INVESTIGATING THE RECOGNITION AND INTERACTIONS OF NON-POLAR α HELICES IN BIOLOGY

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Life Sciences

2016

James Alexander Baker

Contents

Word count 8496

The University of Manchester

James Alexander Baker
Doctor of Philosophy
Investigating the Recognition and Interactions of Non-Polar α Helices in Biology
December 15, 2016

Non–polar helices figure prominently in structural biology, from the first protein structure (myoglobin) through trans-membrane segments, to current work on recognition of protein trafficking and quality control. Trans-membrane α helix containing proteins make up around a quarter of all proteins, as well as two thirds of drug targets, and contain some of the most critical proteins required for life as we know it. Yet they are fundamentally difficult to study experimentally. This is in part due to the very features that make them so biologically influential: their non–polar trans-membrane helix regions. What is missing in the current literature is a nuanced understanding of the complexities of the helix composition beyond a hydrophobic region of around 20 residues. Currently it is known that the properties of trans-membrane protein α helices underpin membrane protein insertion mechanisms.

By leveraging large datasets of trans-membrane proteins, this thesis is focused on characterizing features of α helices en masse, particularly regarding their topology, membrane–protein interactions, and intra-membrane protein interactions.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's Policy on Presentation of Theses.

Acknowledgements

I would like to thank all members of both the Eisenhaber research group, as well as the the Curtis and Warwicker research group for discussion, but in particular Jim Warwicker, Frank Eisenhaber, Birgit Eisenhaber, and Wing-Cheong Wong for supervision and guidance during my research. I would also like to thank The University of Manchester and the A*STAR Singapore Bioinformatics Institute for funding the project. Furthermore I would like to extend my gratitude to the research group of Professor Stephen High.

List of publications

Journal Articles

Posters

Baker, J. A. and Warwicker, J. A Bioinformatic Method to Identify Potential SNARE Proteins. *40th FEBS Congress* Late Breaker (2015)

List of Abbreviations

Endoplasmic Reticulum (ER)

Molecular Dynamics (MD)

Protein Data Bank (PDB)

Plasma Membrane (PM)

Palmitoyloleoylphosphatidylcholine (POPC)

Soluble N-Ethylmaleimide-Sensitive Factor Attachment Receptor (SNARE)

Signal Peptide (SP)

Tail Anchor (TA)

Trans-membrane (TM)

Trans-membrane Helix (TMH)

Trans-membrane Protein (TMP)

Trans-membrane Segment (TMS)

Chapter 1

Introduction

Trans-membrane (TM) biology is a huge and varied field that is ultimately the study of the interface between compartments of the cell; one of the fundamental pillars of life as we know it. Trans-membrane Protein (TMP)s include some of the most critical to life proteins as well as a large number of drug targets. However, the experimental inaccessibility of the Trans-membrane Helix (TMH) has hampered progress of study compared to their globular structural counterparts. Despite progress over the last decade, the understanding of the relationship between the sequence and function of a TMH is incomplete.

In this chapter we will place the TMH problem in context, then describe the important biological aspects of the TMH (the traversing Trans-membrane Segment (TMS) as well as the membrane itself), and discuss tools and methods that allow us to analyze and describe the nuanced differences between these TMH sequences.

1.1 α Helices; Structure And Function

1.1.1 Trans-membrane Helix Sequence Composition

Measurements of the TMH regions have found that they are roughly 20 residues in length; 17.3±3.1 from 160 TMHs [1], 27.1±5.4 residues based on 129 TMHs [2], 26.4 residues based on 45 TMHs [3], 25.3±6.0 residues based on 702 TMHs [4], 24.6±5.6 from 837 TMHs [5], and 28.6±1.6Å to 33.5 ±3.1Å from 191 proteins depending on membrane types [6]. There are a couple of reasons for this variation. Primarily is that

the boundaries of TMHs are extremely hard to precisely identify since it is unclear exactly how far the TMH rises into the water-interface region [7]. Secondly is that it is emerging that different membranes have different thicknesses [8], and that this is directly reflected in the hydrophobic lengths of the TMH [6, 9].

