PREDICTION OF TRANSMEMBRANE REGIONS OF G PROTEIN-COUPLED RECEPTORS USING MACHINE LEARNING TECHNIQUES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MUAZZEZ ÇELEBI ÇINAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
MOLECULAR BIOLOGY AND GENETICS

SEPTEMBER 2019

Approval of the thesis:

PREDICTION OF TRANSMEMBRANE REGIONS OF G PROTEIN-COUPLED RECEPTORS USING MACHINE LEARNING TECHNIQUES

submitted by **MUAZZEZ ÇELEBİ ÇINAR** in partial fulfillment of the requirements for the degree of **Master of Science in Molecular Biology and Genetics Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşe Gül Gözen
Head of Department, **Biological Sciences** _____

Assoc. Prof. Dr. Çağdaş Devrim Son
Supervisor, **Biological Sciences Dept., METU** _____

Prof. Dr. Tolga Can
Co-Supervisor, **Computer Engineering Dept., METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Özlen Konu
Molecular Biology and Genetics Dept., Bilkent University _____

Assoc. Prof. Dr. Çağdaş Devrim Son
Biological Sciences Dept., METU _____

Prof. Dr. Tolga Can
Computer Engineering Dept., METU _____

Assoc. Prof. Dr. Tunca Doğan
Health Informatics Dept., Hacettepe University _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Health Informatics Dept., METU _____

Date: 02.09.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Muazzez Çelebi Çınar

Signature :

ABSTRACT

PREDICTION OF TRANSMEMBRANE REGIONS OF G PROTEIN-COUPLED RECEPTORS USING MACHINE LEARNING TECHNIQUES

Çınar, Muazzez Çelebi
M.S., Department of Molecular Biology and Genetics
Supervisor: Assoc. Prof. Dr. Çağdaş Devrim Son
Co-Supervisor
: Prof. Dr. Tolga Can

September 2019, 56 pages

G protein-coupled receptors (GPCRs) are one of the largest and the most significant membrane receptor families in eukaryotes. They transmit extracellular stimuli to the inside of the cell by undergoing conformational changes. GPCRs can recognize a diversity of extracellular ligands including hormones, neurotransmitters, odorants, photons, and ions. These receptors are associated with a variety of diseases in humans such as cancer and central nervous system disorders, and can be proclaimed as one of the most important targets for the pharmaceutical industry. They have seven transmembrane helices that contain essential regions such as ligand binding sites, actuator protein (e.g. G protein) binding sites and cholesterol binding sites. There is a large gap in topology data for membrane proteins due to the experimental limitations resulting from unstability of the membrane. In UniProt, which is a freely available database of protein sequences and structural and functional information, only 29 GPCRs among the thousands have experimentally solved transmembrane (TM) region data. The topology information of other membrane proteins is provided using

the TMHMM prediction tool, which is based on hidden Markov models. However, it incorrectly predicts the total number of TM regions for 6 of the 29 experimentally determined GPCRs. With this study, we try to develop a GPCR-specific TM prediction algorithm using machine learning techniques. The algorithm is based on hydrophobicity of each amino acid in the protein sequence and the secondary structure. As hydrophobicity scale, both Moon-Fleming and Kyte-Doolittle hydrophobicity scales are implemented separately. The secondary structures are derived from the JPred server. With this algorithm, we obtain more than 85% accuracy with higher true positive rate. The results obtained could shed light on many other scientific researches and facilitate structure-based drug discovery with further therapeutic opportunities for many diseases.

Keywords: GPCR, transmembrane, hydrophobicity, classification

ÖZ

ODTÜ TEZ ŞABLONU

Çınar, Muazzez Çelebi
Yüksek Lisans, Moleküler Biyoloji ve Genetik Bölümü
Tez Yöneticisi: Doç. Dr. Çağdaş Devrim Son
Ortak Tez Yöneticisi
: Prof. Dr. Tolga Can

Eylül 2019 , 56 sayfa

G proteine-kenetli reseptörler (GPKRler), ökaryotlardaki en büyük ve en önemli membran reseptörü ailelerinden biridir. Konformasyonel değişikliklerle hücre dışı uyarıcıları hücrenin içine iletirler. GPKRler, hormonlar, nörotransmiterler, odorantlar, fotonlar ve iyonlar dahil olmak üzere çeşitli hücre dışı ligandları tanıyabilir. Bu reseptörler, kanser ve merkezi sinir sistemi bozuklukları gibi insanlarda çeşitli hastalıklarla ilişkilendirilir ve farmasötik endüstrisi için en önemli hedeflerden biri olarak kabul edilirler. Ligand bağlanma, aktüatör proteini (örneğin G proteini) bağlanma ve kolesterol bağlanma bölgeleri gibi önemli bölgeleri içeren yedi adet transmembran helisi vardır. Membranın dinamik yapısından kaynaklanan deneysel sınırlamalar nedeniyle, membran proteinlerinin topoloji bilgisinde eksiklikler vardır. Ücretsiz bir protein veritabanı olan UniProtta, binlerce GPKR arasından sadece 29 tanesinin deneysel olarak çözülmüş topoloji verisi mevcuttur. Diğer membran proteinlerinin topoloji bilgileri, TMHMM tahmin aracı tarafından sağlanır. Ancak, bu 29 reseptör arasından 6 tanesinin toplam TM sayısını yanlış tahmin etmiştir. Bu çalışma ile, makine öğrenme tekniklerini kullanarak GPKRye özgü bir TM tahmin algoritması geliştirmeye çalı-

şıyoruz. Algoritma, protein sekansındaki her bir amino asidin hidrofobisite ve ikincil yapısına dayanmaktadır. Hidrofobisite ölçeği olarak, hem Moon-Fleming hem de Kyte-Doolittle hidrofobisite ölçekleri ayrı ayrı kullanılmıştır. İkincil yapılar JPred ile üretilmiştir. Bu algoritma ile %85 gibi yüksek doğruluk oranının yanı sıra yüksek gerçek pozitif oranı elde edebiliyoruz. Elde edilen sonuçlar birçok başka bilimsel araştırmaya ışık tutabilir ve birçok hastalık için daha fazla tedavi imkânı içeren yapı bazlı ilaç keşiflerini kolaylaştırabilir.

Anahtar Kelimeler: GPKR, transmembran, hidrofobisite, sınıflandırma

To my source of peace

ACKNOWLEDGMENTS

I would like to thank my advisor, Assoc. Prof. Dr. Çağdaş Devrim Son, for his patience, encouragement and for giving me an opportunity to meet the bioinformatics world. I would like to thank my co-advisor, Prof. Dr. Tolga Can, for his support, never-ending guidance and most importantly invaluable friendship.

I would like to thank the rest of thesis committee members, Assoc. Prof. Dr. Özlen Konu, Assoc. Prof. Dr. Tunca Doğan and Assoc. Prof. Dr. Nurcan Tunçbağ for their precious time and comments.

I would like to have a special thank to my source of peace, Ali Çınar, for his continuous support in every possible way. With his precious love and eternal understanding, he was by my side through the hardest and darkest time.

I would like to thank my sister, Ayça Çelebi, for teaching me how to be patient and responsible. I would also like to thank my older sister, Tuğba Uçan, for disallowing me to feel lonely throughout my childhood.

I would like to my aunt, Fatma Bozkurt, for being like a mother to me.

I would like to thank my precious friends, Elif Bozlak, Cansu Demirel, Cansu Dinçer, Evrim Fer, Meriç Kınalı and Gökçe Senger, for being patient with me, for making me much less prejudiced and for becoming my other family.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xv

LIST OF ABBREVIATIONS xvi

CHAPTERS

1 INTRODUCTION 1

1.1 G Protein-Coupled Receptors 1

1.1.1 Transmembrane Regions of GPCRs 4

1.2 Hydrophobicity Scales of Amino Acids 5

1.3 UniProt: Protein Database 6

1.3.1 Prediction of Topology Data of Membrane Proteins in UniProt 6

1.4 Literature Review 6

1.5 Proposed Method 9

1.5.1 The Objective of the Proposed Method 10

1.6 Contributions and Novelties 10

1.7 The Outline of the Thesis 10

2 MATERIALS AND METHOD 13

2.1 Background 13

2.2 Data Selection 13

2.3 Smoothing the Data Set 14

2.4 Data Preparation 17

2.5 Baseline Method 17

2.6 Method 18

2.6.1 Cost Matrix 19

2.7 WEKA 21

2.7.1 Classification Methods 21

2.7.1.1 SMO and Random Forest as Classification Approaches . 22

2.7.2 Evaluation of Algorithm Performance 23

2.7.2.1 Parameters for Evaluating Algorithm Performance . . . 23

3 RESULTS AND DISCUSSION 25

3.1 Data 25

3.2 Smoothing Data 25

3.2.1 Hydrophobicity Scales 26

3.3 Mispredictions of TMHMM 26

3.4 Baseline Method 27

3.5 Parameters and Algorithms for TM Start Residue 28

3.6 Baseline, TMHMM and the Proposed Method 32

3.7 The Known GPCRs 34

3.7.1 GPCRs Having 8 TMs as TMHMM Result 35

3.7.2 GPCRs Having 6 TMs as TMHMM Result 35

3.7.2.1 The Last TM Missing 35

3.7.2.2 The Third TM Missing 36

3.7.3 GPCR with Incorrect Prediction of TM Region Locations . . . 36

4 CONCLUSION AND FUTURE WORK 41

4.1 Conclusion 41

4.2 Future Work 42

REFERENCES 43

APPENDICES

A HYDROPHOBICITY SCALES 55

LIST OF TABLES

TABLES

Table 1.1 Available tools for prediction of topology information of the proteins 9

Table 2.1 GPCRs with experimentally solved topology information 15

Table 2.2 The training sets generated for the TMstart prediction algorithms . . 20

Table 2.3 The training sets generated for the TMend prediction algorithms . . 21

Table 2.4 The cost matrices 22

Table 2.5 The representation of a confusion matrix 23

Table 3.1 Evaluation results of the baseline 28

Table 3.2 Effect of cost-sensitive approach 30

Table 3.3 Evaluation of TMHMM results with our algorithm 30

Table 3.4 Cross-validation results of the models with mf for TM starts 32

Table 3.5 Cross-validation results of the models with kd for TM starts 33

Table 3.6 Cross-validation results of the models with mf for TM ends 34

Table A.1 Numerical Values of Hydrophobicity Scales 56

LIST OF FIGURES

FIGURES

Figure 1.1 GPCR positioned in the membrane 3

Figure 2.1 The workflow 14

Figure 2.2 Smoothing method 16

Figure 2.3 The data preparation 18

Figure 3.1 MCC of SMO and Random Forest 31

Figure 3.2 MCC of baseline, TMHMM, SMO and RF 34

Figure 3.3 GPCR with an extra TM predicted by TMHMM 35

Figure 3.4 GPCRs, the last TM missed by TMHMM 37

Figure 3.5 GPCRs, the third TM missed by TMHMM 38

Figure 3.6 GPCR with incorrect prediction of TM region locations 39

LIST OF ABBREVIATIONS

GPCR	G protein-coupled receptor
TM	Transmembrane region
kdHydrophobicity	Kyte-Doolittle hydrophobicity
mfHydrophobicity	Moon-Fleming hydrophobicity
SMO	Sequential Minimal Optimization
RF	Random Forest

CHAPTER 1

INTRODUCTION

1.1 G Protein-Coupled Receptors

Guanine nucleotide-binding protein (G-protein) coupled receptors (GPCRs) represent a superfamily of cell membrane proteins in eukaryotes [1]. GPCRs are regarded as the largest receptor family, which are encoded by about 2% of the coding genes, more than 800 genes, in the human genome [2, 3].

GPCRs function in signal transduction. Upon activation by ligand binding, the receptor transmits the signal through the cell membrane; as a result, synthesis rates of the secondary messenger molecules inside the cell will be altered to initiate an array of downstream signaling pathways [4].

Signal transfer is achieved by conformational changes in the receptor, which can affect the spatial arrangement of the transmembrane regions of the protein [5]. G protein-coupled receptors do not complex with only heterotrimeric G proteins, but also increasing studies have shown that beta arrestins 1 and 2 act as multifunctional partner proteins in signaling [6]. Different signaling pathways are triggered by distinct actuator types [7].

G protein-coupled receptors can be categorized into odorant/sensory and non-odorant. As the name suggests, odorant GPCRs play roles in pheromone signaling and sensorial activities such as taste, light perception, and olfaction while non-odorant GPCRs contribute to hemostasis, reproduction, metabolism, neurotransmission, and cardiac and immune functions [8].

In contrary to those whose functions are characterized, there are some other GPCRs, functions of which remain unknown. Orphan GPCRs are the GPCR molecules that have not yet be associated with endogenous ligands [9]. Most of the orphan recep-

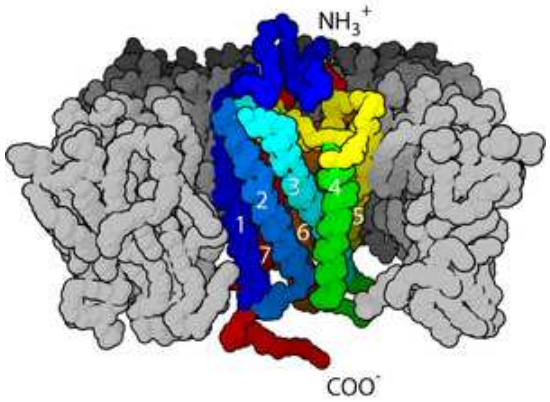
tors act in the brain affecting cognition, mood, movement control, and reward-based learning [10]. Apart from this, the most of the remaining orphan GPCRs affect vision processes [11]. Exploring function and the physiological importance of the orphan GPCRs elucidates which one is associated with which diseases, its exact role and potential or novel therapeutic ways.

