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 Chemistry

2018

James Alexander Baker
orcid.org/0000-0003-0874-2298

Contents

Word count 22,000

List of Tables

List of Figures

The University of Manchester

James Alexander Baker
Doctor of Philosophy
Investigating the Recognition and Interactions of Non-Polar α Helices in Biology
July 30, 2018

Abstract

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 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 data-sets of trans-membrane proteins, this thesis is focused on characterising features of α helices en masse, particularly regarding their topology, membrane–protein interactions, and intramembrane protein interactions.

In this thesis, I make the argument that there are different classifications of trans-membrane α helices. These have markedly different evolutionary pressures, these different classes interact differently with the membrane, and each class serve the protein differently.

Lay Abstract

The survival of each of our cells relies on a cellular barrier to separate themselves from the surrounding environment. The barrier works by being chemically very different from both the outside environment and to the inside of the cell, which in both cases are mostly water. The membrane is fatty, and because of that, the membrane repels water.

Proteins are the molecular machinery that forms much of the cell structure and shape as well as carrying out many of the cell's routine tasks. Around a third of our genome codes for proteins that are permanently embedded in the membrane, but because these proteins are adapted for a life in the water repelling cell wall, they are very hard to study in laboratories which often rely on methods that hold proteins in water-based solutions.

In this thesis, we focus particularly on the parts of the protein that are embedded in the water repelling cellular skin. Traditionally, these regions are hard to study, because we must first remove them from the cellular wall, which causes problems since the embedded regions also repel water and this often causes them to stick to one another, making them hard to work within a laboratory setting.

We analyse thousands of proteins to further our understanding of electrical charges in the embedded regions and find that negative charge on the outside of the cell has been evolutionarily selected across bacteria, animals, and plants. This is especially true for regions that specifically anchor the protein into the cellular wall. Where the embedded regions have an additional function, for example ferrying something in or out of the cell, the negative charge "bias" can no longer be seen.

This thesis demonstrates the radically different evolutionary story that transmembrane regions have compared to other proteins; the sacrifices they make for their stability in order to maintain their function, and their optimisation through evolutionary timescales to become mould to the membrane as best they can.

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

Acknowledgements

I would like to thank all members of both the Eisenhaber research group, as well as 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.

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 [**Ladokhin2015**]. 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 TMH has hampered the progress of study compared to their globular structural analogues. 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 analyse and describe the nuanced differences between these TMH sequences.

1.1 α Helices; Structure And Function

1.1.1 transmembrane 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 [**Hildebrand2004**], 27.1±5.4 residues based on 129 TMHs [**Ulmschneider2001**], 26.4 residues based on 45 TMHs [**Bowie1997**], 25.3±6.0 residues based on 702 TMHs [**Cuthbertson2005a**], 24.6±5.6 from 837

TMHs [**Baeza-Delgado2013**], and 28.6±1.6Å to 33.5 ±3.1Å from 191 proteins depending on membrane types [**Pogozheva2013**]. 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 [**VonHeijne2006**]. Secondly is that it is emerging that different membranes have different thicknesses [**VanMeer2008**], and that this is directly reflected in the hydrophobic lengths of the TMH [**Sharpe2010**, **Pogozheva2013**].