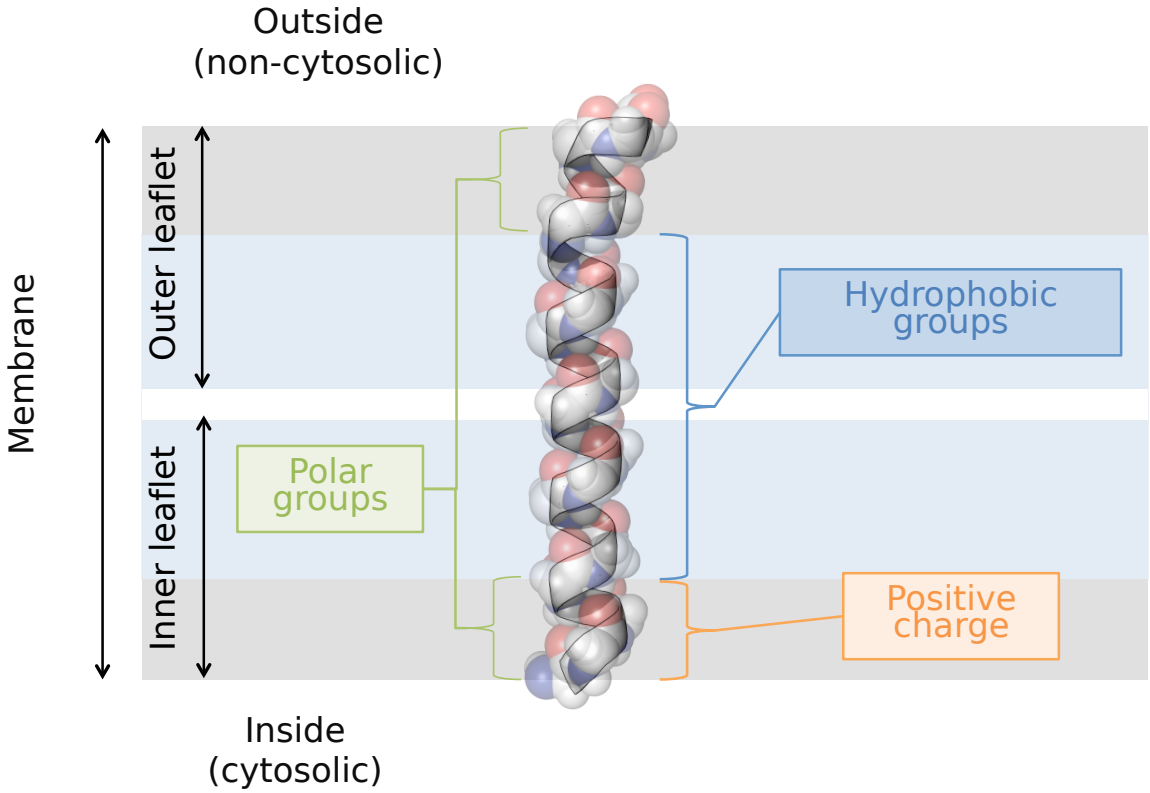

Figure 1.1: A cartoon showing the general components of the membrane and a typical TMH. The example used here for illustrative purposes is trans-membrane region of therein (PDB 2LK9) [10]. Dark gray areas denote the area of lipid head groups. The residues found in these areas are often described as flanking regions, and are often in contact with the aqueous interface of the membrane. The helix core is mostly composed of hydrophobic residues. More recently the hydrophobic group region has been associated with cell localization and a broad range of biochemical functions [11, 12]. Although the regions labeled here generally hold true in terms of the statistical distribution of polar, non-polar, and charged groups, it is by no means absolute laws and many proteins break these "rules" [5, 6, 9].

Sequence properties that can be analyzed by bioinformatics, the sequence complexity and hydrophobicity, of the TMH have been used to predict the role of the TMH as either functional or structural, and as a discrete cluster from other SCOP annotated helices [12]. Those findings demonstrated that the sequence of the TMH holds valuable information regarding biological roles, and forms the basis of our interest in the link between the polarity of a helix and functional activity beyond structural anchorage.

The language used to describe TMHs varies somewhat across the literature, primarily due to a changing understanding of TMH general structure and relevance to function over the last 15 years or so. There is a general composition of a TMH despite specific protein and membrane constraints [9].

A study by Baeza–Delgado *et al.* from 2013 [5] looked at TMHs in 170 integral membrane proteins from a manually maintained database of experimentally confirmed TMPs; MPTopo [13]. The group examined the distribution of residues along the TMHs. As expected, half of the natural amino acids are equally distributed along Trans-membrane (TM) helices whereas aromatic, polar, and charged amino acids along with proline are biasedly near the flanks of the TM helices [5]. It has been noted that transitions between the polar and non-polar groups at the ends of the hydrophobic core occur in a more defined edge on the cytoplasmic side than at the extra-cytoplasmic face when counting from the middle of the helix outwards [5]. This is probably reflecting the different lipid composition of both leaflets of biological membranes [5]. A larger previous study using 1192 human and 1119 yeast predicted TMHs that were not structurally validated further explored the difference in TMH and leaflet structure by exploiting the evolutionarily conserved sequence differences between the TMH in the inner and outer leaflets [9]. TMHs from vertebrates and invertebrates were found to be reasonably similar compositionally. The differences in consensus TMH structure implies that there are general differences between the membranes of the Golgi and Endoplasmic Reticulum (ER). The abundance of serines in the region following the lumenal end of Golgi TMSs probably reflects the fact that this part of many Golgi enzymes forms a flexible linker that tethers the catalytic domain to the membrane [9].

The "Positive-Inside" Rule

Two publications by von Heijne coined the "Positive-Inside" rule demonstrated the practical value of positively charged residue sequence clustering in topology prediction of TMHs in bacteria [14, 15]. It was clearly defined and shown that positively charged residues more commonly were found on the "inside" of the cytoplasm rather than the periplasm of *E. coli*. More recently still large scale sequence analysis of TMHs from different organelle membrane surfaces in eukaryotic proteomes, show the clustering of positive charge being cytosolic [5, 6, 9].

The Aromatic Belt

Tyrosine and tryptophan residues commonly are found at the interface boundaries of the TMH and this feature is called the "aromatic belt" [5, 9, 16–18]. Not all aromatic residues are not found in the aromatic belt; phenylalanine has no particular preference for this region [17, 19]. However it still remains unclear if this is to do with anchorage or translocon recognition [5].

Snorkeling

Broadly speaking, TMHs are non-polar. However some contain polar and charged residues in the helix itself. Whilst this might seem thermodynamically unstable at first glance, a molecular dynamic feature called the "snorkel" effect explains in part how this is possible [20, 21]. Simply put, the snorkeling effect involves the long flexible side chain of leucine reaching the water interface region to interact with the polar head-groups of the bilayer even when the α helix backbone is pulled into the hydrophobic layer [22]. This has also been suggested to allow helices to adapt to varying thicknesses of the membrane [23].

1.1.2 Hydrophobicity of Trans-membrane Segments

Over the last 50 years or so, there have been many attempts to use hydrophobicity scales of residues to predict structural classifications of proteins. Due to the vast amounts of scales, major efforts have been made to compare them to identify which ones are better for which tasks of identifying structural elements [24, 25]. Simm *et al.* 2016 [24] compared 98 scales and found that the accuracy of a scale for secondary structure prediction depends on the spacing of the hydrophobicity values of certain amino acids but generally that the methods behind the scales don't affect the separation capacity between β sheets or α helices.

Throughout this thesis several scales are used to evaluate and estimate hydrophobic values of peptide chains. All the scales aim for quantifying the hydrophobic values of each residue. There are several key differences in their methodology, assumptions, and aims. Ultimately, all the scales are attempting to allow estimation of ΔG_{whf}; the free energy of a folded helix (f) from the water (w) into the membrane core (h). This free

energy measurement is regarded as being currently experimentally inaccessible [26].

Kyte & Doolittle Hydropathy Scale

A scale based on the water–vapor transfer free energy and the interior-exterior distribution of individual amino acids [27].