GPCRs can recognize a vast diversity of extracellular stimuli and these ligands can be ions, amines, purines, chemokines, lipids, hormones, neurotransmitters, photons, and even organic odorants; therefore, they are associated with a broad range of conditions in the human, including central nervous system disorders such as Parkinson's disease, depression, and schizophrenia, inflammatory diseases, metabolic imbalances, cardiac diseases, monogenic diseases, chronic kidney disease, cancer as well as drug addiction [3, 12, 13]. Moreover, they play essential roles in normal aging [14]. G protein-coupled receptors serve as an important target for pharmaceutical industry. It is estimated that a few GPCRs, as small as only 15% of all the GPCRs in human, are targeted by 35% of the approved therapeutic drugs [15]. More studies in pharmaceutical area can increase the number of GPCRs in terms of developing novel therapeutic ways for treatment of various disorders.

GPCR superfamily consists of five families based on sequence and structural similarities; rhodopsin, secretin, glutamate, adhesion, and frizzled/taste [16]. Despite different families, G protein-coupled receptors share a common topology, a single polypeptide spanning the cell membrane seven times. With regard to this, they have seven hydrophobic transmembrane domains (TM1-TM7), three intracellular loops (ICL1-ICL3) and three extracellular loops (ECL1-ECL3) which link membrane-embedded helices, an intracellular C-terminal tail and an extracellular N-terminus [17, 18]. Partner protein binding site is involved in the intracellular side of transmembrane domain [19]. The Figure 1.1 shows GPCR structure and how to be located in the cell membrane.

G protein-coupled receptors can present and function in the form of either homodimer, heterodimer, higher oligomer or monomer [21, 22]. The various studies assume that the oligomerization is required for internalization and recycling of the GPCRs while some others demonstrate that dimerization is necessary for many GPCRs to appropriately localize in the membrane [23, 24]. It has been established that homo- or hetero-dimerization of the glutamate family GPCRs is required for proper function

Figure 1.1: GPCR positioned in the membrane



The figure is taken from [20].

[25]. A number of studies suggest that when constructing heterodimer structures, GPCR complexes are likely to exhibit different functional properties compared with monomeric or homo-dimeric states of each receptor, resulting in altered pharmacology [26]. In addition to necessity for cellular organization and function, dimerization, especially heterodimerization, and oligomerization of GPCRs can provide opportunity for novel drug discoveries and therapeutic ways.

The interaction of membrane-embedded proteins is achieved along the axis parallel to the membrane normal due to the limited rotations in the membrane [27]. Oligomerization of G protein-coupled receptors mostly depends on the interaction of transmembrane regions. On the other hand, there are two possible scenarios of GPCR dimerization, domain-swapped dimerization and contact dimerization. In the domain-swapped dimerization, a conformational change occurs in each protein structure, through which the substructure of one protein is replaced by that of the other protein. In the contact dimerization, the proteins interact with each other through an interface without any structural changes in the proteins [28]. A variety of experimental studies indicate that it is the transmembrane regions of GPCRs that considerably take part in dimerization and/or oligomerization [29, 30]. For example, adenosine (A2AR) and dopamine (D2R) receptors are known to form heteromer structures [31, 32]. Various experimental and computational studies carried out by different groups have indicated potential roles of transmembrane domains of the receptors in the A2A- D2 heteromeric receptor complexes, especially TM 5 of D2R [33, 34, 35]. Furthermore,

it has been shown that certain molecules which can be used as active ingredient of certain drugs can bind to GPCR dimers through TM domains [36].

1.1.1 Transmembrane Regions of GPCRs

GPCR oligomerization is considerably achieved through TM interfaces [37]. Transmembrane regions of G protein-coupled receptors can take part in the receptor interaction with ligand and membrane as well as dimerization and/or oligomerization. Ligand binding sites of a vast number of GPCRs are located within their TM helices [38]. A number of studies support that aromatic amino acids with the locations on these regions mediate ligand binding [39]. On the other hand, the receptors can contact with the plasma membrane consisting of hydrophobic lipid bilayer through their hydrophobic transmembrane regions. This interaction type ranges from weak to strong and contributes to the receptor stability and dynamics [40]. The transmembrane helices are oriented accordance with the membrane lipid bilayer thickness; in return, the membrane can regulate its thickness for TM positioning [41]. Not only with the lipid bilayer, but GPCRs can also directly interact with cholesterol embedded in lipid bilayer through a group of residues on TM regions and some of them have a well-defined cholesterol interaction site [42]. Moreover, studies conducted by various researchers had elucidated that a group of hydrophobic residues on transmembrane helices of rhodopsin and secretin GPCRs are included in G protein binding site [43, 44].

G protein-coupled receptors share a little sequence similarity within the same family between species [45]. However, GPCRs with the same functions share a conserved transmembrane topology pattern [46]. The sequence of transmembrane regions of GPCRs is mostly constituted by hydrophobic amino acids since they are embedded into the lipid membrane. On the other hand, the transmembrane regions also have polar and/or charged amino acids which can allow water molecules to permeate inside these regions because water is required for biomolecules to properly function [47]. Apart from interacting with water, polar amino acids within the transmembrane domains mediate TM-TM interaction (helix-helix contact) as well [39]. Interhelical interactions, supported by polar residues in TM regions, strongly influence the three-dimensional structure of the receptor [48].

Transmembrane regions of G protein-coupled receptors are probably the most significant part of the receptor in terms of both structure and function. GPCRs have a common secondary structure for their transmembrane regions dominated by hydrophobic residues in the core; alpha helices [49, 50]. This folding type is promoted by hydrophobic nature of the membrane in order to substantially exclude water [51]. On the other hand, polar residues can locate within the core of the transmembrane helices, which are moderately isolated from the membrane lipid [52].

1.2 Hydrophobicity Scales of Amino Acids

Each of all the amino acids has distinct hydrophobicity characteristics due to their own unique side chain. Different hydrophobicity values for each amino acid can be determined with different techniques because property of the amino acid is unique to the environment provided by each of the techniques [53]. Different hydrophobicity scales for the common twenty amino acids have been collected and published on the Chimera page of the website of University of California, kdHydrophobicity, wwHydrophobicity, hhHydrophobicity, mfHydrophobicity, and ttHydrophobicity [54]. The transfer free energy of each individual amino acid side chain between water and vapor, and the tendency of each individual amino acid to locate near the interior or the exterior of the protein structure have been used for the calculations of hydropathy values of amino acids in kdHydrophobicity (Kyte-Doolittle hydrophobicity) [55]. wwHydrophobicity (Wimley-White hydrophobicity)scale depends on the interaction energetics of each individual amino acid and lipid membrane, free energy of residue partitioning into electrically neutral membrane interfaces [56]. The study carried out by Wimley and White has also contributed to the determination of hhHydrophobicity scale for which, experimentally determined transfer free energies of each individual amino acid from water to POPC interface and to n-octanol, have been used separately [57, 58, 59]. In mfHydrophobicity (Moon-Fleming hydrophobicity), the twenty natural amino acids have been assigned with a hydrophobicity scale, which regards the transfer free energy of each amino acid side chain from water into lipid bilayer that resembles in structure to the cell membrane [60]. For ttHydrophobicity (Transmembrane Tendency hydrophobicity) scale, alias TM tendency scale, distribution of each

amino acid between soluble proteins and transmembrane sequences was examined by analyzing whole genomic data, and then the results were compared to a number of hydrophobicity scales [61]. Among all the hydrophobicity scales mentioned above, in hhHydrophobicity and mfHydrophobicity scales, more negative values represent the more hydrophobicity [54].

1.3 UniProt: Protein Database

UniProt (the Universal Protein Resource) is a freely accessible collection of datasets belonging to protein sequences, structural and functional information, and supporting data for proteins. The main dataset is UniProt Knowledgebase (UniProtKB) composed of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. The former provides reviewed, meaning experimantally derived and manually annotated by experts, information and the latter involves unreviewed, meaning computationally analyzed and annotated, information [62, 63].

1.3.1 Prediction of Topology Data of Membrane Proteins in UniProt

Currently, Uniprot employs TMHMM, Memsat, Phobius and the hydrophobic moment plot method of Eisenberg and coworkers to predict alpha-helical transmembrane regions of the proteins. UniProt annotates the predicted topology information of a helical membrane protein only if at least two prediction tools' outputs are consistent [64].

1.4 Literature Review

There are several computational methodologies in order to predict topology information of transmembrane proteins. Some of them utilize machine learning techniques such as Hidden Markov Model (HMM) or neural networks.
Tusnády et. al. has proposed a method, named Hidden Markov Model for Topology Prediction (HMMTOP) method, in order to predict the transmembrane topology of helical membrane proteins, which provides localization of membrane proteins. This

method is based on the hypothesis that there are different amino acid distributions on different structural parts of the proteins. It employs Hidden Markov Model. The model consists of five stages; inside loop, inside helix tail, membrane helix, outside helix tail and outside loop. They are characterized according to the amino acid distributions. The predictions are also constrained by the length of transmembrane region according to the width of the cell membrane which is composed of phospholipid bilayer. Tusnády et. al. asserted 79% accuracy of HMMTOP [65].

Other methods developed by Pasquier and Hamodrakas, which are PRED-TMR and PRED-TMR2, are available for prediction of transmembrane segments of the proteins. The algorithms are based on a standard hydrophobicity analysis which is for detecting potential termini of transmembrane regions. The propensity of each residue in the protein to be in a transmembrane segment is calculated using a statistical analysis with the frequency of the residue derived from the known protein data. The algorithm requires only sequence information. Unlike PRED-TMR, PRED-TMR2 has a pre-processing stage in which protein sequences are classified into membrane or non-membrane proteins. The classification is achieved by an artificial neural network. They have asserted 100% accuracy of the classification [66, 67].

Another method, named MEMSAT as old version and MEMSAT2 as a new version, has been proposed by Jones et. al. for prediction of helical membrane protein topology. The algorithm is based on dynamic programming. Expectation maximization is performed for each given model, through which the probability of the model to fit the known membrane protein data is calculated. Propensity of each amino acid to be in a helical transmembrane segment is calculated using the known protein database. They have established a 86-87% accuracy for G-protein coupled receptors by correctly predicting 13 out of 15 GPCRs [68].

Hirokawa et. al. have developed a method, named SOSUI, for prediction of transmembrane helices of membrane proteins and for discrimination of membrane and soluble proteins. The algorithm is based on hydrophobicity and uses Kyte-Doolittle hydropathy scale for calculations. It takes into account the average hydrophobicity of helix segments and the index of polar residues in those segments. They assume that amphiphilic residues are required for stabilization of the transmembrane regions. They have asserted more than 99% accuracy for classification of the proteins as membrane or soluble, and 97% accuracy for transmembrane helix prediction [69].

Bernhofer et. al. have proposed another method, named TMSEG, for prediction of transmembrane helices and and for discrimination of transmembrane helices and other proteins. This method combines evolutionary and empirical information of the proteins and machine learning algorithms which are Random Forest and neural network. The machine learning algorithms constitutes sliding windows of 19 consecutive residues from the protein sequences and predicts the central residue of each window to be either TM residue or non-TM residue or signal peptide residue. The formed windows are undergone multiple sequence alignment to obtain homology-based information. Empirical information used in the work involves average hydrophobicity value (i.e. Kyte-Doolittle hydrophobicity) of each window, and the percentages of positively charged, hydrophobic and polar residues in each window. They have claimed to have around 65% accuracy [70].

Another method, named TMpred, proposed by Hofmann and Stoffel, is used for prediction of transmembrane regions of the proteins and their orientation. The algorithm is based on sequence similarity and on grouping of the protein sequences according to the similarity. Similarity is calculated using statistical analysis with the known transmembrane proteins building weight-matrices for scoring. The method can be performed for not only cell membrane proteins, but also proteins of organelle membranes [71].

Kall et. al. have proposed the method, named Phobius, for prediction of transmembrane topology and signal peptide. One of the aims is the ability of the algorithm to discriminate hydrophobic transmembrane regions and hydrophobic signal peptides. The method is based on a Hidden Markov Model. The model is made up of three compartments; helix core, helix cytoplasmic end, and helix non-cytoplasmic end with different amino acid distributions. It takes into account helix length. The predictions are achieved by comparing amino acid distribution of the query sequence with those of the known transmembrane helices [72].

Sonnhammer et. al. have proposed another method, named TMHMM, for prediction of transmembrane protein topology. It is based on a Hidden Markov Model (HMM). The model is composed of the seven states; one for the helix core, two for caps on either side, one for loop on the cytoplasmic side, two for short and long loops on the non-cytoplasmic side, and one for globular domains in the middle of each loop, which have different amino acid distributions. The method combines HMM with TM hy-

Table 1.1: Available tools for prediction of topology information of the proteins

Tool	URL
HMMTOP	http://www.enzim.hu/hmmtop/
PRED-TMR	http://athina.biol.uoa.gr/PRED-TMR/
MEMSAT	http://bioinf.cs.ucl.ac.uk/software_downloads/memsat/
SOSUI	http://harrier.nagahama-i-bio.ac.jp/sosui/
TMSEG	https://www.predictprotein.org/
TMpred	https://embnet.vital-it.ch/software/TMPRED_form.html
Phobius	http://phobius.sbc.su.se/
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/

drophobicity, expected charged residues on TM helix and helix length. The method is trained with exact transmembrane helix boundaries. They have asserted 97-98% accuracy for the transmembrane helices [73, 74, 75].
The table 1.1 offers currently available tools which can be performed for prediction of helical transmembrane region of the proteins, and their urls.

