Prediction of Lysosomal Membrane Proteins using Machine Learning Techniques

122505
VADKE JAAI VIVEKANAND

Under the guidance of,
DR. JAYASHREE RAMANA

May – 2014

Submitted in partial fulfillment of the Degree of
Master of Technology

DEPARTMENT OF BIOTECHNOLOGY AND
BIOINFORMATICS, JAYPEE UNIVERSITY OF INFORMATION
TECHNOLOGY, WAKNAGHAT

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABBRIVIATIONS

AAC: Amino Acid Composition;

DPC: Dipeptide Composition;

LMP: Lysosomal Membrane Protein;

LOO: Leave-One-Out;

MCC: Matthews Correlation Coefficient;

PSI-BLAST: Position-Specific Iterative-Basic Local Alignment Search Tool;

PSSM: Position-Specific Scoring Matrix;

RBF: Radial Basis Function;

SVM: Support Vector Machine.

CERTIFICATE

This is to certify that the work titled **Prediction of Lysosomal Membrane Proteins using Machine Learning Techniques** submitted by **Miss. Jaai Vadke** in partial fulfillment for the award of degree of Master of Technology of Jaypee University of Information Technology, Waknaghat has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor ……..

Name of Supervisor ……..……

Designation ……..……

Date ……

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Jayashree Ramana, Assistant Professor of Bioinformatics and my guide for final year project, for providing her valuable guidance throughout the project. She not only helped me shape the overall project but also introduced me to a knowledge that I could use in my future.

I wish to thank our Head of Department, Dr, Rajinder Chauhan, Professor of Biotechnology, who fulfilled our project requirements, and also all helping teachers of bioinformatics department, for making me think on my own novel ideas during the lectures which will help me throughout my life.

Special thanks are due my senior, Ms. Tamanna, PhD scholar, for helping me at different stages of project and her patience which helped me completing my project fruitfully.

And undoubtedly my family and my classmates, Rajinder and Anuja who gave me a moral support and suggestions for better project work, I heartily thank them for their encouragement during these 12 months.

Signature of the student ……..……..

Name of Student ……..……..

Date ……..……..

SUMMARY

The existence of public databases with billions of data entries requires a robust analytical approach to represent it with respect to its biological significance. Disease-causing mutations in genes encoding for lysosomal membrane proteins have been only described in the last decade, showing that the rapid progress in this research field is due to the achievements of the human genome project. Very few bioinformatics tools are designed to classify these proteins previously, which either non-specific for such proteins or not available freely. This work presents a machine learning methodology, which classifies the proteins into their classes from their sequences alone: the lysosomal membrane proteins and non-lysosomal membrane proteins. In this study, Support Vector Machine (SVM)-based lysosomal membrane protein prediction system is proposed. Different protein sequence representations are fused to extract the features of a protein sequence, which includes amino acid (AA) composition, dipeptide (DP) composition and normalized Position Specific Scoring Matrices (PSSM). SVM_light software is used as a classifier tool and compared with HMMER package which is a HMM profile based classifier. From this study it is seen that the accuracy of SVM classifier based on PSSM; among other machine learning techniques comes out to be the highest. The overall accuracy of leave one out cross-validation is almost 76% for the data-set. These predicted results suggest that the method can be further modified to discriminate lysosomal membrane proteins from other membrane proteins and Globular proteins, and it also indicates that the PSSM representation of proteins can better reflect the feature of membrane proteins than the classical AA composition.

_____ _____

Signature of Student
Name:
Date:

Signature of Supervisor
Name:
Date:

CHAPTER 1
INTRODUCTION

With the advancements in the field of protein sequencing techniques, we have entered in the era of proteomics which is an important part of bioinformatics studies. The existence of public databases with billions of data entries requires a robust analytical approach to represent it with respect to its biological significance. Therefore, computational tools are needed to analyze the collected data in the most efficient manner. For acquiring knowledge from the sequence data, there are different ways to achieve it like classifying the sequence on the basis of its subcellular location or determining the structure and hence the function of specific protein. In this study I have developed a machine-learning based method for prediction of lysosomal membrane proteins.

Why Lysosomal Membrane Proteins?

To understand Lysosomal membrane proteins firstly we have to know what Lysosomes are. Lysosomes are membrane-enclosed organelles that contain an array of enzymes capable of breaking down all types of biological polymers—proteins, nucleic acids, carbohydrates, and lipids. Lysosomes function as the digestive system of the cell, serving both to degrade material taken up from outside the cell and to digest obsolete components of the cell itself. In their simplest form, lysosomes are visualized as dense spherical vacuoles, but they can display considerable variation in size and shape as a result of differences in the materials that have been taken up for digestion. Lysosomes thus represent morphologically diverse organelles defined by the common function of degrading intracellular material.

Lysosomes are ubiquitous organelles that constitute the primary degradative compartments of the cell. They receive their substrates through endocytosis, phago-cytosis or autophagy. The catabolic function of lysosomes is complemented by lysosome-related organelles (LROs), such as melanosomes, lytic granules, major histo-compatibility complex (MHC) class II

compartments and platelet-dense granules [1]. LROs share many properties with lysosomes, but they also contain cell-type-specific proteins and might require additional cellular machinery for their biogenesis [2, 3]. Lysosomes and LROs are involved in various physiological processes, such as cholesterol homeostasis, plasma membrane repair, bone and tissue remodeling, pathogen defense, cell death and cell signaling (Figure. 1). These complex functions make the lysosome a central and dynamic organelle and not simply the dead end of the endocytic pathway. Two classes of proteins are essential for the function of lysosomes: soluble lysosomal hydrolases (also referred to as acid hydrolases) and integral lysosomal membrane pro-teins (LMPs). Each of the 50 known lysosomal hydrolases targets specific substrates for degradation, and their collective action is responsible for the total catabolic capacity of the lysosome. In addition to bulk degradation and pro-protein processing, lysosomal hydrolases are involved in antigen processing, degradation of the extracellular matrix and initiation of apoptosis [4]. The mammalian lysosome contains ~25 LMPs [5], but additional LMPs are being revealed [5–7]. LMPs reside mainly in the lysosomal limiting membrane and have diverse functions, including acidification of the lysosomal lumen, protein import from the cytosol, membrane fusion and transport of degrada-tion products to the cytoplasm [8] (Figure. 1). The most abun-dant LMPs are lysosome-associated membrane protein 1 (LAMP1), LAMP2, lysosome integral membrane protein 2 (LIMP2; also known as SCARB2) and the tetraspanin CD63.

The lysosomal membrane

Lysosomes are limited by a single 7-10 nm phospholipid-bilayer [9]. A unique feature of the lysosomal membrane is its high carbohydrate content. Lysosomal membrane proteins are generally heavily glycosylated at their luminal domain and form a glycocalyx, which is suggested to protect the membrane from the action of the hydrolytic enzymes contained within this organelle [10]. One crucial role of the membrane limiting lysosomes is to separate the potent activities of lysosomal acid hydrolases from other cellular constituents, thereby

preventing uncontrolled proteolytic damage [9]. The lysosomal membrane also facilitates interaction and fusion with other cellular compartments, including endosomes, autophagosomes and the plasma membrane [11]. In addition to the limiting lysosomal membrane lysosomes have intralysosomal membranes, which represent the main site of membrane degradation within this organelle [12].

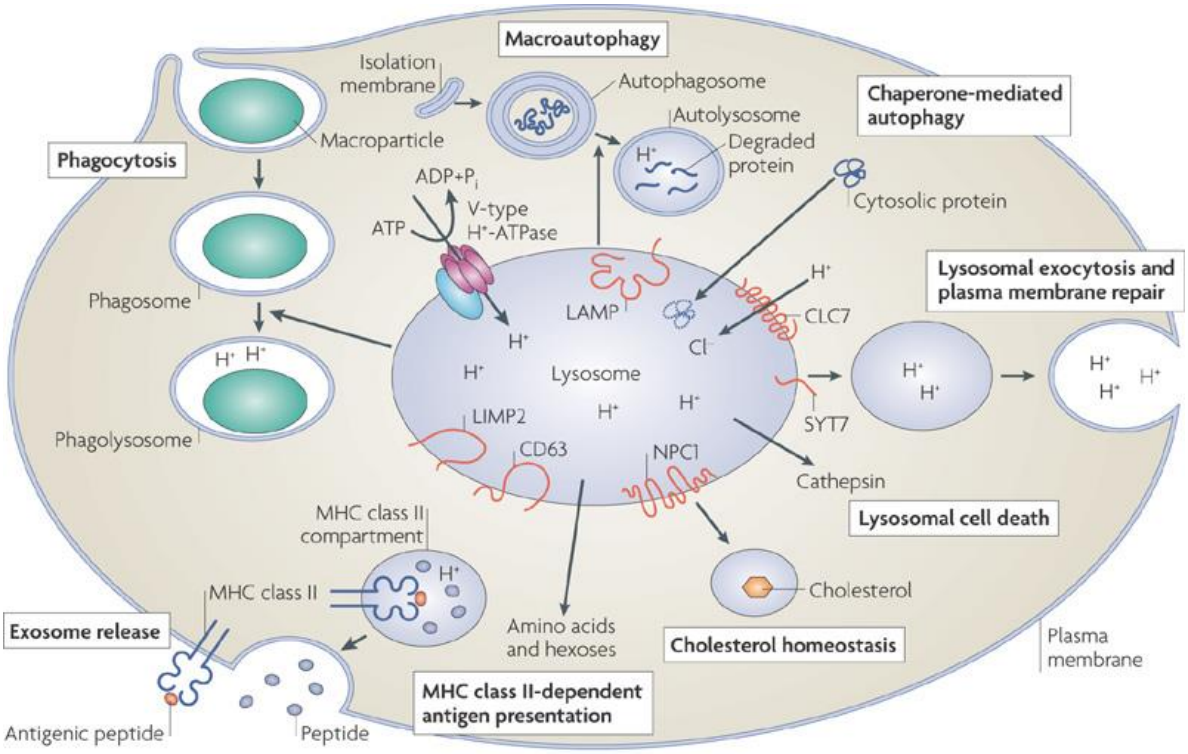


Figure 1: Major functions of lysosomal membrane proteins

Lysosome is a central, acidic organelle that is involved in the degradation of macromolecules through the activity of lysosomal hydrolases. Lysosomes are crucial for the maturation of phagosomes to phago-lysosomes in phagocytosis, which is important for cellular pathogen defense. The macro-autophagy pathway mediates the turnover of cytoplasmic components, such as organelles and large complexes, and is involved in cell death and proliferation.