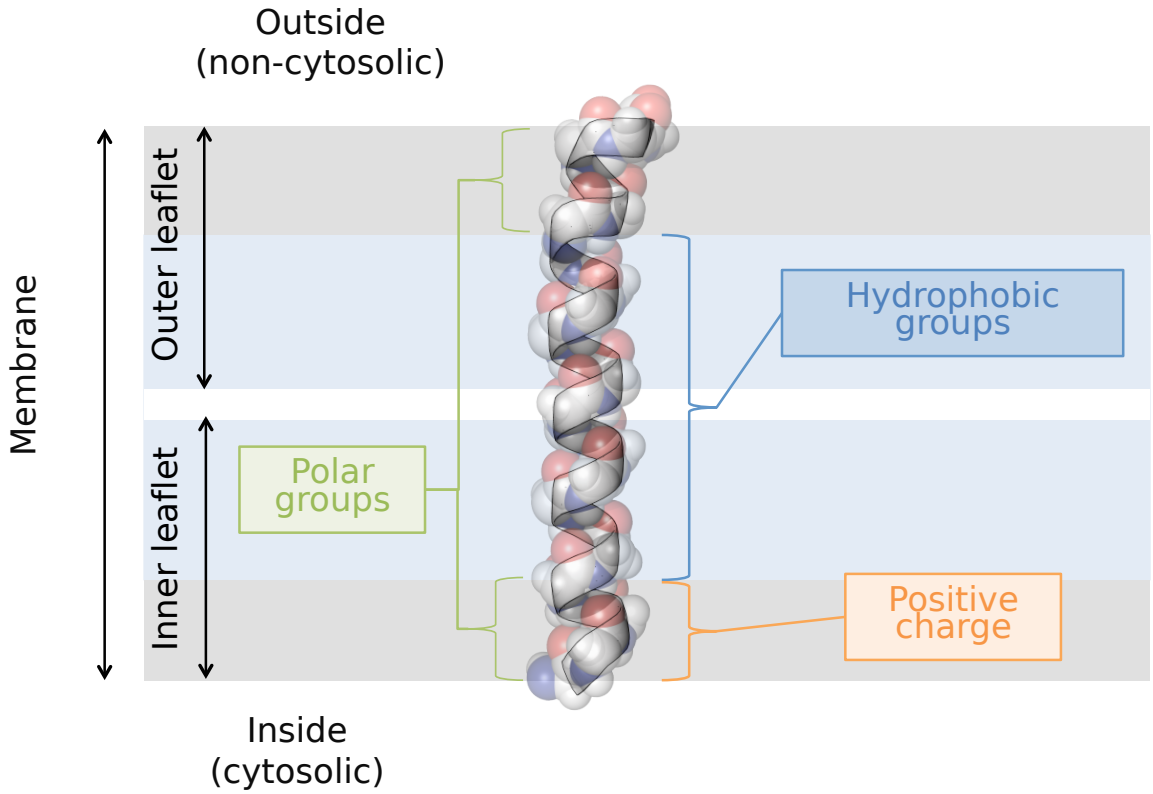

Figure 1.1: A cartoon showing the general components of the membrane and a typical TMH. The example used here for illustrative purposes is the transmembrane region of therein (Protein Data Bank (PDB) 2LK9) [**Skasko2012**]. Dark grey areas denote the area of lipid head groups. The residues found in these 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. Although the regions labelled 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" [**Sharpe2010**, **Baeza-Delgado2013**, **Pogozheva2013**].

From left to right: a typical and traditional TMH, an exceptionally long TMH, a TMH that lies flat in the interface region, a kinked helix that enters and exits the bilayer on the same leaflet, a TMH that is not long enough to span the entire membrane. These exceptional formations present a challenge for topology predictions of the loop regions.

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 [**Sharpe2010**].

A study by Baeza–Delgado *et al.* from 2013 [**Baeza-Delgado2013**] looked at TMHs in 170 integral membrane proteins from a manually maintained database of experimentally confirmed TMPs; MPTopo [**Jayasinghe2001**]. The group examined the distribution of residues along the TMHs. As expected, half of the natural amino acids are equally distributed along transmembrane (TM) helices whereas aromatic, polar, and charged amino acids along with proline are biasedly near the flanks of the TM helices [**Baeza-Delgado2013**]. 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 extracytoplasmic face when counting from the middle of the helix outwards [**Baeza-Delgado2013**]. This is probably reflecting the different lipid composition of both leaflets of biological membranes [**Baeza-Delgado2013**].

A previous study by Sharpe *et al.* from 2010 used 1192 human and 1119 yeast predicted TMHs that were not structurally validated to further explore the difference in TMH and leaflet structure by exploiting the evolutionary conserved sequence differences between the TMH in the inner and outer leaflets [**Sharpe2010**]. 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 [**Sharpe2010**].

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 [**VonHeijne1989**, **Andersson1992**]. 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 [**Sharpe2010**, **Baeza-Delgado2013**, **Pogozheva2013**].

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" [**Hessa2005**, **Granseth2005**, **Sharpe2010**, **Baeza-Delgado2013**, **Nilsson2005a**]. Not all aromatic residues are not found in the aromatic belt; phenylalanine has no particular preference for this region [**Granseth2005**, **Braun1999**]. However, it still remains unclear if this is to do with anchorage or translocon recognition [**Baeza-Delgado2013**].

