ChrSLoc-Net: Machine Learning-Based Prediction of Channelrhodopsins Proteins within Plasma Membrane

Muhammad Nabeel Asim*†, Muhammad Ali Ibrahim*†, Muhammad Imran Malik‡, Andreas Dengel*†, Sheraz Ahmed*
Email: muhammad_nabeel.asim@dfki.de
*German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
†TU Kaiserslautern, Kaiserslautern, Germany
‡National Center for Artificial Intelligence (NCAI), National University of Sciences and Technology, 44000 Islamabad, Pakistan

2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) | 978-1-6654-0358-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/BHI50953.2021.9508615

***Abstract*—There is a rising interest in investigating mechanisms and engineering of integral membrane proteins (MPs) which make crucial contribution in perceiving and controlling cellular response against different external signals. MPs need to be inserted, folded and expressed correctly in lipid bi-layer and transferred to appropriate cellular location to perform its diverse range of functions. Channelrhodopsins (ChRs), light gated ion-channel proteins belonging to microorganisms are imminent for diverse neurobiology applications where expression as well as localization to plasma membrane is a pre-condition for function. Developing robust computational methodologies to accurately identify ChRs localization is an active area of research. Existing computational approaches make use of one-hot-vector encoding or protein embeddings to encode MP sequences that are fed to Gaussian process regression model. These approaches lack to accurately predict the localization of MP proteins. The paper in hand proposes ChRsLoc-Net predictor that makes use of composition-transition-distribution (CTDC) physico-chemical properties based sequence encoder along with Hubber regressor. Over benchmark dataset, proposed ChrSLoc-Net approach outperforms state-of-the-art MP localization predictor with a significant margin of 9% in terms of mean absolute error. We anticipate that this study will largely assist biologist to comprehend diverse biological processes subject to localization patterns of MPs within plasma membrane.**

***Index Terms*—Channelrhodopsins, Proteins, Plasma Membrane, Machine Learning, Robust Regression, Physico-Chemical Properties, Localization of Membrane Proteins**

I. Introduction

Acquiring the precise control on neuronal activity in time and space is the primary goal of experimental and translational neurobiology. With the influx of biological technologies, many attempts to monitor and control the activity of neurons in living tissues have utilized magnetic [1], electrical [9], and ultrasound stimulation [17] with different degrees of effectiveness. Recent years have witnessed major breakthroughs in controlling neuronal activity using light (Optogenetics) [10]. Channelrhodopsins (ChRs), a subfamily of retinylidene proteins (rhodopsins) function as light gated ion channels [12] and have gained huge attention in the field of Optogenetics. ChRs are expressed in diverse organisms where they act as sensory photoreceptors within mono-cellular green algae, governing the response to light like exciting or suppressing the neurons [4], [13]. Further, ChRs enable light to regulate a wide range of cellular processes such as electrical excitability, calcium influx, intracellular acidity and several others [4]. The utility of ChRs primarily depends on their capability to express as well as localize properly to plasma membrane within eukaryotic cells. Rhodopsins especially localized to plasma membrane have outstanding pharmaceutical, engineering, and optogenetic applications such as silencing of different neural activities mainly through light illumination [7]. Considering, changes in amino acid distribution of protein sequences oftenly abrogate localization, predictor for ChRs which express and localize well to plasma membrane is of great value as it can greatly facilitate deeper comprehension of core functionality and pave way for protein engineering.

Both sequence and structural elements important for membrane localization have been a point of interest for extensive investigation [7], [8], [19]. Scientific studies have facilitated enough understanding of membrane protein sequence related factors for localization like signal peptide sequence, higher hydrophobicity in transmembrane domain, and positive charge on membrane cytoplasm interface [18]. However, still these rules are not sufficient for protein engineering because there exist a significant number of protein sequences which pursue aforementioned rules yet fail to locate within plasma membrane. Membrane protein sequence changes not only impacts expression but also the localization. Such impact is largely context dependent, eliminating localization in one particular sequence context does not affect other, and subtle changes in amino acid distribution can produce dramatic effects [2], [6], [7]. Precisely, sequence determinants of expression as well as localization can not be acquired by very simple rules.