Macroautophagy depends on the fusion of lysosomes with autophagosomes to create autolysosomes, in which degradation occurs. Macroautophagy and chaperone-mediated autophagy, a direct lysosomal transport process for cytosolic proteins, are regulated by lysosome-associated membrane proteins (LAMPs). Lysosomal exocytosis and plasma membrane repair are Ca2+ and synaptotagmin 7 (SYT7)-dependent fusion events, which are possibly involved in pathogen entry, autoimmunity and neurite outgrowth. The lysosomal cell death pathway is triggered by a release of lysosomal cathepsins through an unknown mechanism. Cellular cholesterol homeostasis is controlled by lysosomal cholesterol efflux through Niemann–Pick C1 protein (NPC1). Major histocompatibility complex (MHC) class II-dependent antigen presentation requires lysosomal proteases and membrane proteins. The release of exosomes is thought to be involved in adaptive immune responses. Lysosomal membrane proteins are also involved in the transport of newly synthesized hydrolases to the lysosome (for example, lysosomal integral membrane protein 2 (LIMP2)) and across the lysosomal.

Acid hydrolases and lysosomal membrane proteins

Two categories of proteins are essential for the correct function of lysosomes: integral membrane proteins and soluble hydrolytic enzymes. The approximately 60 resident hydrolases have different target substrates, and their collective action permits the degradation of all types of macromolecules [13]. Lysosomal proteins are synthesized at the rough ER, transferred to the Golgi apparatus and targeted to the lysosome by specific sorting mechanisms. Targeting of newly synthesized lysosomal proteins can be direct, from the trans-Golgi network (TGN) to the endosomal system, or indirect, involving transport to the plasma membrane and subsequent endocytosis [14]. The best characterized route is the direct pathway, which is dependent on the mannose-6-phosphate (M6P) receptor, through which the majority of lysosomal hydrolases end up in lysosomes [15]. After synthesis, proteins move to the cis-Golgi network, where they are covalently modified by the addition of M6P residues

[15]. The M6P-tagged lysosomal hydrolases are recognized and bound by M6P receptors in the TGN and sorted into transport vesicles, which bud off from the TGN and fuse with late endosomes. At the low pH of the late endosome, the hydrolases dissociate from the M6P receptors, and the empty receptors are recycled to the Golgi apparatus for further transport [15].

Approximately 25 lysosomal membrane proteins have been identified, which reside primarily in the limiting lysosomal membrane [13, 14]. Proteins residing in the lysosomal membrane are usually highly glycosylated transmembrane proteins, which mediate a number of essential functions for the organelle, including acidification of the lysosomal lumen, import of protein from the cytosol and transport of degradation end products out of the lysosome. The characteristic acidic pH of lysosomes is a result of the action of the vacuolar H+-ATPase, a transmembrane multimeric protein complex [16]. The vacuolar H+-ATPase uses energy from ATP hydrolysis to pump protons from the cytosol against their electrochemical gradient into the lysosomal lumen [16]. Other lysosomal membrane proteins are involved in interactions and fusion with other cell components, including endosomes, phagosomes and the plasma membrane.The most abundant lysosomal membrane proteins are lysosome-associated membrane protein (LAMP)-1 and -2, lysosomal integral membrane protein (LIMP)-2 and CD63 [14].

Bis(monoacylglycero)-phosphate (BMP)

Bis(monoacylglycero)-phosphate (BMP), also known as lyso-bis-phosphatidic acid (LBPA), is an unusual phospholipid that is found mainly in the inner membrane of lysosomes and late endosomes [17]. The unusual stereo conformation of BMP results in higher resistance to the action of phospholipases compared to other phospholipids. In the endolysosomal system, hydrophobic lipids and membranes are digested by hydrophilic enzymes, a process in which BMP serves as an important factor. BMP is negatively charged at the acidic pH of lysosomes, and these negative charges facilitate the adhesion of the soluble positively charged

hydrolases, thus allowing the hydrolases to degrade lipids at the interface of the inner membranes of lysosomes [17]. In addition, evidence suggests that BMP regulates the dynamics of the internal membranes of late endosomes, is involved in protein- and lipid-sorting and plays a critical role in endo/lysosomal cholesterol trafficking.

Lysosome-associated membrane proteins (LAMPs)

LAMP-1 and -2 have been estimated to constitute 50% of lysosomal membrane proteins [9]. LAMPs are transmembrane proteins with a large luminal domain, a transmembrane domain and a short C-terminal cytoplasmic tail [18]. They are heavily glycosylated, as indicated by the increase in the mass of the polypeptide from approximately 40 kDa to 120 kDa after glycosylation. Mice deficient in LAMP-1 are viable and demonstrate a mild phenotype with normal lysosomal morphology and function [9]. Deficiency of LAMP-2 induces a more severe phenotype with extensive accumulation of autophagic vacuoles in many tissues, and degradation of long-lived proteins is severely impaired [9]. These findings indicate that LAMP-2 is critical for autophagy (described later), which is further substantiated by the finding that LAMP-2 deficiency in humans causes Danon's disease. This disease is a lysosomal glycogen storage disease that is associated with the accumulation of autophagic material in striated myocytes, resulting in a pathological condition associated with cardiomyopathy, myopathy and variable mental retardation [19].

Lysosomal-integral membrane protein 2 (LIMP-2)

LIMP-2/LGP85 belongs to the CD36 family of scavenger receptors and is one of the most abundant ubiquitously expressed lysosomal membrane proteins. It spans the membrane twice, with the N- and C-terminus located in the cytosol and exhibits a highly glycosylated loop within the lysosomal lumen [20]. LIMP-2 may be involved in lysosome/endosome biogenesis [21]. Overexpression of LIMP-2/LGP85 was shown to result in the accumulation of large swollen vacuoles that share both early and late endosomal as well as lysosomal features.

These large vacuoles appear electron-lucent with only occasional luminal membranes, suggesting that the invagination of internal vesicles may be impaired. Pulse–chase experiments showed that the large vacuoles were not initially derived from lysosomes. Co-expression of dominant-negative Rab5b with LIMP-2/LGP85 totally inhibited the formation of the large swollen vacuoles, indicating that normal function of Rab5 was necessary for their appearance. These results suggest that LIMP-2/LGP85 may control the balance between vesicle invagination and vesicle budding from the limiting membrane of endosomal compartments. It is possible that overexpression of LIMP-2/LGP85 causes a dispersal of the budding machinery, which might be due to an impaired recruitment of an as-yet-unknown cytoplasmic factor that is involved in vesicular fission and/or fusion. LIMP-2 is the receptor for the mannose 6-phosphate receptor-independent transport of β- GC (β-glucocerebrosidase) to the lysosome [22]. In LIMP- 2-deficient fibroblasts or macrophages, β-GC is no longer effectively transported to the lysosome, but is secreted into the cell culture medium. Also, in vivo missorting of β-GC occurs with low tissue levels of β-GC and increased enzyme activity in serum of LIMP-2-deficient mice. Previous studies indicate that the interaction of β-GC and LIMP-2 takes place very early in the secretory pathway in the ER [9]. From the ER, the receptor–ligand complex then traffics through the Golgi to the lysosome, where its acidic pH probably leads to a dissociation of the ligand from its receptor. Recently, mutations in the human gene encoding the lysosomal integral membrane protein LIMP-2 were found to be responsible for AMRF (action myoclonus–renal failure) syndrome, a fatal autosomal-recessive disorder characterized by focal glomerulosclerosis, progressive myoclonus epilepsy and ataxia [23]. It was found that AMRF disease-causing mutations similar to a disruption of a crucial coiled-coil domain within the luminal part of LIMP-2 abolished β-GC binding [24].

CD63: an unusual lysosomal tetraspanin

CD63, also called LIMP-1, belongs to the family of tetraspanins [25]. This family is composed of 33 members in mammals, spanning the membrane four times and forming a

small and a large extracellular loop. Tetraspanins group specific cell-surface proteins and thereby increase the formation and stability of functional signalling complexes. Such complexes are involved in diverse cellular processes, such as cell activation, adhesion, motility, differentiation and malignancy. CD63 is an exceptional tetraspanin, since, at steady state, it is usually found as a heavily glycosylated protein in late endosomes/lysosomes. The majority of tetraspanins described so far usually reside in the plasma membrane. Despite the existence of abundant data on the presumed role of CD63 in isolated cell types, its function in vivo is largely unknown. The phenotype of CD63-knockout mice [26] suggests a role for CD63 in development and distribution of immune system cells, a regulatory activity in platelet adhesion, and an important role in kidney physiology.

Lysosome biogenesis requires integration of the endocytic and biosynthetic pathways of the cell (Figure. 2). Lysosomal targeting of newly synthesized lysosomal proteins can be direct, from the trans-Golgi network (TGN) to the endosomal system, or indirect, involving transport to the plasma membrane and subsequent endocytosis. The best understood direct pathway is the mannose-6-phosphate receptor (M6PR) mediated transport of lysosomal hydro-lases [27, 28]. By contrast, remarkably little is known about the structural and molecular machinery for the transport of LMPs to lysosomes. The significance of tightly regu-lated LMP trafficking is illustrated by recent findings that describe new and unexpected roles for LMPs in cellular physiology. It is becoming apparent that LMPs can impose specific functions onto the organelles through which they are transported or in which they reside, such as the endoplasmic reticulum (ER), lysosomes and the plasma membrane. Their importance is further highlighted by the discovery of an increasing number of gene mutations that lead to lysosomal dysfunction and disease [29]. In addition, various knockout mice and non-mammalian model organisms have highlighted the role of LMPs in cell. Here, I have discussed the cellular pathways involved in lysosome biogenesis and the putative and emerging roles of LMPs in the transport of proteins and organelles and the consequences of their impaired trafficking for human health.