A study of conserved tryptophan residues during folding of integrin αIIβ3 TM complex demonstrated the anchoring effects of tryptophan (0.4 kcal/mol contribution to membrane stability) in TMHs is greater than the other residues [**Situ2018**]. It was suggested that it's wide amphiphilic range (it's stabilising energetic contribution in either hydrophobic or polar sites) complements the heterogeneity and asymmetry of mammalian membrane lipids in particular.

The Tyrosine side chain is a six-membered aromatic ring with an OH group attached. Tryptophan has two aromatic rings that are fused into one large hydrophobic ring-structure. Phenylalanine, although aromatic, is completely hydrophobic, and is found in the transmembrane part rather than the interfacial parts of MPs. The classical explanation for the preference of Tyrosine and Tryptophan to reside in the interfacial regions is their dipolar character. The side chain must simply seekmpromise. This can be achieved by burying the aromatic ring close to, or within, the hydrophobic core, while the hydrophilic part can interact with the polar lipid head-groups at the interface. Other factors such as the aromaticity, size, rigidity and shape of Tryptophan, rather than its dipolar character, has also been suggested as the primary reasons for its interfacial preference.

Snorkelling

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 [**Chamberlain2004**, **Strandberg2003**]. Simply put, the snorkelling 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 [**Krishnakumar2007**]. This has also been suggested to allow helices to adapt to varying thicknesses of the membrane [**Kandasamy2006**]. More recently it was found that although in simulations the energetic cost of arginine at the centre of the TMH is large, in vivo experimentation with the Sec61 translocon reveals a much smaller penalty [**Ulmschneider2017**]. That same study also found that in simulations, snorkelling, bilayer deformation, and peptide tilting combined to to be sufficient to lower the thermodynamic stability penalty of arginine insertion so that hydrophobic TMHs with a central arginine residue will readily insert into the membrane.

1.1.2 Hydrophobicity of transmembrane Segments

Perhaps the most prevalent and important feature of the transmembrane regions is the membrane spanning region which is composed mostly of non-polar residues. More recently the hydrophobic group region has been associated with cell localisation and a broad range of biochemical functions [**Junne2010**, **Wong2012**].

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 [**Simm2016**, **Peters2014**]. Simm *et al.* 2016 [**Simm2016**] 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 [**Cymer2015**].

Although as a trend most of the scales agree, because of the methodological differences, there are indeed variations of values even after normalisation. Due to these discrepancies, it is preferable and typical amongst the literature to use several scales to verify the observable trends resulting from interpretation from an individual scale. Notably, one of the classic scales, Kyte & Doolittle Hydropathy Scale shows a 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 [**Peters2014**]. 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. Several of the scales used throughout this thesis are outlined below.

Kyte & Doolittle Hydropathy Scale

A scale based on the water–vapour transfer free energy and the interior-exterior distribution of individual amino acids [**Kyte1982**].

Hessa's Biological Hydrophobicity Scale

This is arguably the most biologically relevant scale [**Peters2014**], 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 poly-peptide TMH by the ER Sec61 translocon occurred [**Hessa2005**]. 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 [**Hessa2007**].

White and Wimley Octanol – Interface Whole Residue Scale

This scale is calculated from two other scales; the octanol scale, and the interface scale [**White1999**]. 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 [**Nozaki1971**], Wolfenden [**Rose1993**], Chothia [**Chothia1976**], Janin [**Janin1979**], Wolfenden [**Wolfenden1981**], and the von Heijne scale [**VonHeijne1979**]. The scales are normalised according to serine [**Eisenberg1984**]. The automatic TRANSMEM annotation currently used in UniProt is according to TMHMM [**Krogh2001**], Memsat [**Jones2007**], Phobius [**Kall2004**] and the hydrophobic moment plot method of Eisenberg and coworkers [**Eisenberg1984**].

1.1.3 Sequence Complexity

Sequence properties that can be analysed 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 [**Wong2012**]. 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.