Considering the need of computer-aided program that can precisely predict the localization of proteins by using only sequence information, Yang et al. [21] developed a computational approach that utilized one-hot-vector encoding to transform protein sequences into statistical representation. Encoded sequences were passed to Gaussian process regressor. Following the success of pre-trained protein embeddings, Yang et al. [21] also utilized heterogeneous protein sequence data to train Word2vec model in an unsupervised manner. Using

Word2vec based transfer learning, statistical representation of protein sequences was generated and passed to Gaussian process regression which predicted localization of MPs. Critical analysis of state-of-the-art MPs localization predictor indicates that existing sequence encoding approaches do not capture rich characteristics and relationships of protein residues while generating statistical vectors. While one-hot encoding lacks to capture position and correlation of amino acids, pre-trained embedding usually perform better when generated using task specific training data. However, state-of-the-art predictor makes use of pre-trained embeddings generated using heterogeneous protein sequences where MPs localization data is quite limited, which is why Gaussian process regressor failed to generalize well.

Considering the room for improvement, we focus on building a robust regression model solely using sequence information of ChRs and training data acquired from ChR related sources under standardized conditions. We utilize residue physio-chemical properties to precisely capture the distribution patterns of amino acids and Hubber regressor (ChRsLoc-Net) to predict the localization of Channelrhodopsins within plasma membrane. Empirical evaluation on benchmark membrane protein localization dataset indicates that proposed ChRsLoc-Net raises the previous best performance by a significant figure of 9% in terms of mean absolute error.

II. Materials and Methods

The workflow of proposed computational predictor ChrSLoc-Net is illustrated in Figure 1, core modules of which including feature extraction and regression are briefly described in following sub-sections.

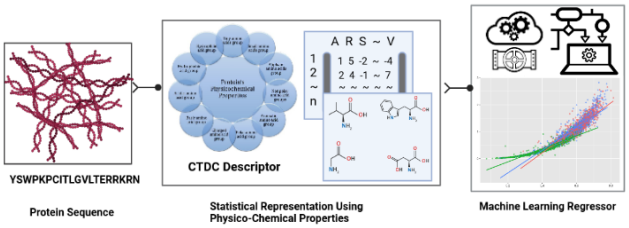

Fig. 1. Workflow of Proposed ChrSLoc-Net based on CTDC Encoding and Hubber Regressor for the Accurate Prediction Channelrhodopsins Proteinswithin Plasma Membrane

A. Membrane Protein Sequence Statistical Representation Learning

In order to represent diverse patterns of amino acid composition having particular physicochemical or structural properties in protein sequences, we have utilized 13 different physicochemical properties to compute these characteristics [3], [5], [20]. These properties are normalised van der Waals volume, hydrophobicity, polarisability, polarity, charge, solvent accessibility, and secondary structures. Protein sequence encoding is learned in 3 different steps: 1) Transforming membrane protein sequence into a sequence of physicochemical or structural properties of residues, 2) Segregating 20 amino acids into 3 groups for 7 distinct physicochemical properties on the basis of major cluster indices generated by Tomii et al. [16], 3) Computing CTDC of residues. Composition is computed by taking the fraction among the number of certain amino acids and total length N of amino acids in membrane protein sequence.

$$Composition(e) = \frac{n_e}{N} \tag{1}$$

Where n_e denotes the sum of number of e, certain amino acid, in the sequence. The value of e could be from 1-to-3, representing the type of amino acid. Considering 2 amino acids are a and b, transition (T) can be computed by taking the fraction between number of ab and ba with membrane protein sequence length $N-1$.

$$Transition(ab+ba) = \frac{n_{ab}+n_{ba}}{N-1} \tag{2}$$

Distribution deals with the position of certain amino acid in the total length of protein sequence which indicates chain length over which first 25%, 50%, and 100% amino acids of specific amino acid are resided. Using 13 physicochemical properties and segregation of amino acids in 3 groups, CTDC generates (13*3) 39-dimensional sequence vectors.

B. Machine Learning Regression: Huber Regression

Considering the possibility of outliers in the dataset, Huber regression penalize uncommon instances by assigning them less wights as compared to other instances of dataset. To handle the outliers in a robust manner, Huber regression utilizes a different loss function rather than standard least square error function. More specifically, in case of too large values which indicates the possibility of being outliers, Huber transform its loss function to linear loss function in order to reduce their impact on the model. Dynamics of huber loss can be mathematically expressed as:

$$HuberLoss = \left\{ \begin{array}{cc} \frac{1}{2}a^2 & if|a| < \delta \\ \delta(|a| - \frac{1}{2}\delta) & otherwise \end{array} \right\} \tag{3}$$

Here δ indicates how large data needs to be in order to activate linear loss criteria. Huber loss is very much identical to trivial least square error penalty function for small residues, however on large datasets, its penalty raises linearly instead of quadratically, which indicates its trait to be more forgiving for outliers. In this manner, their involvement in global cost function gets reduced, which is why hyperplane remains very close to the majority of points despite the presence of outliers.