Hessa's Biological Hydrophobicity Scale

This is arguably the most biologically relevant scale [25], and is often called the ΔG^{aa}_{app} scale. The scale is based on an experimental method where the free energy exchange during recognition of designed polypeptide TMH by the ER Sec61 translocon occurred [16]. These measurements were then used to calculate a biological hydrophobicity scale. The original study reported positional variance in some residues and is strictly valid only for residues in the core of the TMH. A more refined study quantified the positional dependencies of each amino acid type [28].

White and Wimley Octanol – Interface Whole Residue Scale

This scale is calculated from two other scales; the octanol scale, and the interface scale [29]. This scale is fundamentally based on the partitioning of host-guest pentapeptides (acetyl-WL-X-LL-OH) and another set of peptides (AcWLm) between water and octanol, as well as water to Palmitoyloleoylphosphatidylcholine (POPC).

The Eisenberg Hydrophobic Moment Consensus Scale

The Eisenberg scale is a consensus scale based on the earlier scales from Tanford [30], Wolfenden [31], Chothia [32], Janin [33], and the von Heijne scale [34]. The scales are normalized according to serine [35]. The automatic TRANSMEM annotation currently used in Uniprot is according to TMHMM [36], Memsat [37], Phobius [38] and the hydrophobic moment plot method of Eisenberg and coworkers [35].

A Brief Comparison of the Hydrophobicity Scales.

Even at an overview, the methodologies seem strikingly varied. Crucially this results in slightly different scores for some residues. To get an idea of how the methodological

differences of the scales translate to numerical variations, we can normalize the scales so that each residue type is represented as a fraction of the maximum residue type value. If we assume each scale is in principle a linear scale, to calculate the normalised value for a given scale (x_r) we must look at all values within that scale as a set (a).

$$x_r = \frac{a_r + |\min_a|}{|\max_a - \min_a|} \tag{1.1}$$

Where x is the standardized value, r specifies the residue type, and a is the unnormalized hydrophobicity score. This allows us to say, as a fraction, how far between the minimum and maximum value is the value in question.

In the case of the Hessa and White and Wimley scale, the numbers should be inverted. This is because they count the most hydrophobic residues (isoleucine and leucine) as negative values, and more polar residues ascend into the positive numbers. Since the normalisation results in a scale between 0-1, to generate the inverted values we can simply use:

$$x_r = 1 - \frac{a_r + |\min_a|}{|\max_a - \min_a|} \tag{1.2}$$

x_r was compared from different scales to assess the variability of each of the the residue types, r (Figure 1.2).

Although as a trend the scales agree, because of the methodological differences, there are indeed variations of values even after normalization. Due to these discrepancies it is preferable and typical amongst the literature to use several scales to verify the observable trends resulting from an individual scale. Notably, one of the classic scales, Kyte & Doolittle Hydropathy Scale shows striking similarity to the modern Hessa's ΔG^{aa}_{app} scale, and that generally the "better" scales count proline as hydrophilic, and focus on helix recognition rather than amino acid analogues [25]. In α helices from soluble proteins, proline is almost always a helix breaker, and α helix prediction scales don't even attempt to quantify a proline scoring penalty.

1.1.3 Sequence Complexity

Sequence complexity, otherwise referred to as sequence entropy, is essentially an estimate of the linguistic entropy of a string. In the context of biology can be thought of

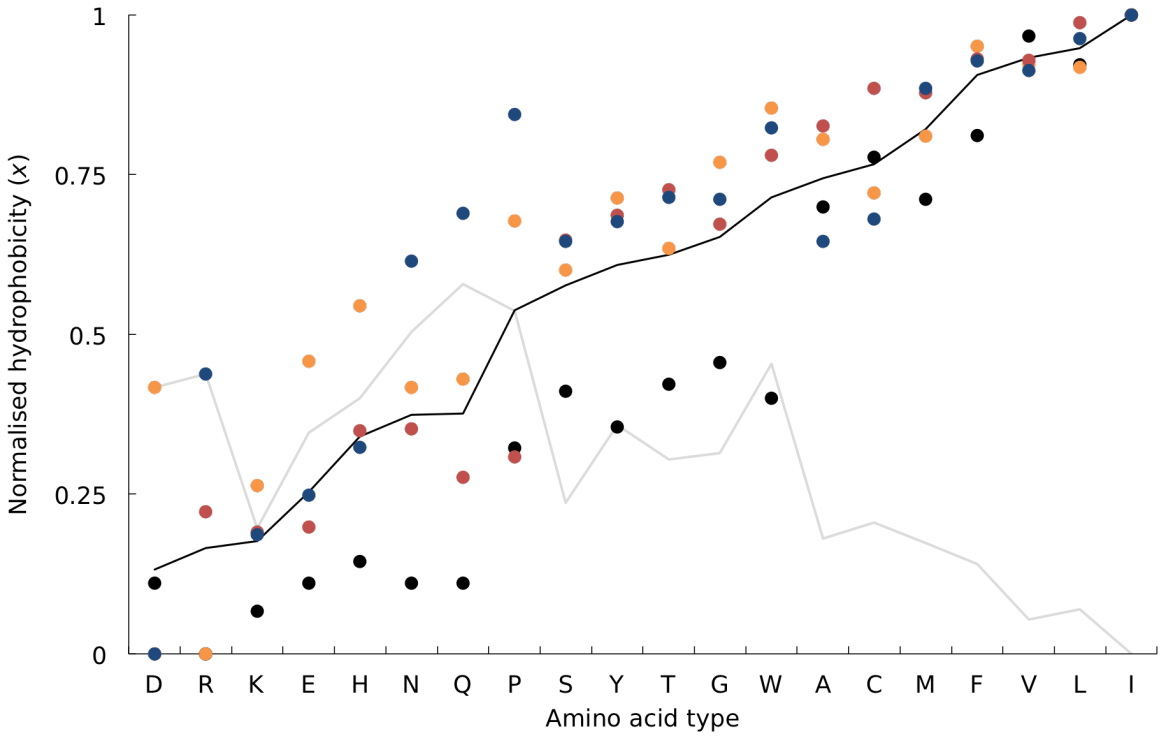

Figure 1.2: A scatter plot comparing normalized values (x) of different hydrophobicity scales. Amino acids on the horizontal axis are arranged according to $\bar{x}_r$ taking into account all 4 scales. The points in blue are the White and Wimley scale [29], in red is Hessa's biological scale [16], in orange is the Eisenberg consensus scale [35], and in black are the Kyte & Doolittle values [27]. The black line connects $\bar{x}_r$ values, whilst the gray line connects $|\max_{x_r} - \min_{x_r}|$, although obviously in both cases the line is only for illustrative purposes and does not show a relationship between the residues.

as an estimation of the non-randomness of a sequence. Sequence complexity can be used to analyze DNA sequences [39–41], however here we will focus on the analysis of the complexity of a sequence in protein sequences.