TMSOC's z-score is able to distinguish between functionally active TMHs and those only associated with anchorage [**Wong2012**]. The z-score is a product of both hydrophobicity and a Shannon like sequence entropy of the character string in the TMH. This term is described below in equation 1.1.

$$z(x_\Phi, x_c) = (-1)^s \left[\frac{(x_\Phi - \mu_\Phi)^2}{\sigma_\phi^2} + \frac{(x_c - \mu_c)^2}{\sigma_c^2} \right] \tag{1.1}$$

Where x_c and x_Φ are moving window averages of c, the sequence entropy [**Wootton1996**]. Φ is the White and Wimley hydrophobicity [**White1999**] for a given segment and μ and σ are the mean and standard deviation of the sequence entropy and hydrophobicity of the functional TMH set, that is those TMHs containing active residues.

Sequence entropy, is essentially an estimate of the linguistic entropy of a string. In the context of biology can be thought of as an estimation of the non-randomness of a

sequence. Sequence complexity can be used to analyse DNA sequences [**Pinho2013**, **Oliver1993**, **Troyanskaya2002**], however here we will focus on the analysis of the complexity of a sequence in protein sequences.

Broadly speaking, the information theory entropy of a linguistic string can be defined as in equation 1.2.

$$H(S) = -\sum_{i=1}^{n} p_i \log_s(p_i) \tag{1.2}$$

Where H is the entropy of a sequence (S), and p_i is the probability of a character i through each position (n) in S. This allows us to quantify the average relative information density held within a string of information [**Shannon1948**].

The compositional complexity is measured over sequence windows. If we have an amino acid composition $\{n_i\}_i = \min i, \ldots, \max i$ 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 [**WOOTTON1994269**, **Wootton1996**] identifies sub-segments of the raw region which have the lowest probability. The algorithm searches for and concatenates sub-threshold segments for the Shannon entropy-like term in equation 1.3

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

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

1.2 α Helices In Membranes

1.2.1 The Transmembrane 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 [**VonHeijne2006**, **Elofsson2007**]; 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 behaviour of TMHs in the lipid bilayer; the native biological environment of TMHs [**Ladokhin2015**].

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 equilibrium and electrostatic interactions [**Cymer2015**, **Elisa2012**, **Ismail2015**]. 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 [**Hessa2011**], and crucially this function beyond anchorage can be revealed by sensitive, careful analysis of the sequence information alone [**Wong2012**].

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 [**Wong2010**].

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 [**Borch2009**].

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 Transmembrane Proteins

Membrane-bound proteins underpin almost every biological process directly, or indirectly, from photosynthesis to respiration. Integral TMP are encoded by between a third to a half of the genes in the human genome which reflects their biological importance [**Hopkins2002**, **Almen2009**, **Wang2013**]. 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 equilibrium. 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} [**Cymer2015**]; the association of the helix in the membrane is typically spontaneous.

1.3 Biological Membrane Composition

"before we discuss the membrane proteins, one must consider the biological reason as for why they exist." The outline that MPs are vital for relaying information and chemistry across the membrane.

1.3.1 Lipids of the Membrane

The compartmentalisation 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 [**Koshland2002**]. 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 transmembrane 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 [**Bondar2010**, **Bondar2009**, **Jardon-Valadez2010**, **Kalvodova2005**, **Urban2005**, **White2001a**, **Jensen2004**, **Henin2014**], as well as protein folding [**Kauko2010**].

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 [**Goetz1998**].

Differences 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 lipopolysaccharide, 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 [**Sharpe2010**]. Furthermore, protein-lipid interactions have been shown to be determinants of membrane curvature [**Jensen2004**], and undertake complex orientations and conformations to allow for hydrophobic mismatch [**Planque2003**].

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 pulls 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 [**Feiner1994**]. 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 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:

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

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 adsorption of either part of the redox pair or the competitive adsorption of a supporting ion in solution [**Feiner1994**]. Therefore one should be cautious to understand the limitations and variability when extrapolating experimentally determined E_0, particularly when using such an idealised 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 75mV to 95mV in the ER membrane [**Qin2011**, **Worley1994**]. Negativity was found in the mitochondrial matrix with a voltage across the mitochondrial membrane at 150mV [**Perry2011**]. No notable membrane potential has been identified in the Golgi [**Schapiro2000**, **Llopis1998**].