C. Membrane Protein Localization Dataset

Membrane Protein (Channelrhodopsin) localization dataset [2] was developed by designing two separate recombination libraries of ten-block using three parent ChRs(CheRiff, CsChrimsonR, and C1C2). Every chimeric ChR variant present in recombination libraries comprised of sequence blocks taken from parental ChRs. This dataset consists of 248 sequences. Genes of 248 sequences were constructed as

well as expressed inside embryonic kidney cells. Membrane localization of 284 sequences was estimated by Bedbrook et al. [2].

III. Experimental Setup and Results

We implement CTDC encoder using Python language and Hubber regressor is employed from Scikit-Learn library. Using standard train-test split of benchmark dataset provided by Yang et al. [21], experimentation is performed with CTDC encoding based hubber regressor as well as 3 other most widely used regression algorithms (RF, SVR, XGBosst). For RF, initial estimator range is defined 100-300, sub-sampling rate of 0.2-0.8, lambda 1e-1 to 1e-8 for XGBoost, and SVR is evaluated using linear, RBF, and polynomial kernel using degree of 2-5. For Hubber regressor, initial epsilon range is defined as 1.10 to 1.40, alpha of 0.0001 to 0.0005, and maximum iteration of 50 to 300. Considering the success of grid search for automated parameters search, optimal values of different hyperparameters are found using Grid search. Following evaluation measures used by Yang et al. [21], a fair performance comparison of proposed and existing methodology is performed in terms of mean absolute error (MAE) and Kendall rank correlation coefficient (Tau). Further, to showcase proposed CTDC-Hubber regressor is not biased towards certain evaluation measure, we also report the performance in terms of 2 other most widely used evaluation measures namely mean squared error (MSE) and R-squared score.

A. Performance Comparison of ChrSLoc-Net with Baseline and Existing Regressors

In order to prove the specialty of physico-chemical property based encoding and Hubber loss which integrates the advantages of different loss functions (MAE, MSE) yet avoiding their disadvantages and better handle outliers, we compare the performance of proposed Huber-CTDC with 3 other regressors including Random Forest, Support Vector Regressor (SVR) and XGBoost regressor (XGB).

Regressors	**Evaluation Measures**			
	MAE	MSE	Tau	R square
Baseline				
RF	0.8274	1.2312	0.2870	0.5530
XGB	1.4854	3.5734	0.2272	-1.069
SVR	1.0558	1.6921	0.4583	0.0201
Existing Approaches				
OneHot_GPR	0.76	–	0.59	–
OneHot_Struct_GPR	0.76	–	0.60	–
AAIndex_GPR	0.76	–	0.55	–
ProFET_GPR	1.03	–	0.32	–
Word2vec_GPR	0.73	–	0.60	–
Proposed Approach				
ChrSLoc-Net	**0.6398**	**0.7365**	**0.6136**	**0.5735**

TABLE I
Performance Figures Produced by 4 Different Regressors for Membrane Protein Sub-cellular Localization Prediction

Table I describes the performance of RF, SVR, XGB, and ChrSLoc-Net in terms of 4 different evaluation measures including MAE, MSE, Tau, and R-square. Analyzing the performance of different approaches reveals that among 3 baseline regressors, overall RF achieves better performance followed by SVR, where XGB produces the worst performance across all evaluation metrics. Among all 4 approaches, the proposed ChrSLoc-Net produced the most promising performance, outperforming RF by the figure of 19%, 49%, 32%, and 2% in terms of MAE, MSE, Tau, and R-squared score.