The compositional complexity is measure over sequence windows. If we have an amino acid composition $\{n_i\}_i = \min i, ..., \max i$ with a window length of $L = \Sigma n_i$,the total number of sequences can be calculated by dividing a factorial of the length by the product of the compositions, i.e $N = L!/\Pi n_i$ possible sequences. The SEG algorithm [42, 43] identifies subsegments of the raw region which have the lowest probability. The algorithm searches for and concatenates sub-threshold segments for the Shannon entropy-like term:

$$K_2 = -\Sigma \frac{n_i}{L} \log \frac{n_i}{L} \tag{1.3}$$

The lowest probability subsegment can be defined as $K_1 = \log N/L$. By altering the window lengths, and the thresholds SEG can be optimized to search for subtle compositional deviations, such as coil-coiled regions.

1.2 α Helices In Membranes

1.2.1 The Trans-membrane Protein Problem

Because of the experimental hindrance, TMP biology has been relatively slow to emerge. Throughout the 1990s the concept of a TMH was simple and fairly assured: they were greasy peptides of around 30Å in length, often bundled together and oriented perpendicularly to the membrane. By 2006, crystallography had elucidated more than 60 high resolution structures. Although the classic TMH structures were broadly prevalent, these structures contained a plethora of unusual TMHs. TMSs are capable of partial spanning of the membrane, spanning using oblique angles, and even lying flat on the membrane surface [7, 44]; the classical model was incomplete. Even recently, there is a contingency in the membrane biology field that despite progress over the last decade there is still a lack of information regarding the relationship between TMH sequences and function, TMH structure, intra-membrane TMP assembly, and the behavior of TMHs in the lipid bilayer; the native biological environment of TMHs [45].

Furthermore, the insertion and formation of the unusually orientated TMHs and of the more traditional TMHs have been shown to be underpinned by complex thermodynamic equilibriums and electrostatic interactions [26, 46, 47]. As well as being a biophysically convoluted system, TMHs are biologically functional beyond anchorage in many cases. TMSs have been identified as regulators of protein quality control and trafficking mechanisms, shifting the idea away from TMHs broadly exclusively functioning as anchors [48], and crucially this function beyond anchorage can be revealed by sensitive, careful analysis of the sequence information alone [12].

When predicting the function of any protein, one follows the dicta that function is facilitated by form, and form is determined by the sequence; the more similar the sequences, the more likely that the function is similar. For globular soluble proteins having the same folds induces strict biochemical restrictions on the packing of a

hydrophobic protein core which requires similarity of non-polar residue patterns. Sequence analysis of non-globular TMPs has not been studied to nearly the same extent yet homology paradigms are silently extended and applied to them. In the case of Signal Peptide (SP)s or TMSs the physical constraints are similar for all TMPs, and so matching is indeed merely a reflection of the physical environment of the bilayer, not the common ancestry. Worryingly, because of this oversight it appears that between 2.1% and 13.6% of Pfam hits for SPs or TMSs are indeed false positive results [49].

Over the last decade, Nanodiscs have been routinely used to much more easily obtain crystal structures. Nanodiscs overcome some of the major challenges caused by the hydrophobic helices, and a more faithful representation of the biological membranes than alternative model membranes like liposomes [50].

However, critical questions remain: How is the TMH oriented in the membrane, how is the TMH interacting with the membrane, how is the TMH interacting with another TMH in the membrane, does the TMH have functions beyond anchorage and if so what are they?

1.2.2 The Importance Of Trans-membrane Proteins

Membrane bound proteins underpin almost every biological process directly, or indirectly, from photosynthesis to respiration. Integral TMP are encoded by around 30% of the genes in the human genome which reflects their biological importance [51]. These proteins allow biochemical pathways that traverse the various biological membranes used in life.

The relationship between the membrane and TMPs is underpinned by complex thermodynamic and electrostatic equilibria. Once inserted the protein doesn't leave the membrane as a result of the TMH being very hydrophobic. This hydrophobicity, and the hydrophobicity of the lipid tails means that they self associate and this association is entropically driven by water. Another way of describing it is that they fiercely dissociate from the water. The overall ΔG for a TMH in the membrane is -12kcalmol^{-1} [26]; the association of the helix in the membrane is typically spontaneous.

1.3 Biological Membrane Composition

1.3.1 Lipids of the Membrane

The compartmentalization of cellular biochemistry is arguably one of the most significant events to have occurred in evolution, and is certainly one of the fundamental prerequisites for life [52]. The proteins that allow life to use this biochemical barrier are perhaps equally important. Together, the lipid bilayer and proteins therein allow complex biochemical systems that facilitate life as we know it.

It is critical to understand that the lipid bilayer and the trans-membrane α helices are inextricably linked, and often what we observe from the α helices reflect the properties of the much harder to study membranes. The lipid membranes influence the local structure, dynamics, and activity of proteins in the membrane in non-trivial ways [53–60], as well as protein folding [61].

There is a rich variety of lipid molecules that make up the biological membranes. The majority of lipids in higher organism membranes are phospholipids, sphingolipids, and sterols. These are composed of a glycerol molecule. Bonded to the glycerol molecule are two hydrophobic fatty acid tail groups, and a negatively-charged polar phosphate group. The polar phosphate group is modified with an alcohol group. Water entropically drives the self association of the lipid molecules. In other words the bilayer forms from these phospholipid molecules due to the fierce dissociation between the polar water and the hydrophobic tails. Furthermore the bilayer maximises van der Waals interactions between the closely-packed hydrocarbon chains, which contributes to the stability of the bilayer. This can be seen even in relatively early Molecular Dynamics (MD) simulations [62].

