylation sites can be identified with high accuracy using hybrid BBO wrappers in combination with weighted filter methods.

Keywords: Post-translational modifications (PTM), N-myristoylation, Biogeography based Optimization (BBO), Support Vector Machines (SVM), Random Forest classifier, Amino Acid Indices, dbPTM, SwissProt.

* Corresponding author.
1 Introduction

Most proteins undergo covalent modifications during or after assembly of the polypeptide chain. This ensures proper protein conformation or folding [1], or directs the newly formed protein to distinct cellular apparatus. Studies indicate that some other protein translation modifications (PTMs) may occur after folding and localization activities are completed [2]; thus influencing the biological or catalytic activity of the protein (like protein degradation). Due to high data complexity, traditional techniques like mass spectroscopy, chemical proteomics, may turn out to be very costly. Thus, profile computations with computational methods are definitely the key to enhance future research in PTM substrate characterization. Protein N-myristoylation refers to the co-translational or post-translational covalent attachment of myristate, a 14-carbon saturated fatty acid, to the N-terminal glycine of eukaryotic and viral proteins. Myristoylation by the myristoyl-CoA protein \( N\)-myristoyltransferase (\( NMT \)) is an important lipid anchor modification of eukaryotic and viral proteins [3]. \( NMT \) recognizes the sequence motif of suitable substrate proteins at the N-terminus and attaches the lipid moiety to the absolutely essential N-terminal glycine residue [3]. N-Myristoyl proteins include proteins involved in different signal transduction cascades, especially intracellular and those enabling rapid, flexible cell responses [4]. Studies reveal that myristoylation and membrane-binding are known to regulate the activity of kinases and influence c-Src stability [5]. It is also known to be a vital growth and regulation component in development of the leukocytic lineage [6]. Recent scientific studies also indicate that HIV protein Nef is preferentially myristoylated by recombinant human isozyme \( NMT2 \). Thus, selective inhibition of \( NMT2 \) may be termed as a novel means of blocking HIV virulence [4] and chemical proteomics has indicated the scope of developing therapeutics for \( NMT \) regulation [7].

In this regard, we have used the in-depth studies of the amino acid sequence variability of substrate proteins (on binding site analyses in X-ray structures or 3D homology models for NMTs from various taxa), biochemical data extracted from the scientific literature; and employed a hybrid BBO-based filter wrapper algorithm with SVM and RF for feature selection and prediction. It is discovered that, at least within a complete substrate protein, the N-terminal 17 amino-acid residues experience different types of variability restrictions that reveal a physical property pattern [8]. Accordingly, we have considered the relevant properties from N-terminal 17 amino-acid residues, where three motif regions may be identified as follows: Region 1 (positions 1–6) fitting the binding pocket, region 2 (positions 7–10) interacting with the NMT’s surface at the mouth of the catalytic cavity and region 3 (positions 11–17) comprising a hydrophilic linker.

2 Materials and Methods

2.1 Dataset Generation

dbPTM [9] is a database that compiles information on protein post-translational modifications (PTMs). The database includes all of the experimentally validated PTM