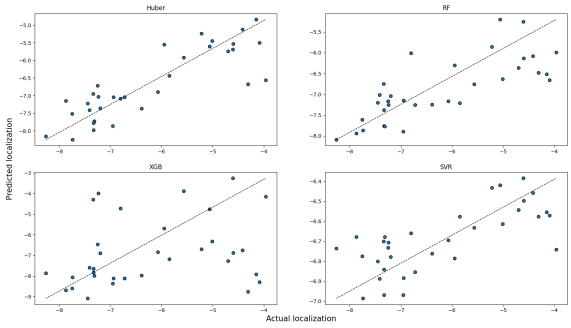

Fig. 2. Decision Surfaces of 4 Different Regressors for Membrane Protein Sub-cellular Localization Prediction

To further analyze the performance of 4 different approaches, decision surfaces of all 4 approaches are shown in scatter plots (Figure 2). As is evident from the Figure 2, among all 3 baseline approaches, RF produces better decision boundary, however among all 4 approaches, proposed ChrSLoc-Net decision surface is far better than other approaches. This is because, it has the most number of points on the decision boundary or close to the decision boundary.

In order to prove the integrity of proposed ChrSLoc-Net methodology, we compare the performance of ChrSLoc-Net with existing integral membrane protein sub-cellular localization predictors.

For a fair performance comparison, Table I illustrates the performance of ChrSLoc-Net (Huber-CTDC) and existing predictive methodologies in terms of mean absolute error (MAE) and Kendall rank correlation coefficient (Tau). As is indicated by the Table I, Gaussian process regression (GPR) achieves almost similar MAE value of 0.76 with one-hot sequence encoding, one-hot sequence-structured based encoding, and AAIndex descriptor. This performance improves by the figure of 3% on the induction of word2vec pre-trained embeddings. Among all existing approaches, GPR marks the worst performance using ProFET encoding approach. Whereas, proposed physico-chemical property based predictor Huber-CTDC marks best performance by a significant figure of 9% in terms of MAE. In terms of Tau, existing computational predictor GPR once again achieves better performance with word2vec embeddings followed by one-hot sequence-structure based and one-hot sequence based encoding. Similar to MAE, ProFET marks the lowest Tau score of 32%. Proposed Huber-CTDC approach outperform existing membrane protein sub-cellular localization predictor by comparable margin in terms of Tau.

IV. Discussion

To effectively capture the biological characteristics of residues, researchers have pre-dominantly utilized physico-chemical property based descriptors due to their simplicity and capability to preserve important sequence information [11], [22]. Physico-chemical properties based encodings are showing great promise in diverse Proteomics applications such as determining post-translational modifications of proteins [14], ubiquitylation sites of plant-specific proteins [15], etc. Unlike traditional encoding schemes, physico-chemical properties based descriptors represent precise and similar inherent relations of all residues in a compact feature space. This is due to the fact that physical and chemical properties are well described and highly correlated in terms of diverse core components such as size, hydrophobicity, extent of degenerate triplet codons, priority in a beta strand, and occurrence frequency of residues in beta strand [14]. For the membrane protein localization prediction, empirical evaluation has indicated that 39-dimensional sequence vectors of a physico-chemical properties based encoding descriptor CTDC managed to precisely conserve biophysical characteristics, similarities and differences of residues while diminishing superfluous details. Unlike traditional least squares penalty, hubber penalty increases in linear manner rather than quadratic manner, therefore, the idea of using a robust machine learning regressor (Hubber Regressor) whose loss function is not heavily impacted by the outlier largely assisted the proposed methodology ChrSLoc-Net to achieve promising predictive performance.

V. Conclusion

This research introduces CTDC encoder that makes use of physico-chemical properties of amino acids to encode Channelrhodopsins protein into fixed length statistical vectors. Utilizing CTDC encoded vectors, we explore the potential of 4 different regressors namely Random forest, XGboost, SVR and Hubber for localization prediction of membrane proteins. A comprehensive experimentation over membrane proteins benchmark dataset reveals that, overall, CTDC encoder generates better statistical vectors by incorporating positional and semantic information. Using CTDC encoding, hubber regressor manages to outperform state-of-the-art membrane protein location prediction performance by 9% in terms of mean absolute error. Although performance values of other three regressors by using CTDC encoding is less then Hubber regressor, but their values are marginally comparable with state-of-the-art performance values. We believe that proposed methodology will help biologists to understand different biological processes of membrane protein by precisely predicting their location inside the membrane.

References

[1] Anthony T Barker. The history and basic principles of magnetic nerve stimulation. *Electroencephalogr Clin Neurophysiol Suppl.*, 51:3–21, 1999.