1.5 Proposed Method

Within the study, the aim is the prediction of transmembrane alpha helices of G protein-coupled receptors.The proposed method uses hydrophobicity and secondary structures information about GPCRs. The reason for focusing on hydrophobicity is hydrophobic nature of the membrane; hence, naturally hydrophobic characteristics of the membrane proteins embedded in it [76]. Because mfHydrophobicity and kd-Hydrophobicity scales are graphically shown to be ones the most compatible with the actual transmembrane regions of the GPCRs with the known topology, they are used for estimation of each individual amino acid hydrophobicity, separately. For secondary structures of each protein, JPred is performed. The method is trained with the GPCRs with experimentally solved topology data derived from Uniprot. Its performance is evaluated both using cross validation and test data. Although the scientist

group developing TMHMM assert that its accuracy over the transmembrane helices is 97-98%, it fails to predict TM numbers of 6 proteins when we try it over the 29 GPCRs with known topology data, which causes low accuracy and low true positive rate [73]. With this study, we demonstrate that we can obtain more than 85% accuracy with higher true positive rate.

1.5.1 The Objective of the Proposed Method

The proposed method was developed for the prediction of transmembrane alpha helices of G protein-coupled receptors. Correctly predicting the transmembrane regions of the GPCRs can considerably contribute to the modelling the whole structure of those proteins. Moreover, the results can be utilized for studies on drug development and drug-protein interactions; hence, novel therapeutic ways. The method outputs can be used for protein-protein and protein-drug docking studies.

1.6 Contributions and Novelties

Our contributions with this study can be summarized as follows:

- Effects of hydrophobicity and secondary structures on determination of transmembrane regions of GPCRs were assessed.
- The distinct hydrophobicity scales in the literature were compared.
- The performances of the two machine learning algorithms, SMO and Random Forest, were compared and evaluated.
- An approach for prediction of GPCR transmembrane helices was developed.

1.7 The Outline of the Thesis

The thesis proceeds as follows. Chapter 1 gives general information about G protein-coupled receptors, and their characteristics. In the Chapter 2, the proposed method

was explained in detail. Chapter 3 involves the results with the discussions. The last chapter adresses conclusion and potential future works.

CHAPTER 2

MATERIALS AND METHOD

2.1 Background

In UniProt, the topology information of the unknown membrane proteins is supplied by the prediction tools. UniProt employs a number of prediction tools involving TMHMM, Memsat, Phobius and the hydrophobic moment plot method of Eisenberg and coworkers for prediction of helical transmembrane regions [64]. It was stated that TMHMM tool has a high accuracy, above 90% [73].

UniProt involves records for about 219000 proteins which are classified as G-protein-coupled receptor. However, only 29 of them have experimentally solved topology information. When running TMHMM with those 29 known GPCRs, we observed that it failed to predict the total number of transmembrane regions, which is normally expected to be 7 for GPCRs. The TMHMM tool gave 6 as the prediction output for 5 out of the known GPCRs while it determined to be 8 for one GPCR.

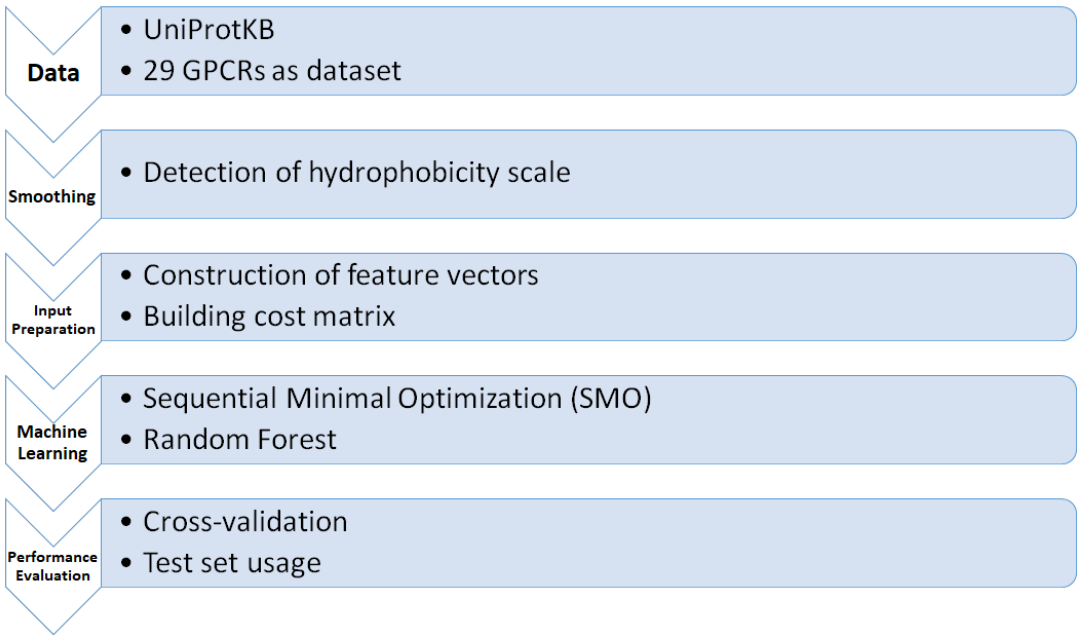
In this thesis study, the aim is to develop a more efficient algorithm for GPCR-specific TM prediction. We focused on the hydrophobicity of primary sequences and the secondary structures since it is known that alpha helical membrane proteins such as G protein-coupled receptors have long hydrophobic transmembrane helices [77].

The workflow of the study can be generalized as in the Figure 2.1.

2.2 Data Selection

To develop a prediction algorithm, a dataset is required as a reference. In this thesis study, the dataset was built with the G-protein-coupled receptors with experimentally

Figure 2.1: The workflow



solved topology information. They were used to train the proposed prediction algorithm. The proteins were selected from the UniProt database with the filters shown below:
"protein family: "g protein-coupled receptor" AND
subcellular location > topological domain evidence: "experimental" AND
reviewed: yes"

The 29 different GPCRs retrieved by the query above are shown in Table 2.1.

2.3 Smoothing the Data Set

At the beginning of the study, we had to determine which hydrophobicity scale to be used in the algorithm for further analyses. Therefore, we constructed a diagram for each known GPCR and examined the compatibility of the hydrophobicity scales with the actual hydrophobic transmembrane regions by plotting. Each residue in the protein sequence was replaced by the corresponding hydrophobicity constant so that a

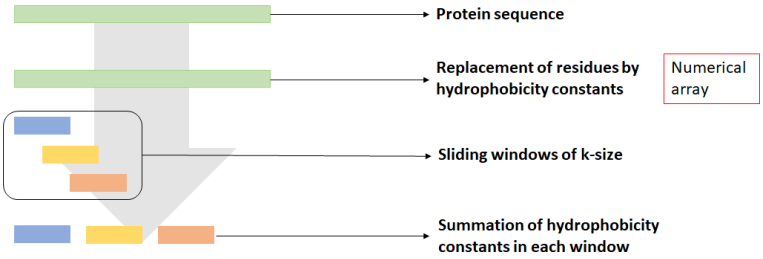
Table 2.1: GPCRs with experimentally solved topology information

Entry	Organism	Protein names
Q96PE1	Homo sapiens (Human)	Adhesion G protein-coupled receptor A2 (G-protein coupled receptor 124) (Tumor endothelial marker 5)
P29274	Homo sapiens (Human)	Adenosine receptor A2a
P07700	Meleagris gallopavo (Wild turkey)	Beta-1 adrenergic receptor (Beta-1 adrenoreceptor) (Beta-1 adrenoceptor) (Beta-T)
P51686	Homo sapiens (Human)	C-C chemokine receptor type 9 (C-C CKR-9) (CC-CKR-9) (CCR-9) (G-protein coupled receptor 28) (GPR-9-6) (CD antigen CDw199)
P08483	Rattus norvegicus (Rat)	Muscarinic acetylcholine receptor M3
P21554	Homo sapiens (Human)	Cannabinoid receptor 1 (CB-R) (CB1) (CANN6)
P41595	Homo sapiens (Human)	5-hydroxytryptamine receptor 2B (5-HT-2B) (5-HT2B) (Serotonin receptor 2B)
P32300	Mus musculus (Mouse)	Delta-type opioid receptor (D-OR-1) (DOR-1) (K56) (MSL-2)
P42866	Mus musculus (Mouse)	Mu-type opioid receptor (M-OR-1) (MOR-1)
P41146	Homo sapiens (Human)	Nociceptin receptor (Kappa-type 3 opioid receptor) (KOR-3) (Orphanin FQ receptor)
P61073	Homo sapiens (Human)	C-X-C chemokine receptor type 4 (CXC-R4) (CXCR-4) (FB22) (Fusin) (HM89) (LCR1) (Leukocyte-derived seven transmembrane domain receptor) (LESTR) (Lipopolysaccharide-associated protein 3) (LAP-3) (LPS-associated protein 3) (NPYRL) (Stromal cell-derived factor 1 receptor) (SDF-1 receptor) (CD antigen CD184)
P20789	Rattus norvegicus (Rat)	Neurotensin receptor type 1 (NT-R-1) (NTR1) (High-affinity levocabastine-insensitive neurotensin receptor) (NTRH)
P21453	Homo sapiens (Human)	Sphingosine 1-phosphate receptor 1 (S1P receptor 1) (S1P1) (Endothelial differentiation G-protein coupled receptor 1) (Sphingosine 1-phosphate receptor Edg-1) (S1P receptor Edg-1) (CD antigen CD363)
P31356	Todarodes pacificus (Japanese flying squid) (Ommastrephes pacificus)	Rhodopsin
O43614	Homo sapiens (Human)	Orexin receptor type 2 (Ox-2-R) (Ox2-R) (Ox2R) (Hypocretin receptor type 2)
P08100	Homo sapiens (Human)	Rhodopsin (Opsin-2)
P43220	Homo sapiens (Human)	Glucagon-like peptide 1 receptor (GLP-1 receptor) (GLP-1-R) (GLP-1R)
P55085	Homo sapiens (Human)	Proteinase-activated receptor 2 (PAR-2) (Coagulation factor II receptor-like 1) (G-protein coupled receptor 11) (Thrombin receptor-like 1) [Cleaved into: Proteinase-activated receptor 2, alternate cleaved 1; Proteinase-activated receptor 2, alternate cleaved 2]
P47900	Homo sapiens (Human)	P2Y purinoceptor 1 (P2Y1) (ADP receptor) (Purinergic receptor)
P02699	Bos taurus (Bovine)	Rhodopsin
Q13255	Homo sapiens (Human)	Metabotropic glutamate receptor 1 (mGluR1)
P41594	Homo sapiens (Human)	Metabotropic glutamate receptor 5 (mGluR5)
O43613	Homo sapiens (Human)	Orexin receptor type 1 (Ox-1-R) (Ox1-R) (Ox1R) (Hypocretin receptor type 1)
P21730	Homo sapiens (Human)	C5a anaphylatoxin chemotactic receptor 1 (C5a anaphylatoxin chemotactic receptor) (C5a-R) (C5aR) (CD antigen CD88)
P35462	Homo sapiens (Human)	D(3) dopamine receptor (Dopamine D3 receptor)
P21917	Homo sapiens (Human)	D(4) dopamine receptor (D(2C) dopamine receptor) (Dopamine D4 receptor)
O14842	Homo sapiens (Human)	Free fatty acid receptor 1 (G-protein coupled receptor 40)
P47871	Homo sapiens (Human)	Glucagon receptor (GL-R)
P35367	Homo sapiens (Human)	Histamine H1 receptor (H1R) (HH1R)

numerical array was obtained for each protein sequence. Due to 5 different hydrophobicity scales, we generated 5 separate numerical arrays for each protein sequence. In the diagram, x-axis refers residue number in the protein sequence, y-axis refers the hydrophobicity constant of each residue. However, since the output was too noisy to infer any biological result, smoothing method was implemented.

Smoothing is a method decreasing sampling error by reduction in noisy data and variation; moreover, sliding window one residue by one residue allows to evaluate all possible arrays of given size in the sequence [78]. The hydrophobicity data was smoothed by assigning the sum of the hydrophobicity values in a sliding window of size k to the amino acid centered in that window. Smoothing method is schematically represented in the Figure 2.2.

Figure 2.2: Smoothing method



For smoothing, we defined the window size as 3-residue, 5-residue, 7-residue, 15-residue, and 20-residue and found that a window size of $k = 7$ worked best in our study.

On each diagram, individual hydrophobicity scale, the actual TM regions derived from UniProt, and the TM regions predicted by TMHMM were depicted along each sequence. Therefore, such issues as which parts of the protein sequence are hydrophobic, how the actual TMs and the predicted TMs are compatible with each other, how the hydrophobicity scales are consistent with the hydrophobic TM regions and which hydrophobicity scale is the most appropriate for further analysis could be observed. The results indicated that mfHydrophobiciy and kdHydrophobicity are the most compatible scales with the actual hydrophobic TM regions.

2.4 Data Preparation

A specific group of features must be identified for training a machine learning algorithm. Hydrophobicity and secondary structure were determined as features in this thesis study because it is known that G protein-coupled receptors have long hydrophobic transmembrane helices [77].