Endocytic pathways to the lysosome

The degradative endocytic pathway starts at the plasma membrane and ends in lysosomes. Between these two ‘stations’, endocytosed cargo passes through a range of endosomal intermediates (Figure. 2) that are distinguished by their content, molecular make-up, morphology and pH and by the kinetics by which endocytic tracers reach them[30]. A constant exchange of incoming and outgoing membranes and multiple fusion events result in the gradual remodeling of an endosomal intermediate into a later-stage endosome [31], a process called maturation [32]. In addition, endosomes can exchange content through vesicular transport carriers and tubular connections [33].The widely used distinction between early endosomes (EEs) and late endosomes (LEs) [30] is based on functional and biological characteristics, but oversimplifies the complexity of the endocytic pathway. This was exemplified by a recent immunoelectron microscopy (IEM) study linking the molecular make-up of endosomes with their ultrastructural characteristics [34]. Distinct EE marker proteins shows different distributions, ranging from a restricted localization on early stage EEs to a more widespread distribution on other EEs and on early-stage LEs. These observations indicate that functionally different intermediates of EEs and LEs can be distinguished and that the transition between EEs and LEs is gradual. The endocytic pathway is therefore best regarded as a spatiotemporal continuum of inter mediates, which continuously exchange their content while under-going gradual molecular and structural remodelling and functional transformation.

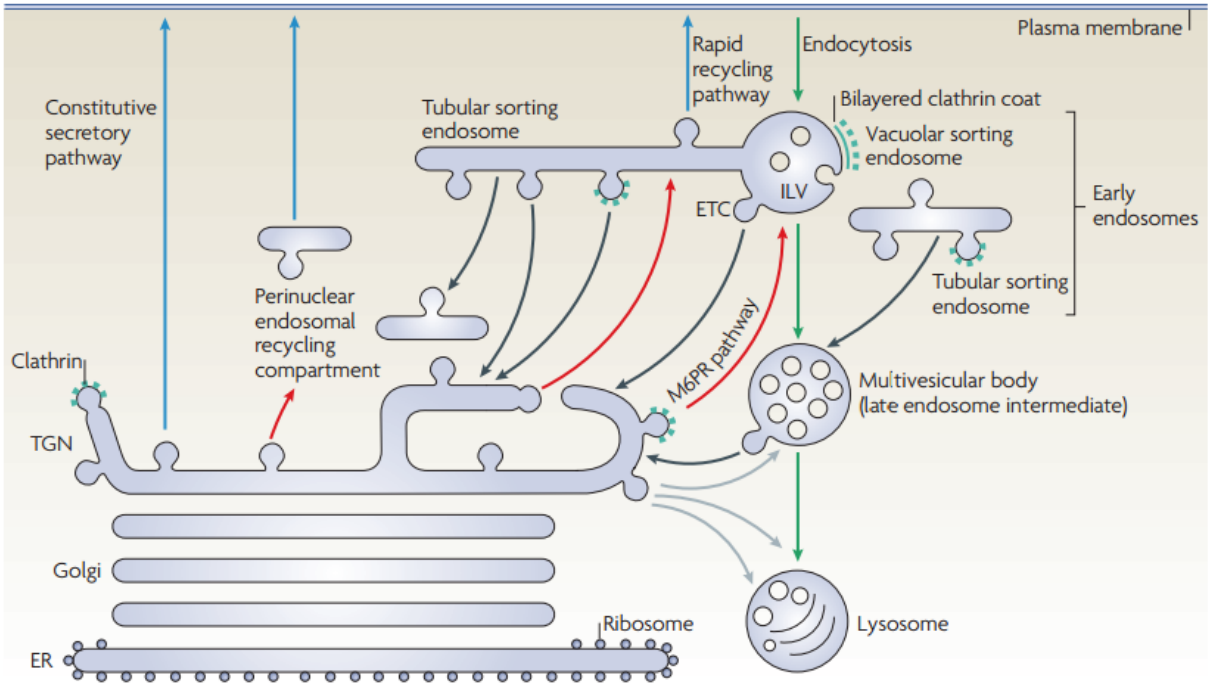


Figure 2 : Possible interactions between the biosynthetic and endocytic pathways.

Lysosome biogenesis requires the concerted involvement of biosynthetic and endocytic pathways. As shown in figure 2, Lysosomes receive cargo for degradation as well as newly synthesized lysosomal proteins by the endocytic pathway (green arrows). Lysosomal proteins are synthesized in the endoplasmic reticulum (ER) and transported through the Golgi complex to the trans-Golgi network (TGN). From the TGN, they can follow the constitutive secretory pathway (blue arrows) to the plasma membrane and subsequently reach lysosomes by endocytosis. In addition, lysosomal proteins can follow a direct intracellular pathway (red arrows) to the endo-lysosomal system. The best-characterized direct intracellular pathway is the clathrin-dependent transport of lysosomal hydrolases by mannose-6-phosphate receptors (M6PRs). The available literature suggests that there are multiple pathways for both lysosomal hydrolases and lysosomal membrane proteins (for example, lysosomal integral membrane protein 2-mediated transport of β-glucocerebrosidase), which may enter the endo-lysosomal pathways at distinct stages of maturation (grey arrows). The black arrows represent multiple retrograde pathways from endosomes.

LAMPs and lysosomal integrity

Although most LMPs predominantly reside in lysosomes, their subcellular distributions can change and are more dynamic than so far appreciated. This has become especially apparent for LAMP1 and LAMP2. LAMPs are type-1 transmembrane proteins with considerable sequence homology that contain a large, heavily glycosylated luminal domain and a short cytosolic tail. For example, there are three LAMP2 isoforms with different transmembrane and cytosolic domains, which show a preferential localization in either endosomes, and lysosomes or the plasma membrane. More generally, the cell surface expression of LAMPs is increased in the activation of platelets, peripheral blood monocytes and cytotoxic T cells [35] and in highly malignant tumour cells [18]. Scientists are only beginning to understand the significance of local LAMP concentrations. The plasma membrane levels of CD63 are of major consequence for other local protein concentrations [36], but elevated plasma membrane levels of LAMP1 and LAMP2 have not yet been linked to a specific phenotype. An important clue to the significance of sustained LAMP levels in lysosomes came from a recent study which showed that LAMP proteins are involved in sensitizing tumour cells to lysosomal cell death (LCD) (Figure. 1). Oncogenic transformation of fibroblasts on the one hand leads to a decrease in the levels of LAMP proteins in lysosomes and on the other hand increases the sus-ceptibility of these cells to the LCD pathway [37]. Likewise, decreased levels of LAMP1 and LAMP2 also contribute to an enhanced sensitivity of transformed cells to anti-cancer drugs that trigger LCD. Overexpression of LAMPs had the opposite effect, indicating that LAMPs can protect cells from the LCD pathway. In addition, LAMP depleted cells showed a redistribution of lysosomes to the cell periph-ery, pointing to a role for LAMPs in lysosomal dynamics. The earlier studies correlating surface expression of LAMP proteins to metastatic potential of carcinoma cells [18, 38], exemplify the importance of LAMP targeting in maintaining lysosomal integrity and in regulating LCD pathways. They also underscore the need to understand the relationship between LAMP trafficking and successful anti-cancer treatment.

Lysosomal disorders

Lysosomal disorders represent a group of at least 40 genetic diseases [39], each of which results from a deficiency of one or more proteins involved in the degradation of macromolecules in lysosomes. . Initially the lysosomal membrane was considered to be only a mechanical border separating the acid lysosomal environment from the neutral surrounding cytoplasm. Since the discovery of a lysosomal cystine carrier, defective in an inherited human disease, more than 20 specific transport systems have been characterized in the lysosomal membrane. Most of them function as exporters and only a few as importers. Several types of lysosomal membrane transporters can be discriminated: solute carriers, pumps and channels. Each of the lysosomal transporters has a high specificity for groups of amino acids, sugars, nucleosides, inorganic ions, and vitamins. Genetic disorders of these transporters cause a wide array of neurological and visceral diseases, ranging from developmental to degenerative disorders. Until recently, all knowledge about lysosomal transport proteins was based on the biochemical (kinetic) characteristics of transport. The molecular and functional properties of the better characterized lysosomal transport systems and the related human diseases are discussed here.

The Lysosomal Membrane and Storage Diseases

Lysosomes are intracellular organelles acidified by a vacuolar-type (V-type) proton pump, which lowers the intraluminal pH to around 5. This acid environment is essential for several lysosomal functions, like enzymatic degradation, proton-coupled transport processes, receptor-ligand interactions, vesicle trafficking and sorting. In lysosomal storage diseases, undegraded macromolecules accumulate in the lysosomal compartment as a consequence of the mutation in one of the lysosomal hydrolases.[40, 41] However, in a few cases the substances accumulated in the lysosome are not undegraded macromoleculas but products of hydrolytic degradation that are supposed to leave the lysosomal compartment for metabolic recycling. In the group of lysosomal storage diseases, transport disorders represent rare

examples of inborn errors of metabolism caused by a defect of an intracellular membrane transport.