1.4 Biogenesis of transmembrane Proteins

1.4.1 Translocation Overview

There are, broadly speaking, 3 types of translocation; BiP-mediated eukaryotic post-translational translocation, bacterial post-translational insertion using the Tat system for folded proteins and the Sec system for unfolded proteins, and co-translational insertion in bacteria through the Holotranslocon (HTL) protein complex or its individual components.

1.4.2 Co-translational Translocation

1.4.3 Post-Translational Translocation

1.5 Aims of This Thesis.

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 analyse large sequence datasets that have rich topological annotation. By analysing 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 *erichia coli* protein, whereas much higher numbers of negatively charged residues are needed to reverse topology [**Nilsson1990**]. 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 [**Granseth2005**, **Nilsson2005a**, **Sharpe2010**, **Baeza-Delgado2013**, **Pogozheva2013**]. However, in *E. coli* negative residues do experience electrical pulling forces when travelling through the SecYEG translocon indicating that negative charges are biologically relevant [**Ismail2015**]. In this chapter, we explore the literature surrounding charged residue distribution in the TMH, and demonstrate that the "negative-outside" skew exists in anchoring TMHs

2.3 Introduction

Two decades ago, the classic concept of a TMH was a rather simple story: Typical TMPs were thought to be anchored in the membrane by membrane-spanning bundles of non-polar α–helices of roughly 20 residues length, with a consistent orientation of being perpendicular to the membrane surface. Although this is broadly true, hundreds of high quality membrane structures have elucidated that membrane-embedded helices can adopt a plethora of lengths and orientations within the membrane. They are capable of just partial spanning of the membrane, spanning using oblique angles, and even lying flat on the membrane surface [**Elofsson2007**, **VonHeijne2006**]. The insertion and formation of the TMHs follow a complex thermodynamic equilibrium [**Moon2013**, **MacCallum2011**, **Cymer2015**]. From the biological function point of view, many TMHs have multiple roles besides being just hydrophobic anchors; for example, certain TMHs have been identified as regulators of protein quality control and trafficking mechanisms [**Hessa2011**]. As these additional biological functions are mirrored in the TMHs sequence patterns, TMHs can be classified as simple (just hydrophobic anchors) and complex sequence segments [**Wong2010**, **Wong2011**, **Wong2012**].

The relationship between sequence patterns in and in the vicinity of TMHs and their structural and functional properties, as well as their interaction with the lipid bilayer membrane, has been a field of intensive research in the last three decades [**Ladokhin2015**]. Besides the span of generally hydrophobic residues in

the TMH, there are other trends in the sequence such as with a saddle-like distribution of polar residues (depressed incidence of charged residues in the TMH itself), an enriched occurrence of positively charged residues in the cytosolic flanking regions as well as an increased likelihood of tryptophan and Tyrosine at either flank edge [**Sharpe2010**, **VonHeijne1986**, **VonHeijne1988**, **VonHeijne1989**, **Baeza-Delgado2013**, **Granseth2005**]. Such properties vary somewhat in length and intensity between various biological organelle membranes, between prokaryotes and eukaryotes [**Ojemalm2013**] and even among eukaryotic species studied due to slightly different membrane constraints [**Sharpe2010**, **Pogozheva2013**]. These biological dispositions are exploitable in terms of TM region prediction in query protein sequences [**Beuming2004**, **Zhao2006**] and tools such as the quite reliable TMHMM [**Krogh2001**, **Sonnhammer1998**], Phobius [**Kall2004**, **Kall2007**] or DAS-TMfilter represent todays prediction limit of TMHs hydrophobic cores within the protein sequence [**Cserzo2002**, **Cserzo2004**, **Kall2002**]. The prediction accuracy for true positives and negatives is reported to be close to 100% and the remaining main cause of false positive prediction are hydrophobic α–helices completely buried in the hydrophobic core of proteins. To note, reliable prediction of TMHs and protein topology is a strong restriction for protein function of even otherwise noncharacterised proteins [**Eisenhaber2016**, **Eisenhaber2012**, **Sherman2015**] and thus, very valuable information.