FASTA format of each protein sequence was retrieved from UniProtKB. We assigned the corresponding hydrophobicity constant to each individual amino acid along the sequence; therefore, each residue in the sequence was represented by an associated numerical value. A numerical array was formed instead of protein sequence. Using both kdHydrophobicity and mfHydrophobicity scales separately, we could obtain two separate numerical arrays for each protein sequence.

The secondary structures of the proteins were formed using the JPred prediction server, because it has a high accuracy with 82% [79]. In the JPred prediction results, each residue building an alpha helix structure is depicted by 'H', those involved in extended structures are symbolized by 'E', and all remaining residues are denoted by dashed lines, '-' [80, 81]. We marked each residue,which is observed to be a component of helix regions, as '1' and the remaining part as multiple '0's [82]. Therefore, a binary array was obtained with the same size of the protein sequence.

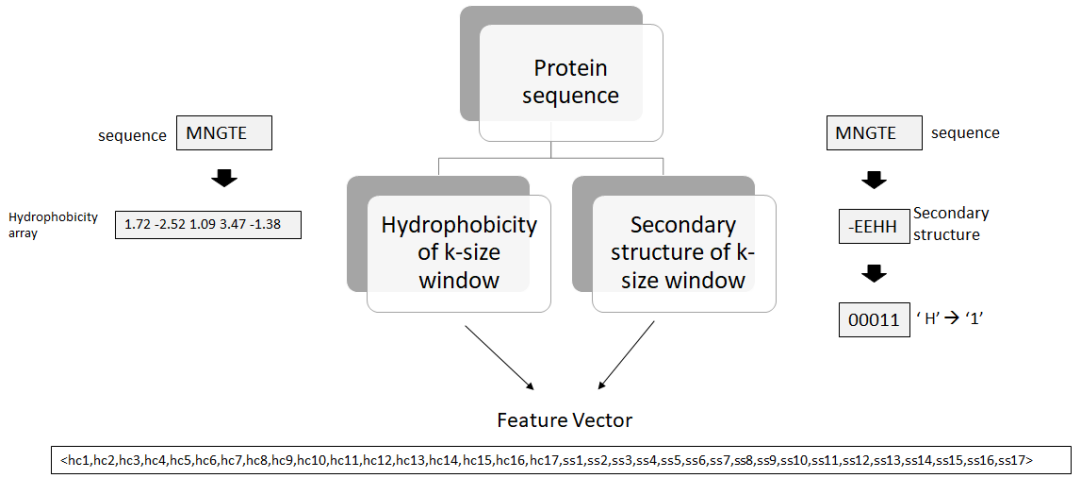
We tried to learn machine learning algorithms with the quantitative data.

The data preparation is schematically represented in the Figure 2.3. In the figure, a feature vector which was constructed of 17-residue sliding window is exemplified.

2.5 Baseline Method

Before running the proposed method with the data set, a baseline method was performed. A threshold was set to the hydrophobicity values which were smoothed. In our study, baseline method is the most basic method taking into account only hydrophobicity. TM start points were detected, whose hydrophobicity values were value of the threshold. The results of the proposed method were compared with the output of the baseline method.

Figure 2.3: The data preparation



2.6 Method

A sliding window approach was used because sliding window one residue by one residue allows to evaluate all possible arrays of given size in the sequence [78]. Given a specific size of a window as an input (i.e., as a parameter), the window was moved across the whole protein sequence one residue by one residue. Thus, $(N - k + 1)$ windows can be obtained if the sequence has N residues and the subsequence, or window, size is k.

We used three different window sizes of 11-residue, 17-residue, and 21-residue in order to be able to compare them and to select the best size for further analysis. For each protein in the data set, we performed the sliding window method along the related numerical arrays, one of which was composed of hydrophobicity constants and the other of which was the associated binary array representing the secondary structure. We implemented each of the three sliding windows with different sizes for each of the two hydrophobicity scales. We classified samples whose center residue was the actual starting point of the GPCR transmembrane region as "True" and all other samples as "False". It was expected that there are totally 203 "True"s in each data set due to existence of the 29 GPCRs with 7 TM regions.

We repeated the procedure for TM end points with optimal features involving 17-residue windows and mfHhydrophobicity to create a separate database.

The training sets of TM start points built using 11-residue windows were composed of 13499 samples, i.e., subsequences, 203 of which were True as expected. The training sets built using 17-residue windows were composed of 13325 samples, 202 of which were True. The training sets constructed using 21-residue windows were composed of 13209 samples, 201 of which were True.

The number of "True" samples was less than the expected in the cases of 17-residue windows and 21-residue windows, because the indexes of some TM start residues did not match those of center residues of the windows due to length of the window sizes; thus, they could not be involved in any window center. Different from TM start data sets, the number of minority classes in TM end data sets was as expected, 203.

Taking into account any potential experimental error rates, we created additional training sets by classifying the subsequences whose center residue was either the actual TM start point or the next residues on the both sides of the actual TM start point as "True" and the others as "False". In this scenario, 21 True samples for each protein was expected. In order to obtain more balanced classes, another training set was built. We classified subsequences whose centered residue is either the actual starting point of TM or one of the two residues next to the actual start residue of TM as "True" on both sides and the others as "False". In this scenario, 35 True samples for each protein was expected. We performed this procedure only on TM start data sets.

The training sets consisted of "True" and "False" classes. The machine learning algorithms, SMO and Random Forest, were run with the training data sets. The confusion matrices were created.

In the study, we generated 18 different training sets with different feature vectors for TM start point prediction and 6 different training sets with different feature vectors for TM end point prediction.The training sets are described in the Table 2.2 and the Table 2.3.

2.6.1 Cost Matrix

In this thesis study, totally 24 distinct training sets were created, all of which had a "True" class with much fewer samples than the "False" class since sliding window method generated thousands of samples but only a few of them could be classified as

Table 2.2: The training sets generated for the TMstart prediction algorithms

Training set	Hydropathy Scale	Secondary Structure	Window Size	True Class Size
Training set 1	mf	jPred	11	203
Training set 2	mf	jPred	11	609
Training set 3	mf	jPred	11	1015
Training set 4	kd	jPred	11	203
Training set 5	kd	jPred	11	609
Training set 6	kd	jPred	11	1015
Training set 7	mf	jPred	17	202
Training set 8	mf	jPred	17	606
Training set 9	mf	jPred	17	1010
Training set 10	kd	jPred	17	202
Training set 11	kd	jPred	17	606
Training set 12	kd	jPred	17	1010
Training set 13	mf	jPred	21	201
Training set 14	mf	jPred	21	603
Training set 15	mf	jPred	21	1006
Training set 16	kd	jPred	21	201
Training set 17	kd	jPred	21	603
Training set 18	kd	jPred	21	1006

"True". A classifier is biased toward the minority class [83]. In the literature, there are two ways to compensate imbalanced classification; resampling and cost-sensitive approaches. In our study, due to the inadequate amount of data, we preferred to use a cost-sensitive technique to manage highly imbalanced data [84]. Most standard classification algorithms have poor performance in detection of the minority class. Cost-sensitive learning assigns different weights to each data point to minimize the failure in prediction [85]. In this thesis study, the majority class was the "False" class and the algorithm might tend to predict any given instance as False. Therefore, we introduced a penalty for prevention of any misclassification by using the cost matrix. We determined the penalty values according to rate of the sizes of majority and minority classes. For example, in the data set with 1 True per TM, we described the penalty as 65 for misclassification of any instance in the actual "True" class while

Table 2.3: The training sets generated for the TMend prediction algorithms

Training set	Hydropathy Scale	Secondary Structure	Window Size	True Class Size
Training set 1	mf	jPred	11	203
Training set 2	mf	jPred	17	203
Training set 3	mf	jPred	21	203
Training set 4	kd	jPred	11	203
Training set 5	kd	jPred	17	203
Training set 6	kd	jPred	21	203

we determined the penalty as 1 for misclassification of any instance in the actual "False" class since the size of the actual "False" class is 65-times of the actual "True" class.

2.7 WEKA

Machine learning is a computation type that can promptly characterize patterns in given data. WEKA is an open-source data-mining software tool that includes a vast number of machine learning algorithms for classification, regression, clustering, and association mining problems [86].

2.7.1 Classification Methods

Classification strategies can be described as solving the given problem by training themselves with a training data set, predicting the test set using the solution and assigning it to one of the predefined classes. Logistic regression, SVM, Random Forest and Decision Tree can be given as examples of the most popular classification methods [87].

Table 2.4: The cost matrices

(a) The cost matrix for 1 True per TM

		Actual TMs	
		TM	non-TM
Predicted TMs	TM	0	1
	non-TM	65	0

(b) The cost matrix for 3 True per TM

		Actual TMs	
		TM	non-TM
Predicted TMs	TM	0	1
	non-TM	21	0

(c) The cost matrix for 5 True per TM

		Actual TMs	
		TM	non-TM
Predicted TMs	TM	0	1
	non-TM	12	0

2.7.1.1 SMO and Random Forest as Classification Approaches

SMO (Sequential Minimal Optimization) is a SVM (Support Vector Machine), which is mostly performed for the problems with two classes [88, 89]. SVM creates the solution with a weighted linear combination of the samples in the training set [90]. Therefore, it can recognize irrelevant attributes and increase the final accuracy of the algorithm [89]. SMO can deal with the Quadratic Programming (QP) problem, which is the minimization of a quadratic functional subject to linear constraints, arising during training [91]. SMO partitions the large QP problem into the smaller sub-problems and solves the smallest one at every step [92, 93]. On the other hand, Random Forest is another classification algorithm which consists of many weakly-correlated decision

trees. Each individual tree predicts, or votes, the possible classification of a given input, and then the final result is identified as the correct classification [94, 95]. Random Forest is capable of balancing error in unbalanced data sets [96].

2.7.2 Evaluation of Algorithm Performance

Cross validation is an evaluation method applied for assessing performance of a classifier. For N-fold cross validation, the whole data is split into N components. The classifier is trained with $(N-1)$ parts and the left one part is used as test set to evaluate the trained classifier. This is repeated N times unless any data the in test set is remained as unpredicted [97]. In this thesis study, 10-fold cross validation was an evaluation method. We permitted the Weka to randomly split the training set. Apart from cross validation, the other method was using test set. We divided the data set into training and test sets. For each time, the test set was made up of the subsequences of one of the GPCR which was incorrectly predicted by TMHMM. We evaluated the resultant confusion matrices.
The structure of a confusion matrix is illustrated in the Table 2.5 [98].

Table 2.5: The representation of a confusion matrix

		Actual classes	
		True	False
Predicted classes	True	TP	FP
	False	FN	TN

2.7.2.1 Parameters for Evaluating Algorithm Performance

In order to assess the model performance, there are a number of parameters involving true positive, true negative, false positive, false negative, and accuracy [99, 98].

- True positive (TP) is the outcome which is the real positive and predicted as positive.

- True negative (TN) is the outcome which is the real negative and predicted as negative.
- False positive (FP) is the outcome which is the real negative but predicted as positive.
- False negative (FN) is the outcome which is the real positive but predicted as negative.
- Accuracy is the rate of total true positive and true negative instances to total sample number.
- Recall, true positive rate (TPR), or sensitivity, is the proportion of total true positives to sum of total true positives and total false negatives.
- False positive rate (FPR) is the proportion of total false positives to sum of total true negatives and total false positives.
- MCC (Matthews correlation coefficient) is the correlation between actual and predicted classifications and ranges from -1 to +1. The worst value is 1 and the best value is +1. The Equation 21 shows how to calculate MCC.

$$\mathbf{MCC} = \frac{(\mathbf{TP} * \mathbf{TN}) - (\mathbf{FP} * \mathbf{FN})}{\sqrt{(\mathbf{TP} + \mathbf{FP}) * (\mathbf{FN} + \mathbf{FP}) * (\mathbf{TN} + \mathbf{FP}) * (\mathbf{FN} + \mathbf{TN})}} \quad (21)$$

In this thesis study, a true positive could be identified as the sample, or subsequence, whose centered residue was an actual TM start point and the algorithm also predicted it as actual TM start point. A true negative could be identified as the sample, or subsequence, whose centered residue was not an actual TM start point and the algorithm also predicted it as non-TM start point. A false positive could be identified as the sample, or subsequence, whose centered residue was not an actual TM start point but the algorithm also predicted it as TM start point. A false negative could be identified as the sample, or subsequence, whose centered residue was an actual TM start point but the algorithm also predicted it as non-TM start point.

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Data

For the this thesis study, we identified the GPCRs with the experimentally solved topology information, totally 29 GPCRs. We used these receptors as the dataset to train the machine learning algorithms. Because the aim of the study is to develop a more efficient algorithm than TMHMM to predict the topology information for GPCRs, we run TMHMM prediction tool with these known GPCRs to observe its efficiency. Although a GPCR has naturally 7 transmembrane regions, TMHMM gave 6 as the prediction output for 5 out of them while it determined to be 8 for one of these GPCRs.

3.2 Smoothing Data

We constructed graphics for the known GPCRs showing the hydrophobicity scales, the TM regions as output of TMHMM and the actual TM regions to be able to clearly observe which hydrophobicity scales are consistent with the actual hydrophobic TM regions, and to compare the actual TM areas and the predicted TM areas by TMHMM. Each residue in the protein sequence was replaced by the hydrophobicity value of each scale and the graphics consists of these one residue by one residue values. The graphics can be found in `https://github.com/mzzclb/GPCR-TM-Prediction/issues/6`. However, they were too noisy to infer any biological result. That is why to use smoothing method. Smoothing is a method decreasing sampling error by reduction in noisy data and variation [78]. To find the most

suitable and informative one, we tried different window sizes; 3-residue, 5-residue, 7-residue, 15-residue, and 20-residue, and the graphics were constructed using these window sizes for smoothing and the graphics for all of the proteins are available in the following links:

- https://github.com/mzzclb/GPCR-TM-Prediction/issues/2
- https://github.com/mzzclb/GPCR-TM-Prediction/issues/3
- https://github.com/mzzclb/GPCR-TM-Prediction/issues/4
- https://github.com/mzzclb/GPCR-TM-Prediction/issues/5

The graphics can be examined in detail and the window sizes can be compared. 7-residue was the most informative graphics so we used the graphics with 7-residue windows. The sliding windows whose sizes are less than 7 lead to noisy data and variation, and those whose sizes are larger than 7 may result in lack of data observed in the graphics.