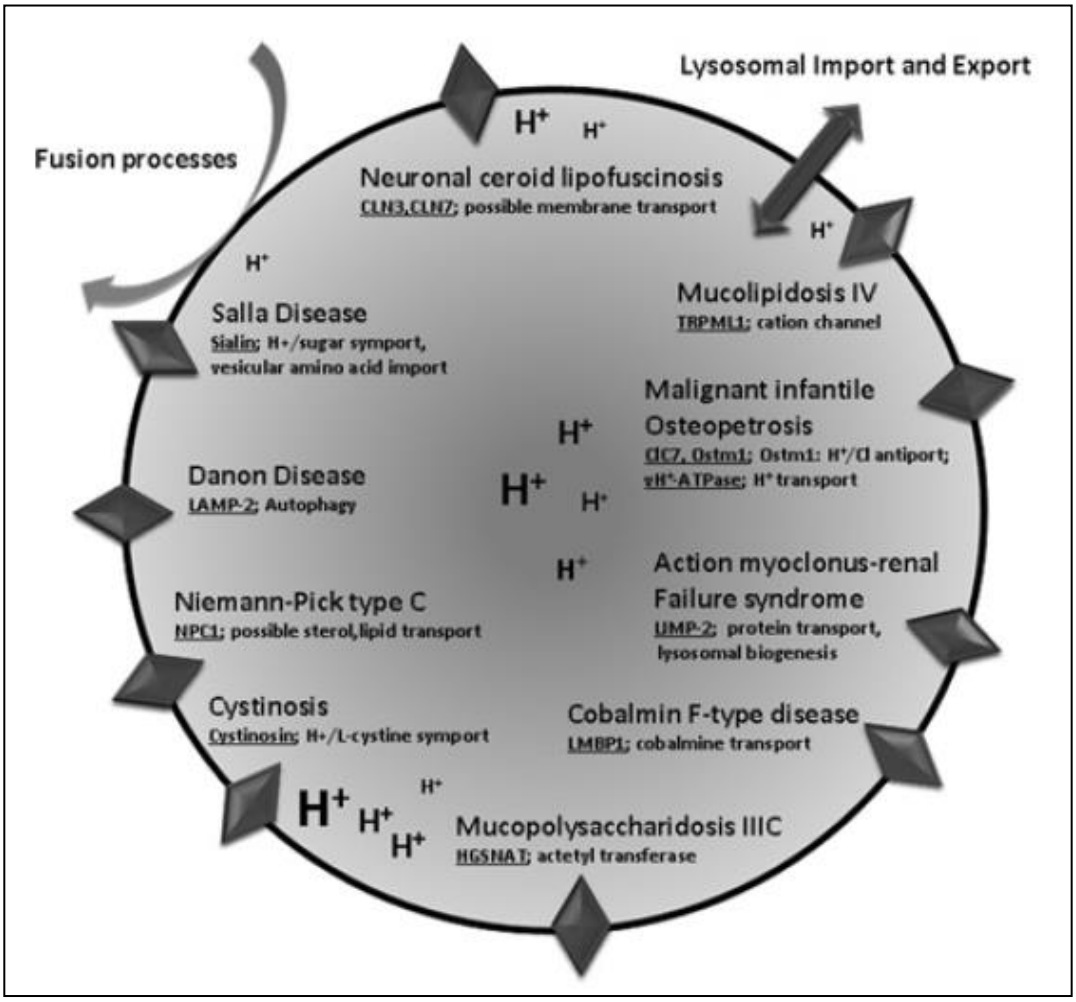


Figure 3: Schematic overview of the most relevant human diseases associated with mutations in lysosomal membrane proteins [9]. The arrows indicate that lysosomal membrane proteins are involved in transport processes through the lysosomal membrane and in the regulation of fusion of lysosomes with other cellular membranes.

Disease-causing mutations in genes encoding for some of these proteins have been only described in the last decade, showing that the rapid progress in this research field is due to the achievements of the human genome project. Here I have discussed the disorders caused by mutations in the solute carriers, ion channels and proton pump. The only disorder caused

by a defect of the heavily glycosylated integral membrane protein LAMP-2, Danon's disease is a defect in organelle transport and communication; it is not a defect in a transport process across the lysosomal membrane.

Solute Carriers and Transport Defects

The lysosomal membrane contains several specific carriers for the transport of solutes across the membrane. Most of the substrates transported by the lysosomal carriers are products of enzymatic degradation of macromolecules (single amino acids, dipeptides, monosaccharides, and lipids), but also specific carriers transport vitamins, heavy metals and drugs.[42] Many carriers with selective substrate specificity function as uniporters (passive transporters) following the Michaelis-Menten kinetics of transport along the substrate concentration gradient, or cotransporters (symportes and antiporters, secondary active transporters) coupled to an ion gradient, which provides the driving force for the direction of transport. This is usually the proton gradient generated by the energy-dependent vacuolar proton pump. Although more than 20 carriers have been characterised, only eight genes are known of which seven are coupled to a human disease.

CTNS encode for cystinosin the transporter defective in cystinosis, a lysosomal storage disease caused by intralysosomal storage of cystine crystals. SLC17A5 is the gene encoding for sialin, the sialic acid transporter defective in sialic acid storage disease. CLN3 encodes for a multimembrane-spanning protein, which is mainly localized in lysosomes in nonneuronal cells and in endosomes in neuronal cells. This protein is affected in Batten disease, a juvenile form of ceroid lipofuscinosis. NPC1 encodes for a new type of human permeases and is mutated in Niemann-Pick type C1 patients. [43] SLC36A1 encoding for a lysosomal transporter, LYAAT-1, of small neutral amino acids, like alanine, proline and GABA, has recently been identified as a member of the eukaryotic specific amino acids/auxin permease (AAAP) family, but is so far not coupled to a human disease [44].

This study emphasizes on the discrimination of lysosomal membrane protein types from various other types of membrane proteins as well as from the globular proteins on the basis of specific characteristics of membrane proteins.

Comparison of previous work on sub-cellular localization problem

Several methods have been proposed to discriminate membrane proteins from amino acid (AA) sequence information. These methods include statistical analysis, hidden markov models, and machine-learning techniques [45].

Methods	Jackknife test (%)
CDA (Chou & Elrod1999)	77.8
CDA and PseAA (Chou 2001)	76.58
AA composition and SVM (Cai et al. 2004)	86.79
Low frequency Fouriers pectrum (Liu et al. 2005)	81.5
Weighted u-SVM using PseAA (Wang et al. 2004)	89.5
PseAA and stacking (Wang et al. 2006)	88.7
Wavelet and cascade neural network (Rezaei et al. 2008)	86.8
Discrete wavelet and SVM (Qiu Sun Huang & Liang 2010)	78.13

Table 1: Comparison of previous work on subcellular localization prediction systems [45]

The pseudo-amino acid compositions (PseAAC) are used for the prediction of membrane protein types. Covariant discrimination algorithm was used by the Chou [46] in conjunction with pseudo-amino acid composition (PseAAC)-based feature extraction. For the improvement of the prediction accuracy of membrane protein types, Chou has carried out a series of works. Earlier days most of the machine-learning techniques were used for the prediction of membrane protein types like Yang et al. predicted membrane protein types on the basis of dipeptide as well as AA composition, Cai et.al [47] used support vector machine

and AA composition, and Sonnhammer et.al [48] have used the hidden Markov model for predicting topology of membrane protein types. Similarly, Liu [49] have employed the Fourier spectrum and SVM, while Wang et.al [50] have used weighted SVM and PseAA composition. Wang et.al [51] have used PseAA and stacked generalization. Chou and Shen [52, 53] developed a web server for the prediction of membrane protein types.

None of these projects are specific for Lysosomal membrane proteins. Those proteins are classified either as Lysosomal proteins or misclassified as plasma membrane proteins. Tripathi V. et.al [45] proposed the ANN approach for classifying LMPs having good accuracy, but this project is not available on web.

In this work, I have analyzed the AA composition and dipeptide composition along with PSSM matrices in order to improve the prediction quality.

CHAPTER 2

MATERIALS AND METHODS

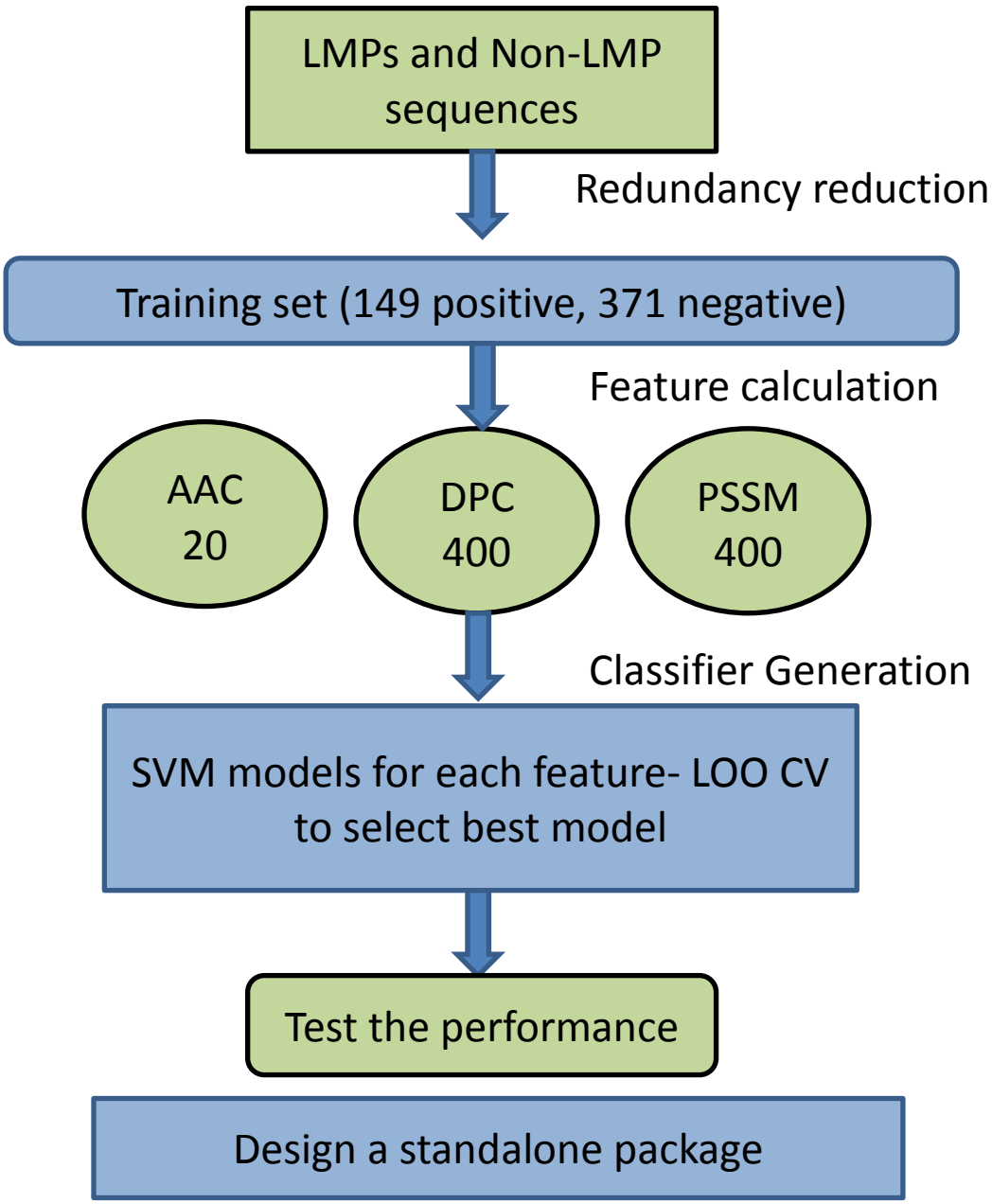


Figure 4: Flowchart for the method discussed in this study

Step 1: Data set collection

Data Sources for Positive and negative datasets

UniProtKB:

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.

The UniProt Knowledgebase consists of two sections: a section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis, and a section with computationally analyzed records that await full manual annotation. For the sake of continuity and name recognition, the two sections are referred to as "UniProtKB/Swiss-Prot" (reviewed, manually annotated) and "UniProtKB/TrEMBL" (unreviewed, automatically annotated), respectively.