The "positiveinside rule" reported by von Heijne [**VonHeijne2006**, **VonHeijne1989**] postulates the preferential occurrence of positively charged residues (lysine and arginine) at the cytoplasmic edge of TMHs. The practical value of positively charged residue sequence clustering in topology prediction of TMH was first shown for the plasmalemma in bacteria [**VonHeijne1989**, **Sipos1993**]. As a trend, the "positive-inside rule" has since been confirmed with statistical observations for most membrane proteins and biological membrane types [**Baeza-Delgado2013**, **Gavel1991**, **Nilsson2005a**, **Wallin1998**]. However, more recent evidence suggests that, in thylakoid membranes, the "positive-inside rule" is less applicable due to the co-occurrence of aspartic acid and glutamic acid residues together with positively charged residues [**Pogozheva2013**].

The positive-inside rule also received support from protein engineering experiments that revealed conclusive evidence for positive charges as a topological determinant [**VonHeijne1989**, **Beltzer1991**, **Kida2006**, **Nilsson1990**]. Mutational experiments demonstrated that charged residues, when inserted into the centre of the helix, had a large effect on insertion capabilities of the TMH via the translocon. Insertion becomes more unfavourable when the charge was placed closer to the TMH core [**Hessa2005**].

It remains unclear exactly why and how exactly the positive charge determines topology from a biophysical perspective. Positively charged residues are suggested to be stronger determinants of topology than negatively charged residues due to a dampening of the translocation potential of negatively charged residues. This dampening factor is the result of protein-lipid interactions with net zero charged phospholipid, phosphatidylethanolamine and other neutral lipids. This effect favours cytoplasmic retention of positively charged residues [**Bogdanov2014**].

The recent accumulation of TMP sequences and structures allowed revisiting the problem of charged residue distribution in TMHs (see also `http://blanco.biomol.uci.edu/mpstruc/`). For example, whilst β–sheets contain charged residues in the TM region, -helices generally do not (38). Large-scale sequence analysis of TMH from various organelle membrane surfaces in eukaryotic proteomes confirm the clustering of positive charge having a statistical bias for the cytosolic side of the membrane. At the same time, there are many TMH exception examples to the positive-inside rule; however as a trend, topology can be determined by simply looking for the most positive loop region between helices [**Sharpe2010**, **Baeza-Delgado2013**].

When the observation of positively charged residues preferentially localised at the cytoplasmic edge of TMHs emerged, it was also asked whether negatively charged residues work in concert with TMH orientation. It was shown that a single additional lysine residue can reverse the topology of a model *Escherichia coli* protein, whereas a much higher number of negatively charged residues is needed to achieve the same [**Nilsson1990**]; nevertheless, a sufficiently large negative charge can overturn the positive-inside rule [**Andersson1993**, **Kim1994**] and, thus indeed, negative residues are topologically active to a point. Negatively charged residues were observed

in the flanks of TMHs [**Baeza-Delgado2013**], especially of marginally hydrophobic TM regions [**Delgado-Partin1998**]. It is known that the negatively charged acidic residues in TM regions have a non-trivial role in the biological context. In *E. coli*, negative residues experience electrical pulling forces when travelling through the SecYEG translocon indicating that negative charges are biologically relevant during the electrostatic interactions of insertion [**Ismail2012**, **Ismail2015**].

Unfortunately, there is a problem with statistical evidence for preferential negative charge occurrence next to TMH regions. Early investigations indicated overall both positive and negative charge were influential topology factors, dubbed the charge balance rule. If true, one would also expect to see a skew in the negative charge distribution if a cooperation between oppositely charged residues orientated a TMH [**Sipos1993**, **Hartmann1989**]. It might be expected that, if positive residues force the loop or tail to stay inside, negative residues would be drawn outside and topology would be determined not unlike electrophoresis. Yet, there is plenty of individual protein examples but no conclusive statistical evidence in the current literature for a negatively charged skew [**Sharpe2010**, **Baeza-Delgado2013**, **Granseth2005**, **Pogozheva2013**, **Nilsson2005a**, **Andersson1992**].

There are many observations described in the literature that charged residues determine topology more predictably in single-pass proteins than in multipass TMH [**Kim1994**, **Harley1998**]. It is thought that the charges only determine the initial orientation of the TMH in the biological membrane; yet, the ultimate orientation must be determined together with the totality of subsequent downstream regions [**Sato1998**].