3.2.1 Hydrophobicity Scales

Numerical values of the hydrophobicity scales which are kdHydrophobicity, mfHydrophobicity, ttHydrophobicity, wwHydrophobicity, and hhHydrophobicity are given in the Table A.1 in the Appendix A.

3.3 Mispredictions of TMHMM

The proteins with the experimentally solved topology information were displayed as graphical representations. Each individual graphic includes information of actual TM regions derived from UniProt, TMHMM prediction results, helix regions as secondary structures and the hydrophobicity scales along each protein sequence. On the other hand, the graphics of the GPCRs whose total TM number is incorrectly predicted contain also the results of the proposed algorithms. TMHMM incorrectly predicted the total TM number of 6 GPCRs.

There are several reasons for incorrect identification of the prediction tools, stated in literature. Firstly, the last TM regions of GPCRs are relatively less hydrophobic than the other TM domains; thus, the prediction tools may fail to detect the last TM [100]. On the other hand, according to the study mentioned in [101], presence of signal peptides considerably affects TM prediction, and it was observed that TMHMM accuracy increased after removal of signal peptide sequence from the protein sequence in that study. Signal peptides are short hydrophobic sequences of amino acids located at N-terminus of the protein, which play role in trafficking of the protein to its destination, cell membrane for GPCRs, and is cleaved after localization [102]. Besides, some GPCR molecules have an eighth amphipathic helix located at the C-terminus. They have a hydrophobic core. The functional role of these short helices is recognition of membrane surface, stabilization and orientation [103]. Presence of this helix may misguide the prediction tools. It is stated that most of the TM prediction tools, including TMHMM, tend to incorrectly identify hydrophobic residue clusters in protein sequences as helical transmembrane segments, that can increase false positive rate [104]. After all, another reason for the failure of TMHMM predictions may be that TMHMM has an inclination to describe TM length as 21 residues [77]. Therefore, it can fail to identify transmembrane regions with larger lengths.

3.4 Baseline Method

A threshold was set as a hydrophobicity value for a residue to be a TM start point. This threshold was applied in the smoothed data. In this thesis study, the baseline method takes into account only hydrophobicity. We tried to determine how hydrophobicity alone has effect on determination of transmembrane regions; thus, to determine how many actual transmembrane start point has the threshold as the hydrophobicity value.

To determine the best threshold, we tried different threshold values. The threshold and their evaluation parameters can be observed in the Table 3.1. Both mfHydrophobicity and kdHydrophobicity were utilized separately. For mfHydrophobicity, -4 was the threshold providing the highest MCC, 0.042 while for kdHydrophobicity, 5 was

Table 3.1: Evaluation results of the baseline

Hydrophobicity	Threshold	MCC	TP rate	Accuracy
kd	5	0.043	0.089	95.8%
kd	6	0.000	0.035	95.2%
kd	7	0.008	0.039	95.7%
kd	8	0.039	0.089	95.5%
kd	9	0.033	0.069	96.1%
kd	10	0.009	0.0394	95.9%
mf	-5	0.005	0.035	95.8%
mf	-4	0.042	0.044	94.9%
mf	-3	0.032	0.089	94.9%
mf	3	0.007	0.064	93.6%
mf	4	0.001	0.049	93.9%
mf	5	0.011	0.079	93.0%

the threshold providing the highest MCC, 0.043. MCC was required to evaluate the thresholds because the data was highly imbalanced.
Generally, accuracy was high. The reason could be high true negatives.

3.5 Parameters and Algorithms for TM Start Residue

In order to develop an efficient prediction algorithm, a number of parameters was used. Firstly, mfHydrophobicity and kdHydrophobicity were determined to be the most significant as the hydrophobicity scale after examining all of the graphics of the known GPCRs, and it was decided to use them separately. The second parameter was optimal length of subsequences creating feature vectors. We used 11-residue, 17-residue, and 21-residue, separately, and the most effective one to find a pattern was 17-residue. Determination of the size is critical for the algorithm performance. Although small window accelerates the algorithm, accuracy performance may decreases due to lack of data; on the contrary, large window may increase latency [105]. In general, TP rate was the highest for 11-residue of window size due to lack of data and

21-residue of window size led noisy data for a pattern. A feature vector for the optimal method is composed of 17-residue of window from numerical hydrophobicity array and 17-residue of window from binary secondary structure array. After determination of optimal feature vector size, it was noticed that any experimental error could result in miscalculation of TM start point. By taking into account any errors, we decided to identify the residues next to the TM start point as "True", that means they also can be TM start point. The next residues on the both sides and the next two residues on the both sides were used separately. We run the machine learning algorithms, SMO and Random Forest, with the models derived from the different combinations of these parameters separately. In order to evaluate the algorithms, 10-fold cross validation was performed, in which the training sets were randomly split. The Tables 3.4, 3.5 and 3.6 show the cross-validation results for each model with cost-sensitive approach.

Due to highly imbalanced data and minority of "True" class, MCC is one of the most important evaluation parameters [106]. The machine learning algorithms have a tendency to predict any given instance to be "False" since they aim to minimize the overall error rate without paying any special attention to the minority class [107]. Therefore, true positive rate might be more important than total accuracy for our study. Another performance metric is false positive rate. How many of non-TM residues were predicted as TM start point is important. Penalty should not negatively affect decision making mechanism of the algorithm. However, in this study, highly imbalanced data forced us to take into consideration only MCC.

Most standard classification algorithms have a poor performance in detection of the minority class. Cost-sensitive learning assigns different weights to each data point to minimize the failure in prediction [85]. In this thesis study, the majority class is the "False" class and the algorithm may tend to predict any given instance as False. Therefore, we introduced a penalty for prevention of any misclassification by using the cost matrix. The effect of cost-sensitive approach on the proposed algorithm can be observed in Table 3.2. The cost-sensitive approach was implemented to the model having mfHydrophobicity, 17-residue as window size and 1 "True" for each TM start residue. Without any penalty for misclassification, the algorithm tended to predict any given instance as "False" and none of given samples as "True"; thus, the algorithm without cost-sensitive approach has 0 TP rate despite of higher accuracy than that with cost-sensitive approach. It can be inferred that cost-sensitive approach is a

Table 3.2: Effect of cost-sensitive approach

	With Cost-sensitive	**Without Cost-sensitive**
TPR	0.891	0.000
Accuracy	86.57%	98.48%

significant method for handling with imbalanced data.

We tried to evaluate TMHMM results performing the proposed method. We used the optimal model having mfHydrophobicity, 17-residue as window size and 1 True for each TM start residue with cost-sensitive approach. The resultant confusion matrix is given in the Table 3.3. According to the outcomes, the reason for very low true positive rates despite of the high accuracy can be clarified with the large number of true negative instances. TMHMM also has a tendency to identify a given instance as in the majority, "False" class.

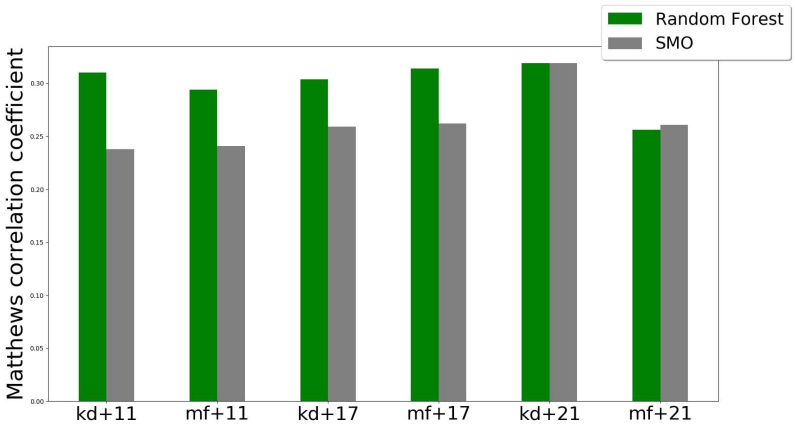
When observing the Table 3.4, a number of inferences can be made. Several studies claim that Random Forest has a high prediction performance [108, 109]. The Figure 3.1 shows Matthews correlation coefficients of the Random Forest and SMO performed with the models. MCC is a significant parameter to evaluate performance of a prediction algorithm trained with a imbalanced data and our data set is highly imbalanced. In general, Random Forest had a higher MCC value, which means that Random Forest can more efficiently handle imbalanced data. The reason why false positive rate was higher in SMO can be that giving penalty for misclassification of

Table 3.3: Evaluation of TMHMM results with our algorithm

		Actual TMs	
		TM	non-TM
Predicted TMs	TM	48	149
	non-TM	154	12974

TP rate = 0.238, accuracy = 97.7%, MCC = 0.231

Figure 3.1: MCC of SMO and Random Forest



minority class negatively affected the predictions of the algorithm. SMO avoided to identify the given sample as "False". MCC increased as minority class size increases, and Random Forest had a higher MCC for all the models. This can be explained with that Random Forest, comprised of many decision trees, had efficient tree classifiers in this study and good classification results could be obtained [109]. However, because it is unknown that any mistake during experimental procedures of hydrophobicity scales, identifying the next residues as "True" may misguide the algorithm.

When comparing the Table 3.4 with the Table 3.5, it can be proposed that mfHydrophobicity could better reflect the natural hydrophobic environment of the proteins embedded in the membrane. The experimental procedure for calculation of mfHydrophobicity took into account both hydrophilic and hydrophobic environment and this scale was calculated with transient of amino acid between water and lipid bilayer which can mimic the cell membrane.

The results of Random Forest are more practical and promising in our study.

TM end points were used as the other training set for learning algorithms. In the results, it can be observed that both SMO and Random Forest have less accuracy and true positive rates for TM end residues. The transmembrane start point can be characterized as the residue where the protein sequence enters the cell membrane from the extracellular environment and the end point of the transmembrane area is the residue where the protein sequence passes from the cell membrane inside the cell.

Table 3.4: Cross-validation results of the models with mf for TM starts

	TPR	FPR	Accuracy	MCC
mf+11-residue+1True+SMO	0.877	0.150	85.08%	0.241
mf+11-residue+1True+RF	0.192	0.003	98.46%	0.294
mf+17-residue+1True+SMO	0.891	0.135	86.57%	0.262
mf+17-residue+1True+RF	0.213	0.004	98.46%	0.314
mf+21-residue+1True+SMO	0.881	0.132	86.79%	0.261
mf+21-residue+1True+RF	0.199	0.003	98.48%	0.309
mf+11-residue+3True+SMO	0.846	0.134	86.50%	0.397
mf+11-residue+3True+RF	0.470	0.020	95.66%	0.472
mf+17-residue+3True+SMO	0.894	0.135	86.62%	0.421
mf+17-residue+3True+RF	0.535	0.022	95.81%	0.515
mf+21-residue+3True+SMO	0.900	0.136	86.57%	0.424
mf+21-residue+3True+RF	0.556	0.022	95.90%	0.531
mf+11-residue+5True+SMO	0.842	0.140	85.85%	0.469
mf+11-residue+5True+RF	0.640	0.036	94.00%	0.584
mf+17-residue+5True+SMO	0.901	0.123	87.84%	0.533
mf+17-residue+5True+RF	0.708	0.037	94.39%	0.628
mf+21-residue+5True+SMO	0.902	0.123	87.86%	0.534
mf+21-residue+5True+RF	0.726	0.036	94.59%	0.644

This difference could yield a different pattern for TM end prediction. The proposed algorithm is more successful in TM start residue prediction in overall appearance.