Pfam:

Pfam is a comprehensive collection of protein domains and families, represented as multiple sequence alignments and as profile hidden Markov models. The current release of Pfam (27.0, March 2013) contains 14831 curated protein families. Pfam is now based not only on the UniProtKB sequence database, but also on NCBI GenPept and on sequences from selected metagenomics projects. Pfam is available on the web from the consortium members using a new, consistent and improved website design in the UK (http://pfam.sanger.ac.uk/). ‘Sequence coverage’ is the fraction of protein sequences listed in UniProtKB that has at least one Pfam domain, whilst ‘residue coverage’ is the fraction of protein residues that fall within

Pfam domains, as defined by the sub-sequences included in Pfam-A full alignments. Pfam version 27.0 was produced at the European Bioinformatics Institute using a sequence database called Pfamseq, which is based on UniProt release 2012_06.

- Positive dataset
    - Lysosomal membrane proteins (pfam 27.0)- 173 proteins
    - Human LMPs (Ref: Schroder et.al)- 40 proteins
- Negative dataset
    - Globular proteins(UniprotkB release:2013_10)- 1159 proteins
    - Plasma membrane proteins (Ref: Park et.al)-1674 proteins

Step2: Redundancy reduction

CD-HIT stands for Cluster Database at High Identity with Tolerance. The program takes a fasta format sequence database as input and produces a set of 'non-redundant' (nr) representative sequences as output. In addition CD-HIT outputs a cluster file, documenting the sequence 'groupies' for each nr sequence representative. The idea is to reduce the overall size of the database without removing any sequence information by only removing 'redundant' (or highly similar) sequences. This is why the resulting database is called non-redundant (nr). Essentially, cd-hit produces a set of closely related protein families from a given fasta sequence database.

CD-HIT uses a 'longest sequence first' list removal algorithm to remove sequences above a certain identity threshold. Additionally the algorithm implements a very fast heuristic to find high identity segments between sequences, and so can avoid many costly full alignments. With recent developments; CD-HIT package offers new programs for DNA sequence clustering and comparing two databases. It also has lots of new options for clustering control.CD-HIT was originally written by Weizhong Li and is now an open source project.

- Positive data-

  * Total (173+40)-213
  * Reduction (50%identity)-149
- Negative data–

  * Total (1159+1674)-2833
  * Reduction (40%identity)-1458

Negative data used for further analysis- 371

Step 3: Feature calculation

Amino acid composition

The AA frequency of any protein depends on 20 discrete numbers. In AA composition, proteins can be expressed in 20 dimensional vectors [].

$$AAfreq = [f1, f2, f3, \ldots\ldots, f20]$$

Where, $f1, f2, f3, \ldots, f20$ are the frequencies of the 20 AAs of a protein.

Dipeptide composition

The correlation between the dipeptide composition and the stability of the proteins are well established. The primary determinants of the stability of the protein probably reside in its primary structure is an intrinsic property of a protein. There appears to be a correlation between the sensitivity of a protein to in vivo degradation and the presence of certain dipeptides in it. The composition of all the 400 dipeptides based on the distribution of AA residues along the sequences proteins has been computed using the following expression:

$$\mathrm{Di}_{\mathrm{comp(i,j)}} = \frac{\sum \mathrm{N_{ij}}}{\sum \mathrm{N_i} + \sum \mathrm{N_j}}$$

Where i, j stands for the distribution of 20 AA residues at positions i and i + 1. Nij is the number of residues of type i followed by the residue j. ΣNi and ΣNj are the total number of residues of type i and j, respectively.

PSSM (Position Specific Scoring Matrix)

It is a commonly used representation of motifs (patterns) in biological sequences. PSSMs are often derived from a set of aligned sequences that are thought to be functionally related and have become an important part of many software tools for computational motif discovery.

The position weight matrix was introduced by American geneticist Gary Stormo and colleagues in 1982 [55] as an alternative to consensus sequences. Consensus sequences had previously been used to represent patterns in biological sequences, but had difficulties in the prediction of new occurrences of these patterns. [56] The first use of PWMs was in the discovery of RNA sites that function as translationinitiation sites. The perceptron algorithm was suggested by Polish American mathematician Andrzej Ehrenfeucht in order to create a matrixof weights which could distinguish true binding sites from other non-functional sites with similar sequences. Training the perceptron on both sets of sites resulted in a matrix and a threshold to distinguish between the two sets.[55] Using the matrix to scan new sequences not included in the training set showed that this method was both more sensitive and precise than the best consensus sequence.

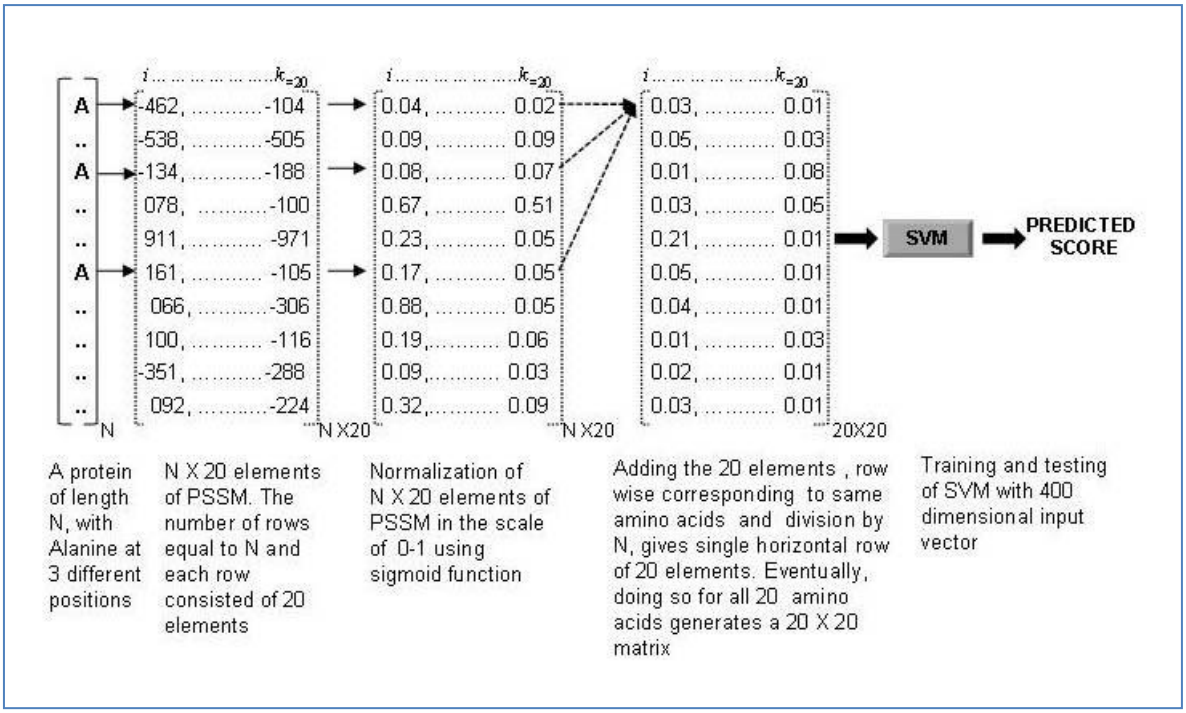


Figure 5: Conversion of PSSM into training vectors. The steps used to convert PSSM profiles generated by PSI-BLAST into a training vector of 400 dimensions [54].

Step 4: Modeling and optimizing classifier

HMMER

HMMER is used to search sequence databases for homologs of protein or DNA sequences, and to make sequence alignments. HMMER can be used to search sequence databases with single query sequences but it becomes particularly powerful when the query is an alignment of multiple instances of a sequence family. HMMER makes a profile of the query that assigns a position-specific scoring system for substitutions, insertions, and deletions. HMMER profiles are probabilistic models called "profile hidden Markov models" (profile HMMs) [57]. Compared to BLAST, FASTA, and other sequence alignment and database search tools based on older scoring methodology, HMMER aims to be significantly more accurate and more able to detect remote homologs, because of the strength of its underlying probability models. In the past, this strength came at a significant computational cost, with profile HMM implementations running about 100x slower than comparable BLAST searches for protein search, and about 1000x slower than BLAST searches for DNA search. With HMMER3.1, HMMER is now essentially as fast as BLAST for protein search.

Procedure:

- Build a profile HMM of positive dataset.
- Iteratively search a protein sequence against a protein sequence database. (PSIBLAST-like)
- Search a protein profile HMM against a protein sequence database.

SUPPORT VECTOR MACHINE

Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier.

Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

SVMs (Support Vector Machines) are a useful technique for data classification. Although SVM is considered easier to use than Neural Networks, users not familiar with it often get unsatisfactory results at first. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels) and several "attributes" (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

Given a training set of instance-label pairs (xi; yi); i $= 1, \ldots, l$ where xi $\in$ Rn and y $\in \{1, -1\}^l$, the support vector machines (SVM) [58] require the solution of the following optimization problem:

$$\min_{w,b,\varepsilon} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \varepsilon_i$$

$$\text{Subject to} \quad y_i(W^T \emptyset(X_i) + b) \geq 1 - \varepsilon_i$$

$$\varepsilon_i \geq 0$$

Here training vectors Xi are mapped into a higher (maybe ifinite) dimensional space by the function $\emptyset$. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term.

Furthermore, $K(xi, xj) \equiv \emptyset(xi)^T \emptyset(xj)$ is called the kernel function.

Basic four kernels of SVM:

- Linear: $K(x_i;\ x_j)\ =\ x_i^T xj$.
- Polynomial: $K(x_i, x_j) = \left(\gamma x_i^T x_j\ +\ r\right)^d,\ \gamma > 0$.
- Radial basis function (RBF): $K(x_i;\ x_j) =\ exp\left(-\gamma\left\|x_i - x_j\right\|^2\right), \gamma > 0.$
- Sigmoid: $K(x_i, x_j)\ =\ tanh(\gamma x_i^T\ x_j\ +\ r)$.

Here,γ, r, and d are kernel parameters.

Procedure:

- Transform data to the format of an SVM package
- Conduct simple scaling on the data
- Consider the RBF kernel $K(x, y)\ =\ e^{-\gamma\|x-y\|^2}$
- Use cross-validation to find the best parameter C and
- Use the best parameter C and to train the whole training set
- Test

Though there are only four common kernels mentioned before, we must decide which one to try first. Then the penalty parameter C and kernel parameters are chosen.