Differenes in Membrane Compositions

It has been known for some time that the outer membranes of Gram negative bacteria are asymmetric in terms of lipid composition. The outer membranes contain lipopolysccharide, whilst the inner is a mixture of approximately 25 phospholipid types. Adding to the membrane asymmetry composition story, a thorough analysis of residue composition in yeast and human TMH regions revealed intra-membrane

leaflet composition asymmetry in the ER, but not the Golgi [9]. Furthermore protein-lipid interactions have been shown to be determinants of membrane curvature [59], and undertake complex orientations and conformations to allow for hydrophobic mismatch [63].

1.3.2 Membrane Potential

Simply put, membrane potential is the voltage across a membrane. If the membrane is permeable to a certain type of ion, then the ion will experience an electrical pulling force during the diffusion process that pull toward the "preferred" biological location. This clearly depends on a chemical component involving both the charge and ion concentration gradient.There are various ways of estimating the membrane potential *ab initio*.

The Nernst equation can be derived directly from the simplified thermodynamic principles (i) the Boltzmann distribution, and (ii) a field charge interaction energy [64]. It is defined as:

$$E_m = \frac{RT}{F} \times \ln \frac{c_{out}}{c_{in}} \tag{1.4}$$

Where charge Em is the membrane potential, z is the ion charge, c is the concentration of an ion in that cell environment.

One problem in a biological membranes is that the compartments always involve multiple ion channels. The Goldman equation aims to solve this problem by accounting for several ions that contribute to c_{out} and c_{in} (such as K^+, Na^+, and Cl^-) simultaneously:

$$E_m = \frac{RT}{F} \times \ln \left(\frac{p_{K^+}.[K^+]_{out} + p_{Na^+}.[Na^+]_{out} + p_{Cl^-}.[Cl^-]_{in}}{p_{K^+}.[K^+]_{in} + p_{Na^+}.[Na^+]_{in} + p_{Cl^-}.[Cl^-]_{out}} \right) \tag{1.5}$$

Where charge Em is the membrane potential, z is the ion charge, $[i]$ is the ion concentration and p_i is the relative membrane permeability for the actual ion.

However, it is rife with caveats caused by the assumptions of the simplified model. Such assumptions include ions having point charge, that the potential is constant throughout the solution. This is compounded because it assumes the constant potential is the same as the point of measurement which can be heavily influenced by,

for example, a specific adsorbtion of either part of the redox pair or the competitive adsorption of a supporting ion in solution [64]. Therefore one should be cautious to understand the limitations and variability when extrapolating experimentally determined E_0, particularly when using such an idealized model in a biological context.

Organelle Membrane Potentials

Several studies have attempted to quantify the various voltages across the intracellular membranes. Negativity was found in the ER, with a voltage between between 75mV to 95mV in the ER membrane [65, 66]. Negativity was found in the mitochondrial matrix with a voltage across the mitochondrial membrane at 150mV [67]. No notable membrane potential has been identified in the Golgi [68, 69].

1.4 Biogenesis of Trans-membrane Proteins

1.4.1 Translocation

Tail-Anchored Proteins Post Translationally Insert

Tail anchored proteins are a topologically distinct class of intracellular proteins defined by their single carboxy-terminal trans-membrane domain with a cytosolic facing amino-terminus. Tail anchored proteins are involved in a range of key cellular functions including protein translocation and apoptosis. Additionally, within the tail anchored class of proteins are a set of vesicle fusion proteins called Soluble N-Ethylmaleimide-Sensitive Factor Attachment Receptor (SNARE) proteins. There is biomedical interest in SNARE drug delivery mechanisms. SNAREs can fuse liposomes containing various drug payloads into the membrane.

Notably, known SNARE TMHs are highly hydrophobic even compared to other tail anchored TMHs.

1.4.2 Translocon Independent Membrane Insertion

Signal anchored proteins, proteins that contain a single hydrophobic segment that serves as both a mitochondrial targeting signal and a membrane anchor, as well as

tail anchored proteins have been shown to be able to spontaneously insert into the membrane independently from the translocon [46, 70, 71].

It is postulated that there are electrostatic factors in the flanking regions that contribute to this spontaneous membrane insertion. Our experimental collaborators in Stephen Highs group are interested in a small group of tail anchored proteins that have very polar trans-membrane domains and are capable of liposome membrane insertion without insertion machinery, also known as spontaneous insertion. They have found that chimeric synaptobrevin, one of the first identified SNARE proteins, is capable of spontaneous insertion if the tail anchor domain is replaced by the TM domains belonging to a protein of known spontaneously inserting domains. Their studies have moved the focus of spontaneous insertion away from the loop regions and onto the biophysicochemical factors of the TMH itself. The idea that SNARE proteins are modular, and capable of spontaneous insertion has significant implications for both biomedical application in liposome based drug delivery and can aid future research for testing complex biological molecular networks [72, 73].

1.5 Aims of This Thesis.

1. Negative not inside rule.
2. SNARE and Tail Anchor (TA) project.
3. Good and bad helices.

Chapter 2

The "Negative-Outside" Rule

The description of a TMH remains incomplete. The understanding of TMP topology is erroneous, and despite a wealth of structures, the general model of helix-helix and helix-lipid interactions remains speculative and requires a great deal of intensive analysis to generate a working model of a particular TMP.

The work presented in this chapter is an expanded version of published work [**Baker2017**]. We use advanced statistical analysis to analyze large sequence datasets that have rich topological annotation. By analyzing these sequences in the context of anchorage, we find that some TMHs are confined to biological constraints of the membrane, whereas others that likely contain function beyond anchorage, are less conforming to the membrane. Specifically, there is further elaboration of statistical definitions in the methods than in the published paper.

2.1 Abstract

2.2 Summary

As the idea of positive residues inside the cytoplasm emerged during the late 1980s, so did the idea of negative residues working in concert with TMH orientation. It was shown that removing a single lysine residue reversed the topology of a model *Escherichia coli* protein, whereas much higher numbers of negatively charged residues are needed to reverse topology [74]. One would also expect to see a skew in negatively charged distribution if a cooperation between oppositely charged residues orientated

a TMH, however there is no conclusive evidence in the literature for an opposing negatively charged skew [5, 6, 9, 17, 18]. However, in *E. coli* negative residues do experience electrical pulling forces when traveling through the SecYEG translocon indicating that negative charges are biologically relevant [47].