3.6 Baseline, TMHMM and the Proposed Method

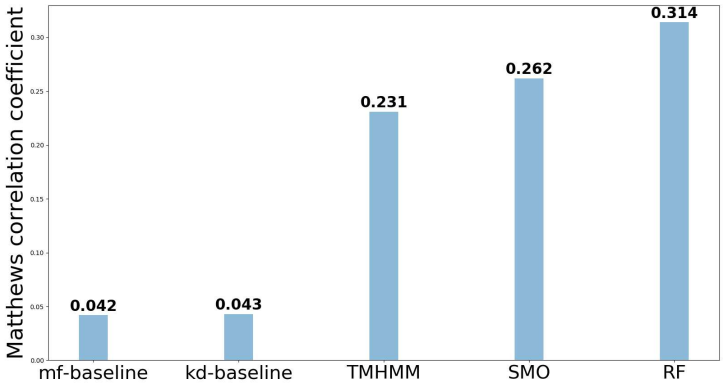
In order to compare baseline, TMHMM, SMO and RF, we used the same model to perform them separately. The model involved 17-residue windows, mfHydrophobicity and regarded only the actual TM start points as "True", not the next residues. The Figure 3.2 clearly demonstrates the comparison of true positive rates of them. According to the results, Random Forest has the highest MCC while the baseline method

Table 3.5: Cross-validation results of the models with kd for TM starts

	TPR	**FPR**	**Accuracy**	**MCC**
kd+11-residue+1True+SMO	0.877	0.153	84.72%	0.238
kd+11-residue+1True+RF	0.212	0.004	98.46%	0.310
kd+17-residue+1True+SMO	0.901	0.141	85.99%	0.259
kd+17-residue+1True+RF	0.193	0.003	98.48%	0.304
kd+21-residue+1True+SMO	0.891	0.140	86.00%	0.256
kd+21-residue+1True+RF	0.204	0.003	98.50%	0.319
kd+11-residue+3True+SMO	0.831	0.139	86.00%	0.383
kd+11-residue+3True+RF	0.504	0.019	95.97%	0.510
kd+17-residue+3True+SMO	0.891	0.138	86.38%	0.416
kd+17-residue+3True+RF	0.551	0.021	95.96%	0.532
kd+21-residue+3True+SMO	0.896	0.139	86.21%	0.416
kd+21-residue+3True+RF	0.526	0.020	95.90%	0.518
kd+11-residue+5True+SMO	0.834	0.143	85.52%	0.460
kd+11-residue+5True+RF	0.652	0.034	94.24%	0.599
kd+17-residue+5True+SMO	0.893	0.126	87.53%	0.523
kd+17-residue+5True+RF	0.709	0.035	94.52%	0.634
kd+21-residue+5True+SMO	0.901	0.129	87.35%	0.525
kd+21-residue+5True+RF	0.720	0.036	94.56%	0.641

has the lowest one. It can be said that hydrophobicity alone is not sufficient to determine transmembrane regions. On the other hand, when examining RF and SMO more closely, it can be noticed that RF can more efficiently handle the imbalanced data than SMO. Apart from these, the diagram reveals that TMHMM performs with the imbalanced data less efficiently than our methods, SMO and RF, but much more efficiently than the baseline method. It can be deduced that like stated in literature, TMHMM does not regard only hydrophobicity for predictions.

Figure 3.2: MCC of baseline, TMHMM, SMO and RF



3.7 The Known GPCRs

The graphics constructed for each known GPCR have all of the hydrophobicity scales in different colors, actual TM regions as dark blue lines, TM regions predicted by TMHMM as red lines, and secondary structures; helices as blue circles and the other secondary structures as pink dots along the sequence. We performed the proposed method, SMO and Random Forest, with the 6 GPCRs whose total TM number was mispredicted by TMHMM. In the graphics, right-facing purple triangles represent TM start points predicted by SMO. Left-facing triangles represent TM end points predicted by SMO. Upside down green triangles represent TM start points predicted by Random Forest. The black triangles represent TM end points predicted by Random Forest. In general, Random Forest tended to make less TM predictions which was forced by penalty. All the graphics of the known GPCRs are available in `https://github.com/mzzclb/GPCR-TM-Prediction/issues/1`. Hydrophobicity

Table 3.6: Cross-validation results of the models with mf for TM ends

	TPR	**FPR**	**Accuracy**	**MCC**
mf+11-residue+1True+SMO	0.803	0.195	80.51%	0.184
mf+17-residue+1True+SMO	0.808	0.178	82.17%	0.198
mf+21-residue+1True+SMO	0.852	0.173	82.74%	0.216

of the proteins was generated using 7-residue sliding windows. The graphics can be examined in detail.

3.7.1 GPCRs Having 8 TMs as TMHMM Result

Human proteinase-activated receptor 2 (P55085) is a GPCR molecule whose topology information is experimentally solved. It belongs to the rhodopsin family. TMHMM gave 8 as total TM number for those proteins. The predicted extra TM was found to be located at the precedent site of the actual first TM. Around the predicted region, there is a very short, less hydrophobic, helix. On the other hand, the proposed method performing SMO identified a TM start point at the beginning of that region while Random Forest made no prediction around that region. The predicted TM region and the results of the proposed method can be observed in the Figure 3.3.

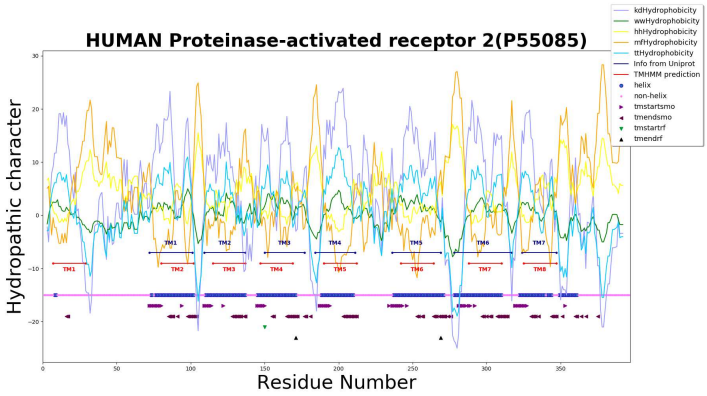


Figure 3.3: GPCR with an extra TM predicted by TMHMM

3.7.2 GPCRs Having 6 TMs as TMHMM Result

TMHMM identified 6 TM regions for 5 GPCRs out of the known GPCRs.

3.7.2.1 The Last TM Missing

Human orexin receptor type 1 (O43613), todpa rhodopsin (P31356) and human P2Y purinoceptor (P47900) are GPCR molecules whose topology information is experi-

mentally solved. These proteins belong to the rhodopsin family. TMHMM failed to predict the last actual TM region. Although at least a helix structure is observed in that region, they are less hydrophobic according to the hydrophobicity scales. Therefore, TMHMM could not predict them as TM, meaning that TMHMM does not take into account only secondary structure. For human orexin receptor type 1 and human P2Y purinoceptor, SMO achieved to detect TM7 start and end points, separately, while Random Forest could identify only actual TM7 start point of the former receptor and only actual TM7 end point of the latter. For todpa rhodopsin, SMO achieved to detect actual TM7 start and end points. However, it identified the start and end points of other 2 non-TM helices located at the end of the protein sequence. Random Forest failed to make any predictions for the actual last TM region. The graphics of all the three proteins can be found in the Figure 3.5.

3.7.2.2 The Third TM Missing

Rat neurotension receptor type 1 (P20789) and human metabotropic glutamate receptor 5 (P41594) are GPCR molecules whose topology information is experimentally solved. While neurotension receptor type 1 is a rhodopsin family protein, metabotropic glutamate receptor 5 belongs to the glutamate family. When examining closely, it can be observed that TMHMM predictions lacks the actual third TM region for both of the proteins. According to the JPred results, a long helix structure is present in that region of each protein but they are less hydrophobic. On the other hand, for both the proteins, SMO achieved to detect TM3 start and end points, separately, while Random Forest could identify only TM3 end point. The Figure 3.5 includes the graphics of both of the proteins to be able to examine in detail.

3.7.3 GPCR with Incorrect Prediction of TM Region Locations

Human D(4) dopamine receptor (P21917) is a GPCR molecule whose topology information is experimentally solved. This protein is also a member of the rhodopsin family. TMHMM gave 7 as total TM number for it. However, there is a mistake about the locations of the predicted transmembrane regions. TMHMM failed to identify the

Figure 3.4: GPCRs, the last TM missed by TMHMM

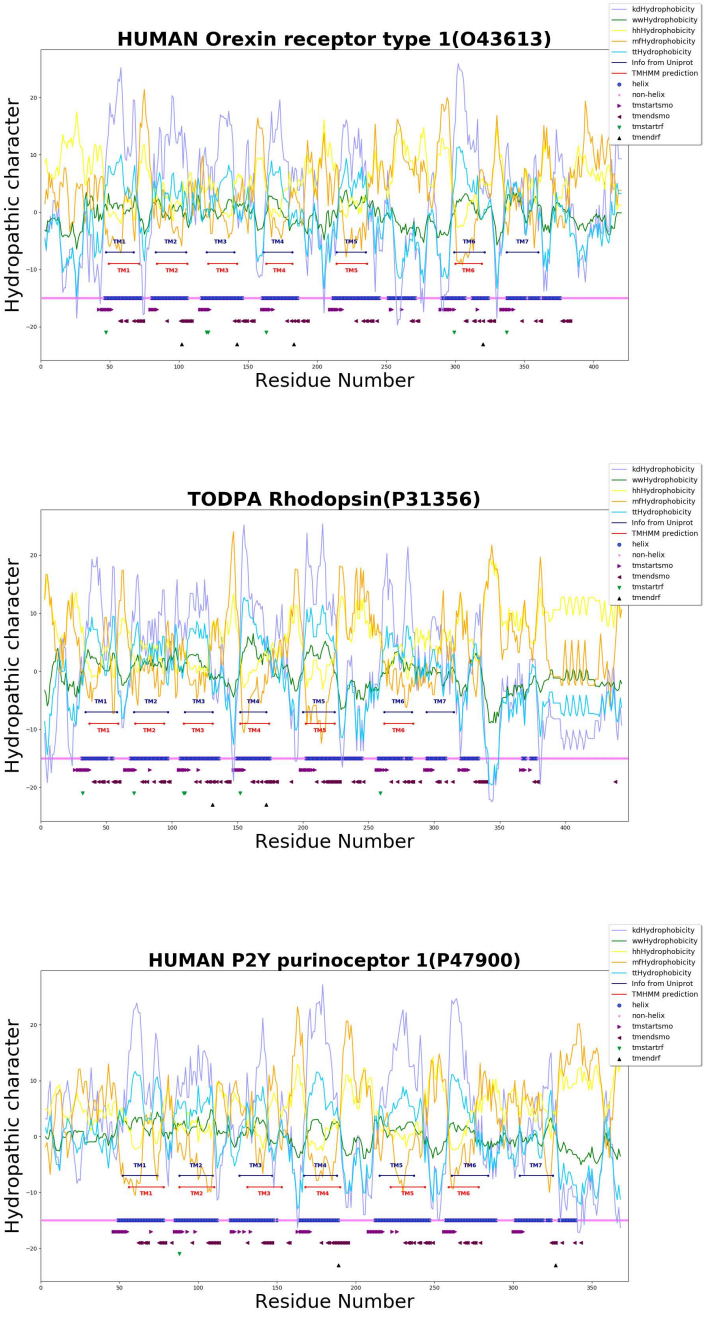
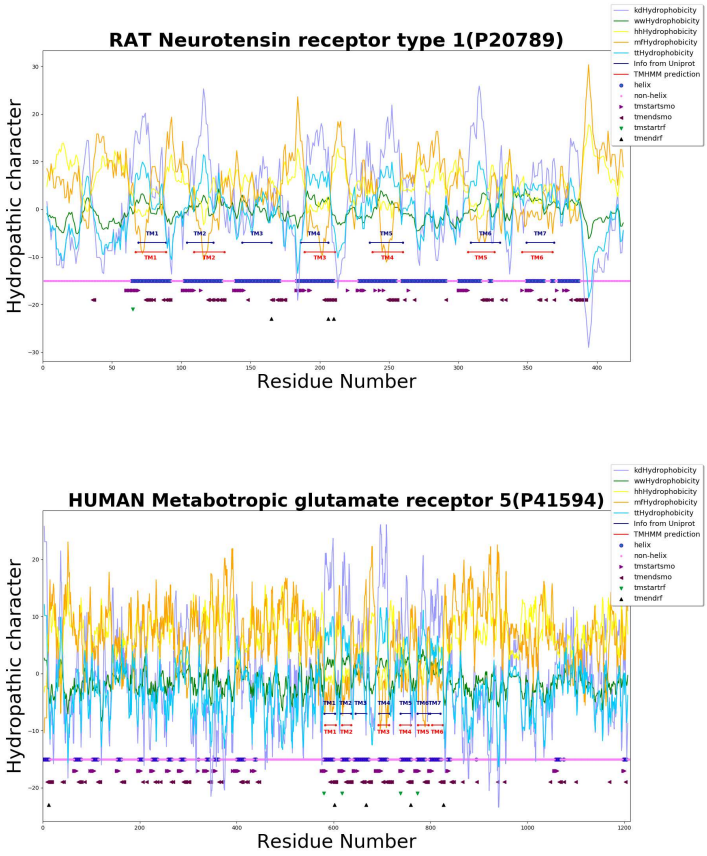


Figure 3.5: GPCRs, the third TM missed by TMHMM



actual TM6 but it predicted another TM region following the actual last TM. Actually, that region is less hydrophobic but has a long helix interrupted 2 times. On the other hand, there is short helix in the actual TM6 region. It can be deduced that TMHMM tends to predict regions of helix with a particular size as transmembrane region. In the Figure 3.6, the difference between the actual TM regions and the predicted TM regions can be more clearly observed.

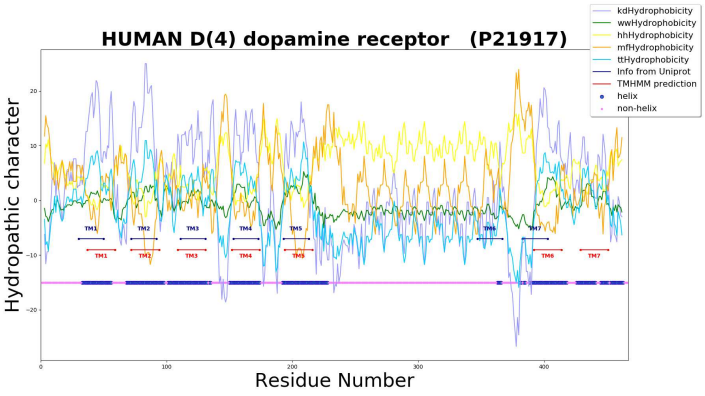


Figure 3.6: GPCR with incorrect prediction of TM region locations

CHAPTER 4

CONCLUSION AND FUTURE WORK

4.1 Conclusion

With this thesis study, we have made a number of contributions. Firstly, it is emphasized that hydrophobicity and secondary structures are key characteristics of the transmembrane regions of GPCRs. In literature, the segments of GPCRs which are embedded in the cell membrane are described as transmembrane helix. Because the cell membrane is hydrophobic due to the lipid building blocks, the proteins embedded in it are naturally expected to be hydrophobic. Hydrophobicity is required for adaptation of the protein to the membrane but according to the baseline method, it alone is not adequate. A pattern for transmembrane regions of GPCRs can be described using hydrophobicity and secondary structures.