With sequence-based hydrophobicity and volume analysis and consensus sequence studies, Sharpe *et al.* [**Sharpe2010**] demonstrated that there is asymmetry in the intramembranous space of some membranes. Crucially, this asymmetry differs among the membrane of various organelles. They conclude that there are general differences between the lipid composition and organisation in membranes of the Golgi and ER. Functional aspects are also important. For example, the abundance of serines in the region following the lumenal end of Golgi TMHs appears to reflect the fact that this part of many Golgi enzymes forms a flexible linker that tethers the catalytic domain to the membrane [**Sharpe2010**].

A study by Baeza-Delgado *et al.* [**Baeza-Delgado2013**] analysed the distribution of amino acid residue types in TMHs in 170 integral membrane proteins from a manually maintained database of experimentally confirmed TMPs (MP-Topo [**Jayasinghe2001**]) as well as in 930 structures from the PDB. As expected, half of the natural amino acids are equally distributed along TMH whereas aromatic, polar and charged amino acids along with proline are biased near the flanks of the TM helices. Unsurprisingly, leucine and other non-polar residues are far more abundant than the charged residues in the TM region [**Sharpe2010**].

In this work, we revisit the issue of statistical evidence for the preferential distribution of negatively charged (and a few other) residues within and nearby TMHs. We rely on the improved availability of comprehensive and large sequence and structure datasets for TM proteins. We also show that several methodical aspects have hindered previous studies [**Sharpe2010**, **Baeza-Delgado2013**, **Pogozheva2013**] to see the consistent non-trivial skew for negatively charged residues disfavouring the cytosolic interfacial region and/or preferring the outside flank. First, we show that acidic residues are especially rare within and in the close sequence environment of TMHs, even when compared to positively charged lysine and arginine. Second, therefore, the manner of normalisation is critical: Taken together with the difficulty to properly align TMHs relative to their boundaries, column-wise frequency calculations relative to all amino acid types as in previous studies will blur possible preferential localisations of negative charges in the sequence. However, the outcome changes when we ask where a negative charge occurs in the sequence relative to the total amount of negative charges in the respective sequence region. Thus, by accounting for the rarity of acidic residues with sensitive normalisation, the "non-negative inside rule/negative-outside rule" is clearly supported by the statistical data. We find that minor changes in the flank definitions such as taking the TMH boundaries from the database or by generating flanks by centrally aligning TMHs and applying some standardised TMH length does not have a noticeable influence on the charge bias detected.

Third, there are significant differences in the distribution of amino acid residues between single-pass and multi-pass TM regions in both the intra-membrane helix and the flanking regions with further variations introduced by taxa and by the organelles along the secretory pathway. Importantly, we find that it is critical to weigh down the

effect of TMHs in multi-pass TMPs with no or super-short flanks to observe statistical significance for the charge bias. To say it bluntly, if there are no flanks of sufficient length, there is also no negative charge bias to be observed.

The charge bias effect is even clearer when a classification of TMHs into so-called simple (which, as a trend, are mostly single-pass and mere anchors) and so-called complex (which typically have functions beyond anchorage) is considered [**Wong2010**, **Wong2011**, **Wong2012**]. We also observe parallel skews with regard to leucine, tyrosine, tryptophan and cysteine distributions. With these large-scale datasets and a sensitive normalisation approach, new sequence features are revealed that provide spatial insight into TMH membrane anchoring, recognition, helix-lipid, and helix-helix interactions.

2.4 Results

2.4.1 Acidic residues within and nearby TMH segments are rare

In order to reliably compare the amino acid sequence properties of TMHs, we assembled datasets of TMH proteins from what are likely to be the best in terms of quality and comprehensiveness of annotation in eukaryotic and prokaryotic representative genomes, as well as composite datasets to represent larger taxonomic groups and with regard to sub-cellular locations (see Table 2.1). In total, 3292 single-pass TMH segments and 29898 multi-pass TMH segments were extracted from various UniProt [**TheUniProtConsortium2014**] text files according to TRANSMEM annotation (download dated 20–03–2016). The UniProt datasets used only included manually curated records; however, it is still necessary to check for systematic bias due to the prediction methods used by UniProt for TMH annotation in the majority of cases without direct experimental evidence. Therefore, a fully experimentally verified dataset was also generated for comparison. The representative 1544 single-pass and 15563 TMHs were extracted from the manually curated experimentally verified TOPDB [**Dobson2015**] database (download dated 21–03–2016) referred to as