2.3 Methods

2.3.1 Normalisation

$$c_r = \frac{(a_{K,r} + a_{R,r}) - (a_{D,r} + a_{E,r})}{N} \tag{2.1}$$

$$p_{i,r} = \frac{a_{i,r}}{\max\limits_{r}(a_r)} \tag{2.2}$$

$$q_{i,r} = \frac{100 a_{i,r}}{a_i} \tag{2.3}$$

2.4 Results

2.4.1 Biophysicochemical differences in multi-pass and single-pass helices

Chapter 3

Tail-Anchored Proteins Revisited; An Up-To-Date Dataset And Biochemical Insights Into Spontaneous Insertion

3.1 Abstract

3.2 Introduction

This study aims to identify SNARE proteins in eukaryotic proteomes by filtering through large datasets using automatically predicted TrEMBL consensus, and manually annotated SWISS-PROT transmembrane regions. The pipeline generates a list of singlepass proteins with a transmembrane domain close to the C terminal, that are not splice isoforms. A previous study predicted 411 tail anchor proteins [75].

3.3 Methods

The original list UniProt protein database was queried for records containing "TRANSMEM" annotation on June 15, 2016, totaling 75826 records from swissprot, and 12322000 records from TrEMBL.

Expression

3.3.1 Filtering the Uniprot database

Steps carried out by Kalbfleisch *et al.* published in Traffic 2007 (8: 16871694) [75], were recreated using up to date tools. The nonredundant human dataset of 145,715 proteins from SwissProt and TrEMBL [76]. 2,478 singlepass proteins were programmatically extracted according to the TRANSMEM count from that list. Then TMDs not within 15AA of the C terminal were removed, resulting in 455 proteins. No splice isoforms were detected according to searching for NON_TER annotation. 195 proteins of the 411 predicted proteins from the previous study were successfully mapped using the Uniprot mapping tools [76]. Duplicate IDs from the previously predicted tail anchored protein were removed from the set. The remaining dataset contained XXX proteins.

3.3.2 Calculating Hydrophobicity

3.3.3 Calculating Sequence Complexity

3.4 Results

3.4.1 An Up To Date Tail-Anchor Dataset

3.4.2 Potential Tail-Anchored SNARE Protein Discovery

3.4.3 Biology of Spontaneously Inserting Tail Anchored Proteins

Chapter 4

The Anchors And The Doers

4.1 Abstract

4.2 Introduction

4.3 Methods

4.4 Results

Chapter 5

Conclusions

5.1 Outlook

5.1.1 The hydrophobicity–sequence complexity continuum

We hypothesize that the hydrophobicity–sequence complexity continuum contains nuanced codes for different functions and that such differentiation of sequence and structural properties will allow assignment to these varying functions. Additionally, we suggest probing functional classification of yet uncharacterized membrane proteins by similarities of combinations of complex TM sets to well studied membrane proteins and finding those classes of TM proteins where this principle is most directly applicable.

Bibliography

1. Hildebrand, P. W., Preissner, R & Frommel, C. Structural features of transmembrane helices. *FEBS Lett* **559,** 145–151 (2004).

2. Ulmschneider, M. B. & Sansom, M. S. P. Amino acid distributions in integral membrane protein structures. *Biochimica et Biophysica Acta - Biomembranes* **1512,** 1–14. ISSN: 00052736 (2001).

3. Bowie, J. U. Helix packing in membrane proteins. *Journal of Molecular Biology* **272,** 780–789. ISSN: 0022-2836 (1997).

4. Cuthbertson, J. M., Doyle, D. A. & Sansom, M. S. P. Transmembrane helix prediction: A comparative evaluation and analysis. *Protein Engineering, Design and Selection* **18,** 295–308. ISSN: 17410126 (2005).

5. Baeza-Delgado, C., Marti-Renom, M. A. & Mingarro, I. Structure-based statistical analysis of transmembrane helices. *European Biophysics Journal* **42,** 199–207. ISSN: 01757571 (2013).

6. Pogozheva, I. D., Tristram-Nagle, S., Mosberg, H. I. & Lomize, A. L. Structural adaptations of proteins to different biological membranes. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1828,** 2592–2608. ISSN: 00052736 (2013).

7. Von Heijne, G. Membrane-protein topology. *Nature Reviews Molecular Cell Biology* **7,** 909–918. ISSN: 1471-0072 (2006).

8. Van Meer, G., Voelker, D. R. & Feigenson, G. W. Membrane lipids: where they are and how they behave. *Nature reviews. Molecular cell biology* **9,** 112–124. ISSN: 1471-0072 (2008).

9. Sharpe, H. J., Stevens, T. J. & Munro, S. A Comprehensive Comparison of Transmembrane Domains Reveals Organelle-Specific Properties. *Cell* **142,** 158–169. ISSN: 00928674 (2010).

10. Skasko, M. *et al.* HIV-1 Vpu protein antagonizes innate restriction factor BST-2 via lipid-embedded helix-helix interactions. *Journal of Biological Chemistry* **287,** 58–67. ISSN: 00219258 (2012).

11. Junne, T., Kocik, L. & Spiess, M. The hydrophobic core of the Sec61 translocon defines the hydrophobicity threshold for membrane integration. *Molecular biology of the cell* **21,** 1662–70. ISSN: 1939-4586 (2010).

12. Wong, W.-C., Maurer-Stroh, S., Schneider, G. & Eisenhaber, F. Transmembrane helix: simple or complex. *Nucleic acids research* **40,** W370–5. ISSN: 1362-4962 (2012).

13. Jayasinghe, S, Hristova, K & White, S. H. MPtopo: A database of membrane protein topology. *Protein Science* **10,** 455–458. ISSN: 0961-8368 (2001).

14. Von Heijne, G. Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. en. *Nature* **341,** 456–458. ISSN: 0028-0836 (1989).

15. Elofsson, A & von Heijne, G. Membrane protein structure: prediction versus reality. *Annu Rev Biochem* **76,** 125–140. ISSN: 0066-4154 (2007).

16. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433,** 377–81. ISSN: 1476-4687 (2005).