Secondly, despite of presence of a variety of hydrophobicity scales in the literature, Moon-Fleming hydrophobicity and Kyte-Doolittle hydrophobicity are the most consistent scales with the real hydrophobic segments in the GPCRs and with the hydrophobic nature of the cell membrane.

Thirdly, SMO and Random Forest are compared with this study. According to the MCC results, it can be said that Random Forest is more efficient with an imbalanced data than SMO.

It is shown that a cost-based approach is an effective way to tackle imbalanced data. On the other hand, accuracy should not be the only parameter for evaluation of a classifier. True positive and false positive rates are as significant as accuracy.

MCC is the most important parameter to evaluate performance of an algorithm trained with an imbalanced data. The proposed method is more efficient for TM-start.

4.2 Future Work

This proposed study can be improved with some future works. First of all, increase in the experimental data will inevitably provide more reliable results and more efficient performance. The proposed method will be performed with the other unknown GPCRs. Moreover, the training data set does not need to be limited to GPCRs. It can be extended to the other membrane proteins. Variety and increase in the data set may give opportunity for a resampling approach which is another way to cope with imbalanced data. Therefore, performance of cost-based approach can be evaluated.
SMO made multiple TM predictions. The successive predictions can be evaluated and the centered one can be regarded as 'True'.
TM start and TM end predictions will be merged to achieve identification of a TM region as a post-process approach.
In order to constraint TM number of prediction, we can use SVM, another ML algorithm, which assigns confidence score to prediction results and the results can be ranked.
In Random Forest, we can analyze the prediction result of each decision-tree and can consider the percentage of the trees which agree on the same class.
Increase in the volume of feature vectors may enhance the accuracy. New features can be added.
TM region prediction can provide accessible surface area calculation and flexibility information.
Amino acid propensities can be used to construct a more improved feature vector.

REFERENCES

[1] N. A. Kratochwil, S. Gatti-McArthur, M. C. Hoener, L. Lindemann, A. D. Christ, L. G. Green, W. Guba, R. E. Martin, P. Malherbe, R. H. P. Porter, J. P. Slack, M. Winnig, H. Dehmlow, U. Grether, C. Hertel, R. Narquizian, C. G. Panousis, S. Kolczewski, and L. Steward, "G protein-coupled receptor transmembrane binding pockets and their applications in gpcr research and drug discovery: a survey," *Current Topics in Medicinal Chemistry*, 2011.

[2] Z. Liao, Y. Ju, and Q. Zou, "Prediction of g protein-coupled receptors with svm-prot features and random forest," *Scientifica*, 2016.

[3] D. Zhang, Q. Zhao, and B. Wu, "Structural studies of g protein-coupled receptors," *Molecules and Cells*, 2015.

[4] W. I. Weis and B. K. Kobilka, "The molecular basis of g protein-coupled receptor activation," *Annual review of biochemistry*, 2018.

[5] D. Bartuzi, A. A. Kaczor, K. M. Targowska-Duda, and D. Matosiuk, "Recent advances and applications of molecular docking to g protein-coupled receptors," *Molecules*, 2017.

[6] J. Storme, A. Cannaert, K. V. Craenenbroeck, and C. P. Stove, "Molecular dissection of the human a3 adenosine receptor coupling with b-arrestin2," *Biochemical Pharmacology*, 2018.

[7] A. D. Mancini, G. Bertrand, K. Vivot, Éric Carpentier, C. Tremblay, J. Ghislain, M. Bouvier, and V. Poitout, "b-arrestin recruitment and biased agonism at free fatty acid receptor," *Journal of Biological Chemistry*, 2015.

[8] H. Komatsu, M. Fukuchi, and Y. Habata, "Potential utility of biased gpcr signaling for treatment of psychiatric disorders," *International Journal of Molecular Sciences*, 2019.

[9] X. long Tang, Y. Wang, D. li Li, J. Luo, and M. yao Liu, "Orphan g protein-coupled receptors (gpcrs): biological functions and potential drug targets," *Acta pharmacologica Sinica*, vol. 33, no. 3, p. 363, 2012.

[10] N. Ye, B. Li, Q. Mao, E. A. Wold, S. Tian, J. A. Allen, and J. Zhou, "Orphan receptor gpr88 as an emerging neurotherapeutic target," *ACS Chemical Neuroscience*, 2018.

[11] T. Ngo, I. Kufareva, J. L. Coleman, R. MGraham, R. Abagyan, and N. J. Smith, "Identifying ligands at orphan gpcrs: current status using structure-based approaches," *British Journal of Pharmacology*, 2016.

[12] R. J. Summers, "Molecular pharmacology of g protein-coupled receptors," *British Journal of Pharmacology*, 2016.

[13] K. Jonas and A. Hanyaloglu, "Impact of g protein-coupled receptor heteromers in endocrine systems," *Molecular and Cellular Endocrinology*, 2017.

[14] P. G. de Oliveira, M. L. S. Ramos, A. J. Amaro, R. A. Dias, and S. I. Vieira, "Gi/o-protein coupled receptors in the aging brain," *Frontiers in Aging Neuroscience*, 2019.

[15] P. A. Insel, K. Sriram, M. W. Gorr, S. Z. Wiley, A. Michkov, C. Salmerón, and A. M. Chinn, "Gpcromics: an approach to discover gpcr drug targets," *Trends in Pharmacological Sciences*, 2019.

[16] A. A.Kaczor, J. Selent, and A. Poso, "Structure-based molecular modeling approaches to gpcr oligomerization," *Methods in Cell Biology*, 2013.

[17] S. P. Alexander, A. Christopoulos, A. P. Davenport, E. Kelly, N. V. Marrion, J. A. Peters, E. Faccenda, S. D. Harding, A. J. Pawson, J. L. Sharman, C. Southan, J. A. Davies, and C. Collaborators, "The concise guide to pharmacology 2017/18: G protein-coupled receptors," *British Journal of Pharmacology*, 2017.

[18] G. Milligan, "Oligomerisation of g-protein-coupled receptors," *Journal of cell science*, 2001.

[19] O. N. Vickery, J.-P. Machtens, and U. Zachariae, "Membrane potentials regulating gpcrs: insights from experiments and molecular dynamics simulations," *Current Opinion in Pharmacology*, 2016.

[20] "G protein–coupled receptor." `https://commons.wikimedia.org/wiki/File:PDB_1hzx_7TM_Sketch_Membrane.png`. last visited on July 2019.

[21] G. Navarro, A. Cordomí, M. Zelman-Femiak, M. Brugarolas, E. Moreno, D. Aguinaga, L. Perez-Benito, A. Cortés, V. Casadó, J. Mallol, E. I. Canela, C. Lluís, L. Pardo, A. J. García-Sáez, P. J. McCormick, and R. Franco, "Quaternary structure of a g-protein-coupled receptor heterotetramer in complex with gi and gs," *BMC biology*, 2016.

[22] X.-Y. Meng, M. Mezei, and M. Cui, "Computational approaches for modeling gpcr dimerization," *Current pharmaceutical biotechnology*, 2014.

[23] W. Nemoto, K. Fukui, and H. Toh, "Grip : A server for predicting interfaces for gpcr oligomerization grip : A server for predicting interfaces for gpcr oligomerization," *Journal of Receptors and Signal Transduction*, 2017.

[24] J. González-Maeso, "Family a gpcr heteromers in animal models," *Frontiers in Pharmacology*, 2014.

[25] X. C. Zhang, J. Liu, and D. Jiang, "Why is dimerization essential for class-c gpcr function? new insights from mglur1 crystal structure analysis," *Protein and Cell*, 2014.

[26] N. J. M. Birdsall, "Class a gpcr heterodimers: Evidence from binding studies," *Trends in Pharmacological Sciences*, 2010.

[27] N. Hurwitz, D. Schneidman-duhovny, and H. J. Wolfson, "Structural bioinformatics memdock : an a -helical membrane protein docking algorithm," *Bioinformatics*, 2016.

[28] J. Selent and A. A. Kaczor, "Oligomerization of g protein-coupled receptors : Computational methods," *Current Medicinal Chemistry*, 2011.

[29] M. Filizola and H. Weinstein, "The study of g-protein coupled receptor oligomerization with computational modeling and bioinformatics," *FEBS Journal*, 2005.

[30] L. M. Simpson, B. Taddese, I. DWall, and C. A. Reynolds, "Bioinformatics and molecular modelling approaches to gpcr oligomerization," *IEEE Conference on Birds*, 2010.

[31] V. Casadó-Anguera, J. Bonaventura, E. Moreno, G. Navarro, A. Cortés, S. Ferré, and V. Casadó, "Evidence for the heterotetrameric structure of the adenosine a 2a –dopamine d 2 receptor complex," *Biochemical Society Transactions*, 2016.

[32] A. Soriano, R. Ventura, A. Molero, R. Hoen, V. Casado, A. Corte, F. Fanelli, F. Albericio, C. Lluís, R. Franco, and M. Royo, "Adenosine a2a receptor-antagonist/dopamine d2 receptor-agonist bivalent ligands as pharmacological tools to detect a 2a-d2 receptor heteromers," *Journal of Medicinal Chemistry*, 2009.

[33] D. O. Borroto-Escuela, D. Marcellino, M. Narvaez, M. Flajolet, N. Heintz, L. Agnati, F. Ciruela, and K. Fuxe, "A serine point mutation in the adenosine a2ar c-terminal tail reduces receptor heteromerization and allosteric modulation of the dopamine d2r," *Biochemical and Biophysical Research Communications*, 2010.

[34] K. Fuxe, S. Ferré, M. Canals, M. Torvinen, A. Terasmaa, D. Marcellino, S. R. Goldberg, W. Staines, K. X. Jacobsen, C. Lluis, A. S. Woods, L. F. Agnati, and R. Franco, "Adenosine a2a and dopamine d2 heteromeric receptor complexes and their function," *Journal of Molecular Neuroscience*, 2005.

[35] A. Prakash and P. M. Luthra, "Insilico study of the a2ar-d2r kinetics and interfacial contact surface for heteromerization," *Amino Acids*, 2012.

[36] A. A. Kaczor, M. Jörg, and B. Capuano, "The dopamine d2 receptor dimer and its interaction with homobivalent antagonists: homology modeling, docking and molecular dynamics," *Journal of Molecular Modeling*, 2016.

[37] J. M. Duarte, N. Biyani, K. Baskaran, and G. Capitani, "An analysis of oligomerization interfaces in transmembrane proteins," *BMC structural biology*, vol. 13, no. 1, p. 21, 2013.

[38] M. Bai, "Dimerization of g-protein-coupled receptors: Roles in signal transduction," *Cellular Signalling*, 2004.

[39] M. Eilers, V. Hornak, S. O. Smith, and J. B. Konopka, "Comparison of class a and d g protein-coupled receptors: common features in structure and activation," *Biochemistry*, 2005.

[40] S. Sadiq, R. Guixà-González, E. Dainese, M. Pastor, G. D. Fabritiis, and J. Selent, "Molecular modeling and simulation of membrane lipid-mediated effects on gpcrs," *Current Medicinal Chemistry*, 2013.

[41] E. K. Tiburu, A. L. Bowman, J. O. Struppe, D. R. Janero, H. K. Avraham, and A. Makriyannisa, "Solid-state nmr and molecular dynamics characterization of cannabinoid receptor-1 (cb1) helix 7 conformational plasticity in model membranes," *Biochimica et Biophysica Acta (BBA)*, 2009.

[42] G. Hedger, H. Koldsø, M. Chavent, C. Siebold, R. Rohatgi, and M. S. Sansom, "Cholesterol interaction sites on the transmembrane domain of the hedgehog signal transducer and class f g protein-coupled receptor smoothened," *Structure*, 2019.

[43] A. Cordomi, S. Ismail, M.-T. Matsoukas, C. Escrieut, M.-J. Gherardi, L. Pardo, and D. Fourmy, "Functional elements of the gastric inhibitory polypeptide receptor: Comparison between secretin- and rhodopsin-like g protein-coupled receptors," *Biochemical Pharmacology*, 2015.

[44] H. Hajjhussein, L. A. Gardner, N. Fujii, N. M. Anderson, and S. W. Bahouth, "The hydrophobic amino acid cluster at the cytoplasmic end of transmembrane helix iii modulates the coupling of the ß1-adrenergic receptor to gs," *Journal of Receptors and Signal Transduction*, 2013.

[45] J. A. Port, M. S. Parker, R. B. Kodner, J. C. Wallace, E. V. Armbrust, and E. M. Faustman, "Identification of g protein-coupled receptor signaling pathway proteins in marine diatoms using comparative genomics," *BMC Genomics*, 2013.

[46] Y. Inoue, M. Ikeda, and T. Shimizu, "Proteome-wide classification and identification of mammalian-type gpcrs by binary topology pattern," *Computational Biology and Chemistry*, 2004.

[47] Y. Lee, S. Kim, S. Choi, and C. Hyeon, "Ultraslow water-mediated transmembrane interactions regulate the activation of a2a adenosine receptor," *Biophysical Journal*, 2016.