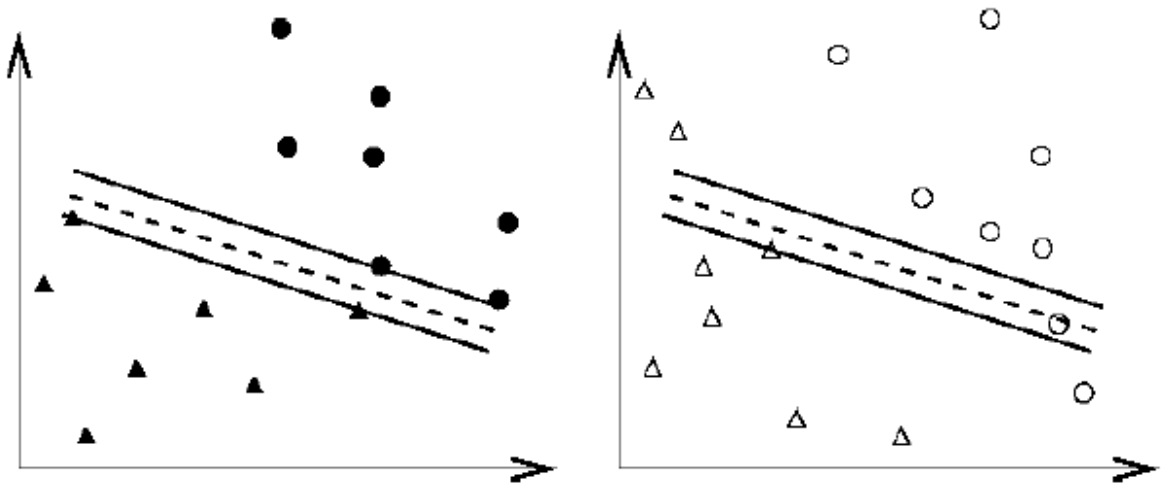
RBF Kernel

In general, the RBF kernel is a reasonable first choice. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. Furthermore, the linear kernel is a special case of RBF [59] since the linear kernel with a penalty parameter ~ C has the same performance as the RBF kernel with some parameters (C, γ). In addition, the sigmoid kernel behaves like RBF for certain parameters [60]. The second reason is the number of hyperparameters which influences the complexity of model selection. The polynomial kernel

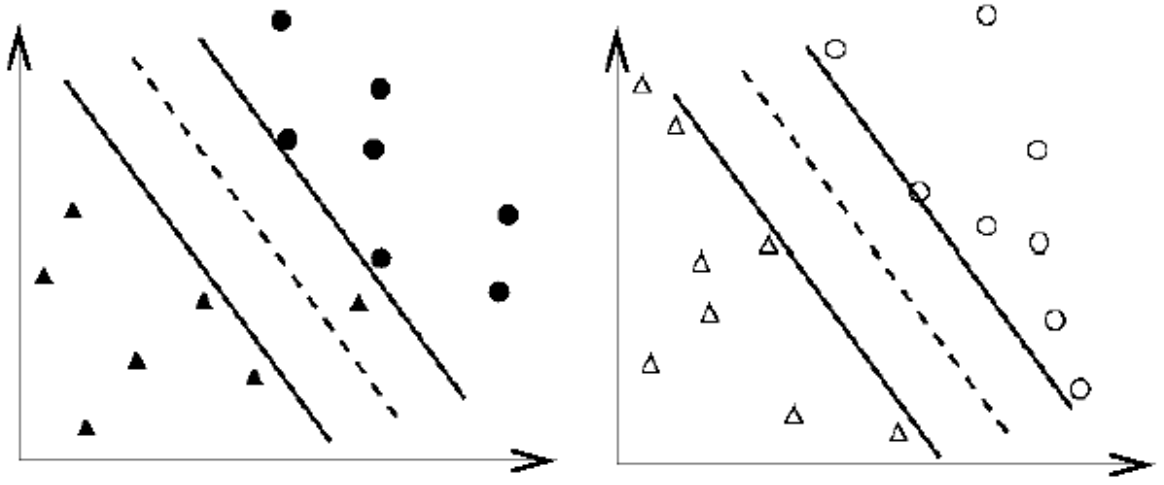
has more hyperparameters than the RBF kernel. Finally, the RBF kernel has fewer numerical dificulties. One key point is 0 < Kij _ 1 in contrast to polynomial kernels of which kernel values may go to infinity $(\gamma x_i^T x_j + r > 1)\ or\ zero\ (\gamma x_i^T x_j + r < 1)$ while the degree is large. Moreover, we must note that the sigmoid kernel is not valid (i.e. not the inner product of two vectors) under some parameters [58]. There are some situations where the RBF kernel is not suitable. In particular, when the number of features is very large, one may just use the linear kernel.

Cross-validation and Grid-search

There are two parameters for an RBF kernel: C and γ. It is not known beforehand which C and γ are best for a given problem; consequently some kind of model selection (parameter search) must be done. The goal is to identify good (C, γ) so that the classifier can accurately predict unknown data (i.e. testing data). Note that it may not be useful to achieve high training accuracy (i.e. a classifier which accurately predicts training data whose class labels are indeed known). As discussed above, a common strategy is to separate the data set into two parts, of which one is considered unknown. The prediction accuracy obtained from the "unknown" set more precisely reflects the performance on classifying an independent data set. An improved version of this procedure is known as cross-validation. In v-fold cross-validation, we first divide the training set into 5 subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining $v - 1$ subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified. The cross-validation procedure can prevent the overfitting problem.



(a) Training data and an overfitting classifier (b) Applying an overfitting classifier on testing data



(c) Training data and a better classifier (d) Applying a better classifier on testing data

Figure 6: An overfitting classifier and a better classifier (● and ▲: training data; O and Δ; testing data).

Figure 6 represents a binary classification problem to illustrate this issue. Filled circles and triangles are the training data while hollow circles and triangles are the testing data. The

testing accuracy of the classifier in Figures 6a and 6b is not good since it overfits the training data. If we think of the training and testing data in Figure 6a and 6b as the training and validation sets in cross-validation, the accuracy is not good. On the other hand, the classifier in 6c and 6d does not overfit the training data and gives better cross-validation as well as testing accuracy.

I have done a "grid-search" on C and γ using cross-validation. Various pairs of (C,γ) values are tried and the one with the best cross-validation accuracy is picked. I found that trying exponentially growing sequences of C and γ is a practical method to identify good parameters.

Cross-validation methods

I performed training testing cycles using self-written perl scripts. Where I used linear, polynomial and radial basis function (RBF) kernels to train and test my SVM models. Each kernel was optimized to yield the best classification by changing the kernel parameters (C, d and γ). This approach was to choose the best parameters in a way so as to maximize accuracy as well as get nearly equal sensitivity and specificity, wherever possible.

Leave-one-out cross validation (LOO CV):

This is a stringent mode of evaluation wherein one dataset sequence is left out for testing, while the rest are used to generate the model. This is iterated on each sequence till each sequence becomes the testing data exactly once. The best parameters as measured by the various performance measures are picked up and then averaged for the final assessment of the model. It has been shown to give an almost unbiased estimator of the generalisation properties of statistical models, and therefore provides a sensible criterion for model selection and comparison.

Performance measures

In order to assess the accuracy of prediction methods, I used several measures, namely-

- **Sensitivity**: percentage of LMP protein sequences that are correctly predicted as LMP,

$$Sensitivity = \frac{TP}{TP + FN} \times 100$$

- **Specificity**: percentage of non- LMP protein sequences that are correctly predicted as non- LMP,

$$Specificity = \frac{TN}{TN + FP} \times 100$$

- **Accuracy**: percentage of correct predictions, for LMP as well as non- LMP, and

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100$$

- **Matthews Correlation Coefficient (MCC):** a measure of both sensitivity and specificity (MCC = 1 indicates a perfect prediction while MCC = 0 indicates a totally random prediction.

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

Where, TP is the number of True Positives, TN is the number of True Negatives, FN is the number of False Negatives, and FP is the number of False Positives for a prediction method.

CHAPTER 3
RESULTS AND DISCUSSION

Performance of standalone SVM models

I began with the LOO Cross-Validation of AAC, DPC and PSSM based classifiers, trained using different kernels like linear, polynomial and RBF (Radial Basis Function). Thereafter, hybrid model using combination of PSSM and AAC features was also developed. The hold-out procedure was performed for the best classifiers to further assess the discriminative quality of the models. Hold-out method provides a further reinforcement about the discriminative power, though because of the random partitioning of the datasets, the results may vary considerably for the different sets. Table 2 summaries the performance of the best SVM classifiers for each module as observed in the cross-validation tests.

Composition based SVM classifiers

I obtained accuracies of approximately ~60% in AAC-based SVM models with different kernels, and 61.15 % with the RBF kernel. The accuracies increased with PSSM usage and reached 68.84% for RBF kernel. However, the DPC model yielded very low accuracies of ~55% for other kernel whereas with RBF, the accuracy touched ~58%. The sensitivity and specificity of this model were also not good. The remarkably better performance of AAC and DPC models can be achieved with the known structural conservation of LMPs

PSSM profile based SVM classifier

PSI-BLAST derived PSSM profiles captures useful information about the residue composition as well as conservation of residues at crucial positions within the protein sequence, because in evolution the amino acid residues with similar physico-chemical properties tend to be highly conserved due to selective pressure. PSSM profiles have been used as SVM input feature for a number of classification problems, e.g. prediction of sub-cellular localization.

I used the PSSM profile, normalized using the logistic function for developing an SVM module. The PSSM profile-based model yielded maximal accuracies of ~67% different kernels, and a remarkably high accuracy of 68.84% with the RBF kernel.

Performance of hybrid SVM models

With an aim to further enhance the prediction accuracy, I developed and evaluated hybrid model using combination of AAC and PSSM. This was the model with the highest overall accuracy of 71.92% better than both the PSSM and AAC models, but with a lower specificity (56.96%) and higher sensitivity (83.42%) as that of the PSSM model. The accuracy was 73.94% for RBF kernel. This model achieved the best overall accuracy amongst all the models.

Feature	Kernel	Parameters			SN(%)	SP(%)	Acc(%)	MCC
		Threshold	C	γ				
AAC	RBF	-0.5	300	0.02	46.3	67.11	61.15	0.13
DPC	RBF	-0.8	500	0.5	53.69	59.83	58.07	0.11
PSSM	RBF	-0.6	600	0.09	53.02	71.42	68.84	0.395
PSSM+AAC	RBF	-0.1	-	0.001	83.42	56.96	73.94	0.665

Table 2: Performance of SVM classifiers for various combinations of training features, kernels and parameters for leave one out cross validation.