17. Granseth, E., Von Heijne, G. & Elofsson, A. A study of the membrane-water interface region of membrane proteins. *Journal of Molecular Biology* **346,** 377–385. ISSN: 00222836 (2005).

18. Nilsson, J., Persson, B. & von Heijne, G. Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins* **60,** 606–616. ISSN: 1097-0134 (2005).

19. Braun, P. & Von Heijne, G. The aromatic residues Trp and phe have different effects on the positioning of a transmembrane helix in the microsomal membrane. *Biochemistry* **38,** 9778–9782. ISSN: 00062960 (1999).

20. Chamberlain, A. K., Lee, Y., Kim, S. & Bowie, J. U. Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *Journal of Molecular Biology* **339,** 471–479. ISSN: 00222836 (2004).

21. Strandberg, E. & Killian, J. A. Snorkeling of lysine side chains in transmembrane helices: How easy can it get? *FEBS Letters* **544,** 69–73. ISSN: 00145793 (2003).

22. Krishnakumar, S. S. & London, E. Effect of Sequence Hydrophobicity and Bilayer Width upon the Minimum Length Required for the Formation of Transmembrane Helices in Membranes. *Journal of Molecular Biology* **374,** 671–687. ISSN: 00222836 (2007).

23. Kandasamy, S. K. & Larson, R. G. Molecular Dynamics Simulations of Model Trans-Membrane Peptides in Lipid Bilayers: A Systematic Investigation of Hydrophobic Mismatch. *Biophysical journal* **90,** 2326–2343. ISSN: 00063495 (2006).

24. Simm, S., Einloft, J., Mirus, O. & Schleiff, E. 50Years of Amino Acid Hydrophobicity Scales: Revisiting the Capacity for Peptide Classification. *Biological Research* **49,** 31. ISSN: 0717-6287 (2016).

25. Peters, C. & Elofsson, A. Why is the biological hydrophobicity scale more accurate than earlier experimental hydrophobicity scales? *Proteins: Structure, Function and Bioinformatics* **82,** 2190–2198. ISSN: 10970134 (2014).

26. Cymer, F., Von Heijne, G. & White, S. H. Mechanisms of integral membrane protein insertion and folding. *Journal of Molecular Biology* **427,** 999–1022. ISSN: 10898638 (2015).

27. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157,** 105–132. ISSN: 00222836 (1982).

28. Hessa, T. *et al.* Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **450,** 1026–1030. ISSN: 0028-0836 (2007).

29. White, S. H. & Wimley, W. C. MEMBRANE PROTEIN FOLDING AND STABILITY : Physical Principles. *Annual Reviews of Biophysics and Biomolecular Structure* **28,** 319–365. ISSN: 1056-8700 (1999).

30. Nozaki, Y. & Tanford, C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *Journal of Biological Chemistry* **246,** 2211–2217. ISSN: 00219258 (1971).

31. Rose, G. D. & Wolfenden, R. Hydrogen Bonding, Hydrophobicity, Packing, and Protein Folding. *Annual Review of Biophysics and Biomolecular Structure* **22,** 381–415. ISSN: 1056-8700 (1993).

32. Chothia, C. The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology* **105,** 1–12. ISSN: 00222836 (1976).

33. Janin, J. Surface and inside volumes in globular proteins. *Nature* **277,** 491–492. ISSN: 0028-0836 (1979).

34. Von Heijne, G. & Blomberg, C. Trans-membrane Translocation of Proteins. The Direct Transfer Model. *European Journal of Biochemistry* **97,** 175–181. ISSN: 0014-2956 (1979).

35. Eisenberg, D. Three-dimensional structure of membrane and surface proteins. *Annual review of biochemistry* **53,** 595–623. ISSN: 00664154 (1984).

36. Krogh, A, Larsson, B, von Heijne, G & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of molecular biology* **305,** 567–580. ISSN: 0022-2836 (2001).

37. Jones, D. T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **23,** 538–544. ISSN: 13674803 (2007).

38. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology* **338,** 1027–1036. ISSN: 00222836 (2004).

39. Pinho, A. J., Garcia, S. P., Pratas, D. & Ferreira, P. J. S. G. DNA sequences at a glance. *PLoS ONE* **8** (ed Gibas, C.) e79922. ISSN: 19326203 (2013).

40. Oliver, J. L., Bernaola-Galván, P, Guerrero-García, J & Román-Roldán, R. Entropic profiles of DNA sequences through chaos-game-derived images. *Journal of theoretical biology* **160,** 457–470. ISSN: 0022-5193 (1993).

41. Troyanskaya, O. G., Arbell, O., Koren, Y., Landau, G. M. & Bolshoy, A. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics* **18,** 679–688. ISSN: 1367-4803 (2002).

42. Wooton, J. C. Non-globular doamins in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **18,** 269–285. ISSN: 0097-8485 (1994).

43. Hayete, B. & Bienkowska, J. *Gotrees: predicting go associations from protein domain composition using decision trees. Methods in Enzymology* **1993,** 127–138. ISBN: 9812560467. doi:`full_text`. ¡`http://europepmc.org/abstract/MED/15759620`¿ (Elsevier, 2005).

44. Elofsson, A & von Heijne, G. Membrane protein structure: prediction versus reality. *Annu Rev Biochem* **76,** 125–140. ISSN: 0066-4154 (2007).

45. Ladokhin, A. S. Membrane Protein Folding & Lipid Interactions: Theory & Experiment. *The Journal of Membrane Biology* **248,** 369–370. ISSN: 0022-2631 (2015).

46. Merklinger, E. *et al.* Membrane integration of a mitochondrial signal-anchored protein does not require additional proteinaceous factors. *Biochem. J.* **442,** 381–389. ISSN: 0264-6021 (2012).

47. Ismail, N., Hedman, R., Lindén, M. & von Heijne, G. Charge-driven dynamics of nascent-chain movement through the SecYEG translocon. en. *Nature structural & molecular biology* **22,** 145–9. ISSN: 1545-9985 (2015).

48. Hessa, T. *et al.* Protein targeting and degradation are coupled for elimination of mislocalized proteins. *Nature* **475,** 394–7. ISSN: 1476-4687 (2011).