[48] R. Abrol, A. R. Griffith, J. K. Bray, and W. A. Goddard, "Structure prediction of g protein-coupled receptors and their ensemble of functionally important conformations," *Methods in Molecular Biology*, 2012.

[49] T. Wang, Y. Wang, L. Tang, Y. Duan, and H. Liu, "7x7 rmsd matrix: a new method for quantitative comparison of the transmembrane domain structures in the g-protein coupled receptors," *Journal of Structural Biology*, 2017.

[50] J. Marino, R. Walser, M. Poms, and O. Zerbe, "Understanding gpcr recognition and folding from nmr studies of fragments," *RSC Advances*, 2018.

[51] D. Juretić, B. Lee, N. Trinajstić, and R. W. Williams, "Conformational preference functions for predicting helices in membrane proteins," *Biopolymers*, 1993.

[52] D. JuretićLarisa, Zoranić, and D. Zucić, "Basic charge clusters and predictions of membrane protein topology," *Journal of Chemical Information and Computer Sciences*, 2002.

[53] W. Hoffmann, J. Langenhan, S. Huhmann, J. Moschner, R. Chang, M. Accorsi, J. Seo, R. Jörg, G. Meijer, B. Koksch, M. T. Bowers, G. von Helden, and K. Pagel, "An intrinsic hydrophobicity scale for amino acids and its application to fluorinated compounds," *Angewandte Chemie - International Edition*, 2019.

[54] "Amino acid hydrophobicity." `https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/hydrophob.html`, 2018.

[55] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathy of a protein," *Journal of Molecular Biology*, 1982.

[56] W. C. Wimley and S. H. White, "Experimentally determined hydrophobicity scale for proteins at membrane interfaces," *Nature Structural Biology*, 1996.

[57] W. C. Wimley, T. P. Creamer, and S. H. White, "Solvation energies of amino acid side chains and backbone in a family of hostguest pentapeptides," *Biochemistry*, 2002.

[58] S. H. White and W. C. Wimley, "Membrane protein folding and stability: physical principles," *Annual Review of Biophysics and Biomolecular Structure*, 1999.

[59] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S. H. White, and G. von Heijne, "Recognition of transmembrane helices by the endoplasmic reticulum translocon," *Nature*, 2005.

[60] C. P. Moon and K. G. Fleming, "Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers," *Proceedings of the National Academy of Sciences*, 2011.

[61] G. Zhao and E. London, "An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity," *Protein Science*, 2006.

[62] S. Pundir, M. J. Martin, and C. O'Donovan, *Protein Bioinformatics*. Humana Press, 2017.

[63] T. U. Consortium, "Uniprot: the universal protein knowledgebase," *Nucleic Acids Research*, 2017.

[64] "Transmembrane." `https://www.uniprot.org/help/transmem`. last visited on September 2019.

[65] G. E. Tusnady and I. Simon, "Principles governing amino acid composition of integral membrane proteins: application to topology prediction," *Journal of molecular biology*, vol. 283, no. 2, pp. 489–506, 1998.

[66] C. Pasquier, V. Promponas, G. Palaios, J. Hamodrakas, and S. Hamodrakas, "A novel method for predicting transmembrane segments in proteins based on a

statistical analysis of the swissprot database: the pred-tmr algorithm," *Protein engineering*, vol. 12, no. 5, pp. 381–385, 1999.

[67] C. Pasquier and S. Hamodrakas, "An hierarchical artificial neural network system for the classification of transmembrane proteins," *Protein engineering*, vol. 12, no. 8, pp. 631–634, 1999.

[68] D. Jones, W. Taylor, and J. Thornton, "A model recognition approach to the prediction of all-helical membrane protein structure and topology," *Biochemistry*, vol. 33, no. 10, pp. 3038–3049, 1994.

[69] T. Hirokawa, S. Boon-Chieng, and S. Mitaku, "Sosui: classification and secondary structure prediction system for membrane proteins.," *Bioinformatics (Oxford, England)*, vol. 14, no. 4, pp. 378–379, 1998.

[70] M. Bernhofer, E. Kloppmann, J. Reeb, and B. Rost, "Tmseg: novel prediction of transmembrane helices," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. 11, pp. 1706–1716, 2016.

[71] "TMbase - A Database of Membrane Spanning Protein Segments." `https://embnet.vital-it.ch/software/tmbase/TMBASE_doc.html`. last visited on September 2019.

[72] L. Käll, A. Krogh, and E. L. Sonnhammer, "A combined transmembrane topology and signal peptide prediction method," *Journal of molecular biology*, vol. 338, no. 5, pp. 1027–1036, 2004.

[73] A. Krogh, B. Larsson, G. V. Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes," *Journal of Molecular Biology*, 2001.

[74] E. L. L. Sonnhammer, G. V. Heijne, and A. Krogh, "A hidden markov model for predicting transmembrane helices in protein sequences," in *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, ISMB '98, pp. 175–182, AAAI Press, 1998.

[75] "Transmembrane Helix Prediction." `http://www.cbs.dtu.dk/~krogh/TMHMM/`, 2000.

[76] K. K. Yoneten, M. Kasap, G. Akpinar, A. Kanli, and E. Karaoz, "Comparative proteomics analysis of four commonly used methods for identification of novel plasma membrane proteins," *The Journal of Membrane Biology*, 2019.

[77] C. Papaloukas, E. Granseth, H. Viklund, and A. Elofsson, "Estimating the length of transmembrane helices using z-coordinate predictions," *Protein Science*, vol. 17, no. 2, pp. 271–278, 2008.

[78] T. M. Beissinger, G. J. Rosa, S. M. Kaeppler, D. Gianola, and N. D. Leon, "Defining window-boundaries for genomic analyses using smoothing spline techniques," *Genetics Selection Evolution*, vol. 47, no. 1, p. 30, 2015.

[79] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton, "JPred4: a protein secondary structure prediction server," *Nucleic Acids Research*, vol. 43, no. W1, 2015.

[80] "Jpred 4 incorporating Jnet: A Protein Secondary Structure Prediction Server." `http://www.compbio.dundee.ac.uk/jpred/faq.shtml#Q12`, 2015. last visited on July 2019.

[81] S. Ruskamo, M. Green, J. Vahokoski, S. Bhargav, F. Liang, I. Kursula, and P. Kursula, "Juxtanodin is an intrinsically disordered f-actin-binding protein," *Scientific reports*, vol. 2, p. 899, 11 2012.

[82] D. P. Staus, L. M. Wingler, D. Pichugin, R. S. Prosser, and R. J. Lefkowitz, "Detergent- and phospholipid-based reconstitution systems have differential effects on constitutive activity of g protein-coupled receptors," *Journal of Biochemical Chemistry*, 2019.

[83] S. Maheshwari, J. Agrawal, and S. Sharma, "New approach for classification of highly imbalanced datasets using evolutionary algorithms," *Int. J. Sci. Eng. Res*, vol. 2, no. 7, pp. 1–5, 2011.

[84] N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data," in *The 2010 International joint conference on neural networks (IJCNN)*, pp. 1–8, IEEE, 2010.

[85] T. Razzaghi, "Cost-sensitive learning-based methods for imbalanced classification problems with applications," 2014.

[86] T. C. Smith and E. Frank, *Statistical Genomics*. Humana Press, 2016.

[87] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[88] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Immprovements to platt's smo algorithm for svm classifier design," *Neural Computation*, 2001.

[89] N. V. Petrova and C. H. Wu, "Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties," *BMC Bioinformatics*, 2006.

[90] F. Pérez-Cruz., P. L. Alarcón-Diana, A. Navia-Vázquez, and A. Artés-Rodríguez, "Fast training of support vector classifiers," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, (Cambridge, MA, USA), pp. 706–712, MIT Press, 2000.

[91] R. Ruiz-Gonzalez, J. Gomez-Gil, F. J. Gomez-Gil, and V. Martínez-Martínez, "An svm-based classifier for estimating the state of various rotating components in agro-industrial machinery with a vibration signal acquired from a single point on the machine chassis," *Sensors*, vol. 14, no. 11, pp. 20713–20735, 2014.

[92] S. Rajamohana, K. Umamaheswari, and R. Karthiga, "Sentiment classification based on lda using smo classifier," *International Journal of Applied Engineering Research*, vol. 55, pp. 1045–1049, 2015.

[93] S. Singaravelan, D. Murugan, and R. Mayakrishnan, "Analysis of classification algorithms j48 and smo on different datasets," *World Engineering & Applied Sciences Journal*, vol. 6, no. 2, pp. 119–123, 2015.

[94] J. K. Williams and J. Abernethy, "Using random forests and fuzzy logic for automated storm type identification," in *AMS Sixth Conference on Artificial Intelligence Applications to Environmental Science*, 2008.

[95] D. Uccellatore, *Logic-in-Memory implementation of Random Forest Algorithm*. PhD thesis, Politecnico di Torino, 2018.

[96] M. E. Sahin, T. Can, and C. D. Son, "Gpcrsort-responding to the next generation sequencing data challenge: prediction of g protein-coupled receptor classes using only structural region lengths," *OMICS-a Journal of Integrative Biology*, 2014.

[97] L. C. Xue, D. Dobbs, A. M. J. J. Bonvin, and V. Honavar, "Computational prediction of protein interfaces: A review of data driven methods," *FEBS Letters*, 2015.

[98] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, 2018.

[99] F. J. W. M. Dankers, A. Traverso, L. Wee, and S. M. J. van Kuijk, *Fundamentals of Clinical Data Science [Internet]"*. Springer, 2019.

[100] D. Kihara, T. Shimizu, and M. Kanehisa, "Prediction of membrane proteins based on classification of transmembrane segments.," *Protein engineering*, vol. 11, no. 11, pp. 961–970, 1998.

[101] M. Ahram, Z. I. Litou, R. Fang, and G. Al-Tawallbeh, "Estimation of membrane proteins in the human proteome," *In silico biology*, vol. 6, no. 5, pp. 379–386, 2006.

[102] M. M. Attwood, A. Krishnan, V. Pivotti, S. Yazdi, M. S. Almén, and H. B. Schiöth, "Topology based identification and comprehensive classification of four-transmembrane helix containing proteins (4tms) in the human genome," *BMC genomics*, vol. 17, no. 1, p. 268, 2016.

[103] T. Sato, T. Kawasaki, S. Mine, and H. Matsumura, "Functional role of the c-terminal amphipathic helix 8 of olfactory receptors and other g protein-coupled receptors," *International journal of molecular sciences*, vol. 17, no. 11, p. 1930, 2016.

[104] M. Cserzö, F. Eisenhaber, B. Eisenhaber, and I. Simon, "On filtering false positive transmembrane protein predictions," *Protein Engineering, Design and Selection*, vol. 15, pp. 745–752, 09 2002.

[105] G. Wang, Q. Li, L. Wang, W. Wang, M. Wu, and T. Liu, "Impact of sliding window length in indoor human motion modes and pose pattern recognition based on smartphone sensors," *Sensors*, vol. 18, no. 6, p. 1965, 2018.

[106] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using matthews correlation coefficient metric," *PloS one*, vol. 12, no. 6, p. e0177678, 2017.

[107] C. Chen, A. Liaw, L. Breiman, *et al.*, "Using random forest to learn imbalanced data," *University of California, Berkeley*, vol. 110, no. 1-12, p. 24, 2004.

[108] E. Kremic and A. Subasi, "Performance of random forest and svm in face recognition.," *Int. Arab J. Inf. Technol.*, vol. 13, no. 2, pp. 287–293, 2016.

[109] S. Bharathidason and C. J. Venkataeswaran, "Improving classification accuracy based on random forest model with uncorrelated high performing trees," *Int. J. Comput. Appl*, vol. 101, no. 13, pp. 26–30, 2014.

APPENDIX A

HYDROPHOBICITY SCALES

Table A.1: Numerical Values of Hydrophobicity Scales

Letter	Residue type	kdHydrophobicity	wwHydrophobicity	hhHydrophobicity	mfHydrophobicity	ttHydrophobicity
I	Isoleucine	4.5	0.31	-0.60	-1.56	1.97
V	Valine	4.2	-0.07	-0.31	-0.78	1.46
L	Leucine	3.8	0.56	-0.55	-1.81	1.82
F	Phenylalanine	2.8	1.13	-0.32	-2.20	1.98
C	Cysteine	2.5	0.24	-0.13	0.49	-0.30
M	Methionine	1.9	0.23	-0.10	-0.76	1.40
A	Alanine	1.8	-0.17	0.11	0.0	0.38
G	Glycine	-0.4	-0.01	0.74	1.72	-0.19
T	Threonine	-0.7	-0.14	0.52	1.78	-0.32
S	Serine	-0.8	-0.13	0.84	1.83	-0.53
W	Tryptophan	-0.9	1.85	0.30	-0.38	1.53
Y	Tyrosine	-1.3	0.94	0.68	-1.09	0.49
P	Proline	-1.6	-0.45	2.23	-1.52	-1.44
H	Histidine	-3.2	-0.96	2.06	4.76	-1.44
E	Glutamic Acid	-3.5	-2.02	2.68	1.64	-2.90
Q	Glutamine	-3.5	-0.58	2.36	3.01	-1.84
D	Aspartic Acid	-3.5	-1.23	3.49	2.95	-3.27
N	Asparagine	-3.5	-0.42	2.05	3.47	-1.62
K	Lysine	-3.9	-0.99	2.71	5.39	-3.46
R	Arginine	-4.5	-0.81	2.58	3.71	-2.57