CHAPTER 4
CONCLUSION

Current Lysosomal membrane protein prediction methods include experimental determination which requires enormous efforts and computational methods which are not specific.. The study presented here represents an initiative towards easy identification of LMPs from other proteins effectively. Apart from solving the LMP identification problem in particular, it advocates and reinforces the rational application of machine-learning algorithms like SVMs to classification problems in biology. The study could be extended to other protein families sharing low pairwise sequence similarity. Though identification of a protein sequence as a LMP would speak little about function because of the high functional versatility, yet it would provide significant clues about the protein structure and hence lead the way towards providing mechanistic insights about the protein. Since user-friendly and publicly accessible web servers represent the future direction for developing practically more useful models or predictors, it shall be included in future work to provide a web server for the method presented here.

REFERENCES

[1] Dell'Angelica, E. C., Mullins, C., Caplan, S. & Bonifacino,J. S. Lysosome-related organelles. FASeB J. 14, 1265–1278 (2000).

[2] Bonifacino, J. S. Insights into the biogenesis of lysosome-related organelles from the study of the Hermansky–Pudlak syndrome. Ann. NY Acad. Sci.1038, 103–114 (2004).

[3] Dell'Angelica, E. C. The building BLOCks of lysosomes and related organelles. Curr. Opin. Cell Biol. 16, 458–464 (2004).

[4] Conus, S. & Simon, H. U. Cathepsins: key modulators of cell death and inflammatory responses. Biochem. Pharmacol. 76, 1374–1382 (2008).

[5] Lübke, T., Lobel, P. & Sleat, D. E. Proteomics of the lysosome. Biochim. Biophys. Acta 1793, 625–635 (2009).

[6] Schroder, B. et al. Integral and associated lysosomal membrane proteins. Traffic 8, 1676–1686 (2007).

[7] Callahan, J. W., Bagshaw, R. D. & Mahuran, D. J. The integral membrane of lysosomes: its proteins and their roles in disease. J. Proteomics 72, 23–33 (2009).

[8] Eskelinen, E. L., Tanaka, Y. & Saftig, P. At the acidic edge: emerging functions for lysosomal membrane proteins. Trends Cell Biol. 13, 137–145 (2003)

[9] Saftig P, Schröder B, Blanz J (2010) Lysosomal membrane proteins: life between acid and neutral conditions. Biochem Soc Trans; 38(6): 1420-1423.

[10] Granger BL, Green SA, Gabel CA, Howe CL, Mellman I, Helenius A (1990) Characterization and cloning of lgp110, a lysosomal membrane glycoprotein from mouse and rat cells. J Biol Chem; 265(20): 12036-12043.

[11] Schröder BA, Wrocklage C, Hasilik A, Saftig P (2010) The proteome of lysosomes. Proteomics; 10(22): 4053-4076.

[12] Schulze H, Kolter T, Sandhoff K (2009) Principles of lysosomal membrane degradation: Cellular topology and biochemistry of lysosomal lipid degradation. Biochim Biophys Acta; 1793(4): 674-683.

[13] Lübke T, Lobel P, Sleat DE (2009) Proteomics of the lysosome. Biochim Biophys Acta; 1793(4): 625-635.

[14] Saftig P, Klumperman J (2009) Lysosome biogenesis and lysosomal membrane proteins: trafficking meets function. Nat Rev Mol Cell Biol; 10(9): 623-635.

[15] Coutinho MF, Prata MJ, Alves S (2012) Mannose-6-phosphate pathway: a review on its role in lysosomal function and dysfunction. Mol Genet Metab; 105(4): 542-550

[16] Mindell JA (2012) Lysosomal acidification mechanisms. Annu Rev Physiol; 74: 69-86.

[17] Möbius W, van Donselaar E, Ohno-Iwashita Y, Shimada Y, Heijnen HF, Slot JW et al. (2003) Recycling compartments and the internal vesicles of multivesicular bodies harbor most of the cholesterol found in the endocytic pathway. Traffic; 4(4): 222-231.

[18] Fukuda M (1991) Lysosomal membrane glycoproteins. Structure, biosynthesis, and intracellular trafficking. J Biol Chem; 266(32): 21327-21330

[19] Danon MJ, Oh SJ, DiMauro S, Manaligod JR, Eastwood A, Naidu S et al. (1981) Lysosomal glycogen storage disease with normal acid maltase. Neurology; 31(1): 51-57.

[20] Vega, M.A., Segui-Real, B., Garcia, J.A., Cales, C., Rodriguez, F., Vanderkerckhove, J. and Sandoval, I.V. (1991) Cloning, sequencing, and expression of a cDNA encoding rat LIMP II, a novel 74-kDa lysosomal membrane protein related to the surface adhesion protein CD36. J. Biol. Chem. 266, 16818–16824

[21] Kuronita, T., Eskelinen, E.L., Fujita, H., Saftig, P., Himeno, M. and Tanaka, Y. (2002) A role for the lysosomal membrane protein LGP85 in the biogenesis and maintenance of endosomal and lysosomal morphology. J. Cell Sci. 115, 4117–4131

[22] Reczek, D., Schwake, M., Schroder, J., Hughes, H., Blanz, J., Jin, X.,Brondyk, W., Van Patten, S., Edmunds, T. and Saftig, P. (2007) LIMP-2 is a receptor for lysosomal mannose-6-phosphate-independent targeting of β-glucocerebrosidase. Cell 131, 770–783

[23] Berkovic, S.F., Dibbens, L., Oshlack, A., Silver, J., Katerelos, M., Vears, D.V., Lullmann-Rauch, R., Blanz, J., Zhang, K.W., Stankovich, J. et al.(2008) Array based gene discovery with 3 unrelated subjects shows SCARB2/LIMP-2 deficiency causes myoclonus epilepsy and glomerulosclerosis. Am. J. Hum. Genet. 82, 673–684

[24] Blanz, J., Groth, J., Zachos, C., Wehling, C., Saftig, P. and Schwake, M. (2010) Disease-causing mutations within the lysosomal integral membrane protein type 2 (LIMP-2) reveal the nature of binding to its ligand β-glucocerebrosidase. Hum. Mol. Genet. 19, 563–572

[25] Metzelaar, M.J., Wijngaard, P.L., Peters, P.J., Sixma, J.J., Nieuwenhuis, H.K.and Clevers, H.C. (1991) CD63 antigen: a novel lysosomal membrane glycoprotein, cloned by a screening procedure for intracellular antigens in eukaryotic cells. J. Biol. Chem. 266, 3239–3245

[26] Schroder, J., Lullmann-Rauch, R., Himmerkus, N., Pleines, I., Nieswandt,B., Orinska, Z., Koch-Nolte, F., Schroder, B., Bleich, M. and Saftig, P.(2009) Deficiency of the tetraspanin CD63 associated with kidney pathology but normal lysosomal function. Mol. Cell. Biol. 29, 1083–1094

[27] Kornfeld, S. & Mellman, I. The biogenesis of lysosomes. Annu. Rev. Cell Biol. 5, 483–525 (1989).

[28] Figura, K. V. & Hasilik, A. Lysosomal enzymes and their receptors. 55, 167–193 (1986).

[29] Ruivo, R., Anne, C., Sagne, C. & Gasnier, B. Molecular and cellular basis of lysosomal transmembrane protein dysfunction. Biochim. Biophys. Acta 1793, 636–649 (2009).

[30] Sachse, M., Ramm, G., Strous, G. & Klumperman, J. Endosomes: multipurpose designs for integrating housekeeping and specialized tasks. Histochem. Cell Biol. 117, 91–104 (2002).

[31] Stoorvogel, W., Strous, G. J., Geuze, H. J., Oorschot, V. & Schwartz, A. L. Late endosomes derive from early endosomes by maturation. Cell 65, 417–427 (1991).

[32] Murphy, R. F. Maturation models for endosome and lysosome biogenesis. Trends Cell Biol. 1, 77–82 (1991).

[33] Luzio, J. P., Pryor, P. R. & Bright, N. A. Lysosomes: fusion and function. Nature Rev. Mol. Cell Biol. 8, 622–632 (2007).

[34] Mari, M. et al. SNX1 defines an early endosomal recycling exit for sortilin and mannose 6-phosphate receptors. Traffic 9, 380–393 (2008).

[35] Kannan, K. et al. Lysosome-associated membrane proteins h-LAMP1 (CD107a) and h-LAMP2 (CD107b) are activation-dependent cell surface glycoproteins in human peripheral blood mononuclear cells which mediate cell adhesion to vascular endothelium. Cell. Immunol. 171, 10–19 (1996).

[36] Pols, M. S. & Klumperman, J. Trafficking and function of the tetraspanin CD63. exp. Cell Res. 315,1584–1592 (2008).

[37] Fehrenbacher, N. et al. Sensitization to the lysosomalcell death pathway by oncogene-induced downregulation of lysosome-associated membrane proteins 1 and 2. Cancer Res. 68, 6623–6633 (2008).

[38] Saitoh, O., Wang, W. C., Lotan, R. & Fukuda, M. Differential glycosylation and cell surface expression of lysosomal membrane glycoproteins in sublines of a human colon cancer exhibiting distinct metastatic potentials. J. Biol. Chem. 267, 5700–5711 (1992).

[39] Hopwood JJ, Brooks DA. An introduction to the basic science and biology of the lysosome and storage diseases. Applegarth DA Dimmick JE Hall JG eds. Organelle diseases 1997:7-35.

[40] Mancini GMS, Verheijen FW. Lysosomal storage diseases. In: Bittar EE, Bittar N, eds. Principles of Medical Biology: Cellular Organelles and the Extracellular Matrix. London: Jai Press inc., 1995:3:133-154.

[41] Gahl WA, Schneider JA, Aula P. Lysosomal transport disorders: Cystinosis and sialic acid storage disorders. In: Scriver CR, Beaudet AL, Sly WS, Valle D, eds. The Metabolic and Molecular Bases of Inherited Disease. 7th ed. New York: McGraw-Hill, 1995:3763-3797.

[42] Mancini GM, Havelaar AC, Verheijen FW. Lysosomal transport disorders. J Inherit Metab Dis 2000; 23:278-92.

[43] Davies JP, Chen FW, Ioannou YA. Transmembrane molecular pump activity of Niemann-Pick C1protein. Science 2000; 290:2295-8.