49. Wong, W. C., Maurer-Stroh, S. & Eisenhaber, F. More Than 1,001 problems with protein domain databases: Transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Computational Biology* **6** (ed Bourne, P. E.) 6. ISSN: 1553734X (2010).

50. Borch, J. & Hamann, T. The nanodisc: A novel tool for membrane protein studies. *Biological Chemistry* **390,** 805–814. ISSN: 14316730 (2009).

51. Almén, M., Nordström, K. J., Fredriksson, R. & Schiöth, H. B. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biology* **7,** 50. ISSN: 1741-7007 (2009).

52. Koshland, D. E. The Seven Pillars of Life. en. *Science* **295,** 2215–2216. ISSN: 00368075 (2002).

53. Bondar, A. N., del Val, C., Freites, J. A., Tobias, D. J. & White, S. H. Dynamics of SecY Translocons with Translocation-Defective Mutations. *Structure* **18,** 847–857. ISSN: 09692126 (2010).

54. Bondar, A. N., del Val, C. & White, S. H. Rhomboid Protease Dynamics and Lipid Interactions. *Structure* **17,** 395–405. ISSN: 09692126 (2009).

55. Jardón-Valadez, E., Bondar, A. N. & Tobias, D. J. Coupling of retinal, protein, and water dynamics in squid rhodopsin. *Biophysical Journal* **99,** 2200–2207. ISSN: 00063495 (2010).

56. Kalvodova, L. *et al.* Lipids as modulators of proteolytic activity of BACE: Involvement of cholesterol, glycosphingolipids, and anionic phospholipids in vitro. *Journal of Biological Chemistry* **280,** 36815–36823. ISSN: 00219258 (2005).

57. Urban, S. & Wolfe, M. S. Reconstitution of intramembrane proteolysis in vitro reveals that pure rhomboid is sufficient for catalysis and specificity. eng. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 1883–8. ISSN: 0027-8424 (2005).

58. White, S. H., Ladokhin, A. S., Jayasinghe, S. & Hristova, K. How Membranes Shape Protein Structure. *Journal of Biological Chemistry* **276,** 32395–32398. ISSN: 00219258 (2001).

59. Jensen, M. & Mouritsen, O. G. Lipids do influence protein function - The hydrophobic matching hypothesis revisited. *Biochimica et Biophysica Acta - Biomembranes* **1666,** 205–226. ISSN: 00052736 (2004).

60. Hénin, J., Salari, R., Murlidaran, S. & Brannigan, G. A predicted binding site for cholesterol on the GABAA receptor. *Biophysical Journal* **106,** 1938–1949. ISSN: 15420086 (2014).

61. Kauko, A. *et al.* Repositioning of transmembrane alpha-helices during membrane protein folding. *Journal of molecular biology* **397,** 190–201. ISSN: 1089-8638 (2010).

62. Goetz, R & Lipowsky, R. Computer simulations of bilayer membranes: Self-assembly and interfacial tension. *Journal Of Chemical Physics* **108,** 7397–7409. ISSN: 00219606 (1998).

63. De Planque, M. R. R. & Killian*, J. A. Proteinlipid interactions studied with designed transmembrane peptides: role of hydrophobic matching and interfacial anchoring (Review). en. *Molecular Membrane Biology* **20,** 271–284. ISSN: 0968-7688 (2003).

64. Feiner, A. & McEvoy, A. The nernst equation. *Journal of chemical education* **71,** 493–494. ISSN: 0021-9584 (1994).

65. Qin, Y., Dittmer, P. J., Park, J. G., Jansen, K. B. & Palmer, A. E. Measuring steady-state and dynamic endoplasmic reticulum and Golgi Zn2+ with genetically encoded sensors. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 7351–6. ISSN: 1091-6490 (2011).

66. Worley, J. F. *et al.* Endoplasmic reticulum calcium store regulates membrane potential in mouse islet ??-cells. *Journal of Biological Chemistry* **269,** 14359–14362. ISSN: 00219258 (1994).

67. Perry, S. W., Norman, J. P., Barbieri, J., Brown, E. B. & Gelbard, H. A. Mitochondrial membrane potential probes and the proton gradient: A practical usage guide. *BioTechniques* **50,** 98–115. ISSN: 07366205 (2011).

68. Schapiro, F. B. & Grinstein, S. Determinants of the pH of the Golgi complex. *Journal of Biological Chemistry* **275,** 21025–21032. ISSN: 00219258 (2000).

69. Llopis, J. *et al.* Measurement of cytosolic, mitochondrial, and Golgi pH in single living cells with green fluorescent proteins. *Proceedings of the National Academy of Sciences* **95,** 6803–6808. ISSN: 0027-8424 (1998).

70. Lan, L, Isenmann, S & Wattenberg, B. W. Targeting and insertion of C-terminally anchored proteins to the mitochondrial outer membrane is specific and saturable

but does not strictly require ATP or molecular chaperones. *The Biochemical journal* **349,** 611–621. ISSN: 02646021 (2000).

71. Colombo, S. F., Longhi, R. & Borgese, N. The role of cytosolic proteins in the insertion of tail-anchored proteins into phospholipid bilayers. *Journal of cell science* **122,** 2383–92. ISSN: 0021-9533 (2009).

72. Allen, T. M. & Cullis, P. R. Liposomal drug delivery systems: From concept to clinical applications. *Advanced Drug Delivery Reviews* **65,** 36–48. ISSN: 0169409X (2013).

73. Nordlund, G., Brzezinski, P. & von Ballmoos, C. SNARE-fusion mediated insertion of membrane proteins into native and artificial membranes. *Nature Communications* **5,** 4303. ISSN: 2041-1723 (2014).

74. Nilsson, I. & von Heijne, G. Fine-tuning the topology of a polytopic membrane protein: Role of positively and negatively charged amino acids. *Cell* **62,** 1135–1141. ISSN: 00928674 (1990).

75. Kalbfleisch, T., Cambon, A. & Wattenberg, B. W. A bioinformatics approach to identifying tail-anchored proteins in the human genome. *Traffic* **8,** 1687–1694. ISSN: 13989219 (2007).

76. Bateman, A. *et al.* UniProt: A hub for protein information. *Nucleic Acids Research* **43,** D204–D212. ISSN: 13624962 (2015).