[2] Claire N Bedbrook, Austin J Rice, Kevin K Yang, Xiaozhe Ding, Siyuan Chen, Emily M LeProust, Viviana Gradinaru, and Frances H Arnold. Structure-guided schema recombination generates diverse chimeric channelrhodopsins. *Proceedings of the National Academy of Sciences*, 114(13):E2624–E2633, 2017.

[3] Wei Chen, Fulei Nie, and Hui Ding. Recent advances of computational methods for identifying bacteriophage virion proteins. *Protein and peptide letters*, 27(4):259–264, 2020.

[4] Karl Deisseroth and Peter Hegemann. The form and function of channelrhodopsin. *Science*, 357(6356), 2017.

[5] Yijie Ding, Jijun Tang, and Fei Guo. Identification of drug-target interactions via multiple information integration. *Information Sciences*, 418:546–560, 2017.

[6] Mylinh T Duong, Todd M Jaszewski, Karen G Fleming, and Kevin R MacKenzie. Changes in apparent free energy of helix–helix dimerization in a biological membrane due to point mutations. *Journal of molecular biology*, 371(2):422–434, 2007.

[7] Assaf Elazar, Jonathan Weinstein, Ido Biran, Yearit Fridman, Eitan Bibi, and Sarel Jacob Fleishman. Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. *Elife*, 5:e12125, 2016.

[8] Karen G Fleming. Energetics of membrane protein folding. *Annual review of biophysics*, 43:233–255, 2014.

[9] Chadwick M Hales, John D Rolston, and Steve M Potter. How to culture, record and stimulate neuronal networks on micro-electrode arrays (meas). *JoVE (Journal of Visualized Experiments)*, (39):e2056, 2010.

[10] Richard H Kramer, Doris L Fortin, and Dirk Trauner. New photochemical tools for controlling neuronal activity. *Current opinion in neurobiology*, 19(5):544–552, 2009.

[11] Timmy Manning and Paul Walsh. The importance of physicochemical characteristics and nonlinear classifiers in determining hiv-1 protease specificity. *Bioengineered*, 7(2):65–78, 2016.

[12] Georg Nagel, Doris Ollig, Markus Fuhrmann, Suneel Kateriya, Anna Maria Musti, Ernst Bamberg, and Peter Hegemann. Channelrhodopsin-1: a light-gated proton channel in green algae. *Science*, 296(5577):2395–2398, 2002.

[13] Oleg A Sineshchekov, Kwang-Hwan Jung, and John L Spudich. Two rhodopsins mediate phototaxis to low-and high-intensity light in chlamydomonas reinhardtii. *Proceedings of the National Academy of Sciences*, 99(13):8689–8694, 2002.

[14] Arslan Siraj, Tuvshinbayar Chantsalnyam, Hilal Tayara, and Kil To Chong. Recsno: prediction of protein s-nitrosylation sites using a recurrent neural network. *IEEE Access*, 9:6674–6682, 2021.

[15] Arslan Siraj, Dae Yeong Lim, Hilal Tayara, and Kil To Chong. Ubicomb: A hybrid deep learning model for predicting plant-specific protein ubiquitylation sites. *Genes*, 12(5):717, 2021.

[16] Kentaro Tomii and Minoru Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engineering, Design and Selection*, 9(1):27–36, 1996.

[17] Yusuf Tufail, Alexei Matyushov, Nathan Baldwin, Monica L Tauchmann, Joseph Georges, Anna Yoshihiro, Stephen I Helms Tillery, and William J Tyler. Transcranial pulsed ultrasound stimulates intact brain circuits. *Neuron*, 66(5):681–694, 2010.

[18] Gunnar von Heijne. Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature*, 341(6241):456–458, 1989.

[19] Stephen H White and William C Wimley. Membrane protein folding and stability: physical principles. *Annual review of biophysics and biomolecular structure*, 28(1):319–365, 1999.

[20] Ke Yan, Xiaozhao Fang, Yong Xu, and Bin Liu. Protein fold recognition based on multi-view modeling. *Bioinformatics*, 35(17):2982–2990, 2019.

[21] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, 2018.

[22] Lulu Yu, Yusen Zhang, Ivan Gutman, Yongtang Shi, and Matthias Dehmer. Protein sequence comparison based on physicochemical properties and the position-feature energy matrix. *Scientific reports*, 7(1):1–9, 2017.