[44] Sagne C, Agulhon C, Ravassard P et al. Identification and characterization of a lysosomal transporter for small neutral amino acids. Proc Natl Acad Sci USA 2001; 98:7206-11.

Bioinformatics Research papers:

[45] Tripathi V. et.al, Discriminating lysosomal membrane protein types using dynamic neural network, Journal of Biomolecular Structure and Dynamics (2013), http://dx.doi.org/10.1080/07391102.2013.827133

[46] Chou, K. C. (2001). Prediction of protein subcellular attributesusing pseudo-amino acid composition. Proteins: Structure, Function, Genetics, 43, 246–255.

[47] Cai, Y. D., Ricardo, P. W., Jen, C. H., & Chou, V. (2004).Application of SVM to predict membrane protein types. Journal of Theoretical Biology, 226, 373–376.

[48] Sonnhammer, E. L. L., Heijne, G. V., & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. In Proceedings of Sixth International Conference on Intelligent Systems for Molecular Biology (pp. 175–182). Menlo Park, CA: AAAI/MIT Press, 6.

[49] Liu, H., Wang, M., & Chou, K. C. (2005). Low-frequency Fourier spectrum for predicting membrane protein types. Biochemical and Biophysical Research Communications, 336, 737–739.

[50] Wang, M., Yang, J., Liu, G. P., XuZ, J., & Chou, K. C. (2004). Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Engineering Design and Selection, 17, 509–516.

[51] Wang, S. Q., Yang, J., & Chou, K. C. (2006). Using stacking generalization to predict membrane protein types based on pseudo-amino acid. Journal of Theoretical Biology, 242, 941–946.

[52] Chou, K. C., & Shen, H. B. (2007). MemType-2L: A web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochemical and Biophysical Research Communications, 360, 339–345.

[53] Chou, K. C., & Shen, H. B. (2009). Review: Recent advances in developing web-servers for predicting protein attributes. Natural Science, 2, 63–92.

[54] Garg A, Gupta D: VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics 2008, 9:62.

[55] Stormo, Gary D.; Schneider, Thomas D.; Gold, Larry; Ehrenfeucht, Andrzej (1982). "Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli".Nucleic Acids Research 10 (9): 2997–3011.

[56] Stormo, G. D. (1 January 2000). "DNA binding sites: representation and discovery".Bioinformatics 16 (1): 16–23.

[57] Krogh, A. (1998). An introduction to hidden Markov models for biological sequences. In Salzberg, S., Searls, D., and Kasif, S., editors, Computational Methods in Molecular Biology, pages 45–63. Elsevier

[58] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages 144-152. ACM Press, 1992.

[59] S. S. Keerthi and C.-J. Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation, 15(7):1667{1689, 2003.

[60] H.-T. Lin and C.-J. Lin. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University, 2003. URL http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf.

Data Sets:

Positive:

1. Schroder et.al (2007), Integral and Associated Lysosomal Membrane Proteins, Traffic, Volume 8, Issue 12, pages 1676–1686, December 2007
2. Pfam- http://pfam.xfam.org/

Negative:

1. Park K, et al. Differentiation between transmembrane and peripheral helices by the deconvolution of circular dichroism spectra of membrane proteins. Protein Sci.1992;1:1032-1049.
2. http://www.uniprot.org/help/uniprotkb

Package Sourse:

i. CD-HIT(-2009-0421-win32): http://www.bioinformatics.org/cd-hit/
ii. HMMER(V3.1b1): http://hmmer.janelia.org/
iii. SVM_light (V 6.02): *http://svmlight.joachims.org/*

APPENDIX

Perl Scripts

---

1. Calculation of amino acid composition

```
#! /usr/bin/perl -w
use strict;
open (S, "negDataOut.fasta") || die "cannot open FASTA file to read: $!";
my %s;
my %seq;
my $key;
while (<S>){
chomp;
   if (/>/){
      s/>//;
      $key= $_;
   }else{
      push (@{$s{$key}}, $_);
   }
}
foreach my $a (keys %s){
   my $s= join("", @{$s{$a}});
   $seq{$a}=$s;
   #print("$a\t$s\n");
}
my @aa= qw(A R N D C Q E G H I L K M F P S T W Y V);
open (FH,'>>aa_neg.txt');
```

```
foreach my $k (keys %seq){
   my %count;
my @seq= split(//, $seq{$k});
   foreach my $r(@seq){
      $count{$r}++;
   }
my @row;
my $i=1;
   foreach my $a (@aa){
      my $final;
     $final.=$i;
     $final.=":";
     $count{$a}||=0;
     $count{$a}= sprintf("%0.3f", ($count{$a})*100/length($seq{$k}));
     $final.=$count{$a};
      push(@row,$final);
   $i++;
   }
my $row= join("\t",@row);
print FH "-1\t$row\n";
}
close FH;
```

2. Calculation of dipeptide composition

```
use Getopt::Std;
getopts('i:o:');
$file1=$opt_i;
$file2=$opt_o;

$aa = "#ACDEFGHIKLMNPQRSTVWY";

   open(FP1,"$file1");
   open(FP2,">$file2");
   while($t1=<FP1>){
   chomp($t1);
   uc($t1);
   $c1 = substr($t1,0,1);
   if($c1 =~ ">")
  {
    @ti = split("##",$t1);
    @ti1 = split("",$ti[1]);
    $le = length ($ti[1]);
    $len=$le-1;
    for($i1=1; $i1 <= 20; $i1++){
       for($i2=1; $i2 <= 20; $i2++)
      {
        $comp[$i1][$i2]=0;
      }
   }
   for($j1 = 0; $j1 < $#ti1; $j1++){
     $c1 = $ti1[$j1];
```

```
      $in1 = index($aa,$c1);
      $c2 = $ti1[$j1+1];
      $in2 = index($aa,$c2);
      $comp[$in1][$in2]++;
    }
    $count=0;
    $svm=1;
  print FP2 "+1\t";
 for($i1=1; $i1 <= 20; $i1++)
 {
    for($i2=1; $i2 <= 20; $i2++)
     {
       $perc=(($comp[$i1][$i2])*100)/$len;
        $count++;
         if($count <= 399)
        { printf(FP2 "%d:%5.3f\t",$svm,$perc); }
          else
         { printf(FP2 "%d:%5.3f\t",$svm,$perc); }
          $svm++;
       }
  }
     print FP2 "\n";
}
}
close FP1;
close FP2;
```

3. Calculation of PSSM

Code1: (scriptforblast.pl)

```
#!/usr/bin/perl -w
#use strict;
my $file = $ARGV[0];
open FH,$file;
my $file1;
my @query = <FH>;
my $i = 0;
my $fastal;
my $fastal1;
my $liness;
my $count =0;
close (FH);
my $count1 =1;
foreach my $xyz (@query)
{
if ($xyz =~ />/ && $count == 0)
{$count = 1;
$liness = $xyz;
$i++;
next;
}
if ($xyz !~ />/)
{
$liness = "$liness$xyz";
next;
}
```

```
if ($xyz =~ />/ && $count == 1)
{
$file1 = "file_$i";
open (FH1,">$file1");
print FH1 "$liness";
close (FH1);
system "blastpgp -d $ARGV[1] -i $file1 -j 3 -h 0.001 -m 0 -C $file1.chk";
system "echo $file1.chk > $count1.pn";
#system "rm $file1";
system "echo  $file1 > $count1.sn";
system "makemat -P $count1";
system "rm file_$i.fasta $count1.pn $count1.sn $file1.chk $file1 $count1.*";
$liness = $xyz;
$count1++;
$i++;
next;
}
}
$file1 = "file_$i";
open (FH1,">$file1");
print FH1 "$liness";
close (FH1);
system "blastpgp -d $ARGV[1] -i $file1 -j 3 -h 0.001 -m 0 -C $file1.chk";
system "echo $file1.chk > $count1.pn";
#system "rm $file1";
system "echo  $file1 > $count1.sn";
system "makemat -P $count1";
system "rm  $file1.chk $file1 $count1.*";
```

Code 2: (scriptforblastcompleterun.pl)

```
#!/usr/bin/perl -w
use strict;
my @aaray1;
system "mkdir tempBlastDatabase";
#system "cp $ARGV[1] tempBlastDatabase";
if ($ARGV[1] =~/\//)
{
@aaray1 = split (/\//,$ARGV[1]);
$ARGV[1] = pop @aaray1;
}

system "perl scriptforblast.pl $ARGV[0] $ARGV[1]";
#system "mkdir tempBlastResultfile";
#system "perl extractBlastResultfile.pl tempBlastResult tempBlastResultfile";
system "rm tempBlastDatabase";
```

4. Leave one out cross validation on SVM classifier

```
#!/usr/bin/perl -w
$file="aaData.txt";
open(FH, $file);
@array=<FH>;
open(FH2,">out1");
print FH2 " "."\n";
print FH2 "Threshold"."\n";

for($i=0.0;$i<=0.9;$i+=.1)
{ print FH2 $i."\n";}
open(FH3,">out2");
for($i=-0.1;$i>=-0.9;$i-=.1)
{ print FH3 $i."\n";}
system("cat out1 out2 >out");
for($i=0;$i<@array;$i++)
{
 for($i=0;$i<520;$i++)
 {
@newarray=@array;
$test=$newarray[$i];
open(FH,">testset");
splice(@newarray,$i,1);
open(FH1, ">trainset");
print FH1 @newarray;
system("/home/student/Desktop/jaai/svm/svm_learn -t 2 -c 100 -g 400 trainset
amino_model");
print FH $test;
```

```
system("/home/student/Desktop/jaai/svm/svm_classify testset amino_model $i.out");
}
}
system("cat *.out >output_amino");
system "rm *.out";
for($i=0.0;$i<=0.9;$i+=.1)
{
system("perl /home/student/Desktop/jaai/svm/threshold_check.pl output_amino $i >predict.$i");
system("perl /home/student/Desktop/jaai/svm/sensitivity.pl >aminopos_output");
}
system "rm predict.*";

for($i=-0.9;$i<=-0.1;$i+=.1)
{
system("perl /home/student/Desktop/jaai/svm/threshold_check.pl output_amino $i >predict.$i");
system("perl /home/student/Desktop/jaai/svm/sensitivity1.pl >aminoneg_output");
}
system("cat aminopos_output aminoneg_output >amino_out");
system("paste -d '\t' out  amino_out >amino_out_1");
system "rm predict.*  aminoneg_output testset trainset amino_model amino_out out out1 out2 aminopos_output";
```