ExpAll here (Table1). TMH organelle residency is defined according to UniProt annotation. To ensure reliability, organelles were only analysed from a representative redundancy-reduced protein dataset of the most well-studied genome: Homo sapiens (referred to as UniHuman herein). The several datasets from UniProt are subdivided into different human organelles (UniPM, UniER, UniGolgi) and taxonomical groups (UniHuman, UniCress, UniBacilli, UniEcoli, UniArch, UniFungi) as described in Table 2.1 (see also Methods section). As will be shown below, these various datasets allow us to validate our findings for a variety of conditions, namely with regard (i) to experimental verification of TMHs, (ii) to origin from various species and taxonomic groups, (iii) to the number of TMHs in the same protein as well as (iv) to subcellular localisation. Data-sets and programs used in this work can be downloaded from `http://mendel.bii.a-star.edu.sg/SEQUENCES/NNI/`.

Table 2.1: Acidic residues are rarer in TMHs of single-pass proteins than in TMHs of multi-pass proteins The statistical results when comparing the number of acidic residues in single-pass or multi-pass TMHs within their database-defined limits and excluding any flanks. The number of helices per dataset can be found in Table 2.2 for single-pass TMHs and Table 3 for multi-pass helices. μ SP is the average number of the respective residues per helix in TMHs from single-pass proteins, while μ MP is the average number of the respective residues per TMH from multi-pass proteins. The Kruskal-Wallis test scores (H statistics) were calculated for the numbers of aspartic acid and glutamic acid residues in each helix from single-pass and the number of aspartic acid and glutamic acid residues in each helix from multi-pass TMHs

Data-set	Acidic residues (D and E)			Aspartic acid (D only)			Glutamic acid (E only)		
	μ SP	μ MP	H statistic P value	μ SP	μ MP	H statistic P value	μ SP	μ MP	H statistic P value
ExpAll	0.086	0.309	148.1 4.50E-34	0.045	0.157	40.3 2.13E-10	0.042	0.161	46.6 8.64E-12
UniHuman	0.076	0.398	316.5 8.31E-71	0.034	0.191	91.6 1.05E-21	0.042	0.207	100.3 1.33E-23
UniER	0.106	0.43	34.4 4.39E-9	0.061	0.161	8.0 4.72E-3	0.045	0.268	26.8 2.24E-7
UniGolgi	0.097	0.381	39.8 2.88E-10	0.043	0.18	19.4 1.05E-5	0.053	0.201	20.2 7.01E-6
UniPM	0.039	0.4	121.0 3.86E-28	0.016	0.187	32.7 1.06E-8	0.022	0.213	36.9 1.26E-9
UniCress	0.062	0.434	163.5 1.99E-37	0.036	0.198	32.5 1.20E-8	0.025	0.241	66.0 4.59E-16
UniFungi	0.177	0.349	43.1 5.14E-11	0.044	0.166	24.5 7.60E-7	0.133	0.183	4.6 0.033
UniBacilli	0.089	0.352	24.1 9.16E-7	0.048	0.185	11.2 8.27E-4	0.04	0.176	12.3 4.54E-5
UniEcoli	0.148	0.315	2.7 0.100	0.111	0.15	0.1 0.729	0.037	0.163	2.2 0.140
UniArch	0.438	0.606	1.8 0.183	0.083	0.344	11.2 8.33E-4	0.354	0.247	3.5 0.0624

The hydrophobic nature of the lipid bilayer membrane implies that, generally, charged residues should be rare within TMHs. For acidic residues, even the location

in the sequence vicinity of TMHs should be disfavoured because of the negatively charged head groups of lipids directed towards the aqueous extracellular side or the cytoplasm. In agreement with the biophysically justified expectations, the statistical data confirms that acidic residues are especially rare in TMHs and their flanking regions. In Figure 1 where we plot the total abundance of all amino acid types in single-pass TMHs and multi-pass TMHs (including their ±5 flanking residues), acidic residues were found to be amongst the rarest amino acids both in UniHuman and ExpAll.

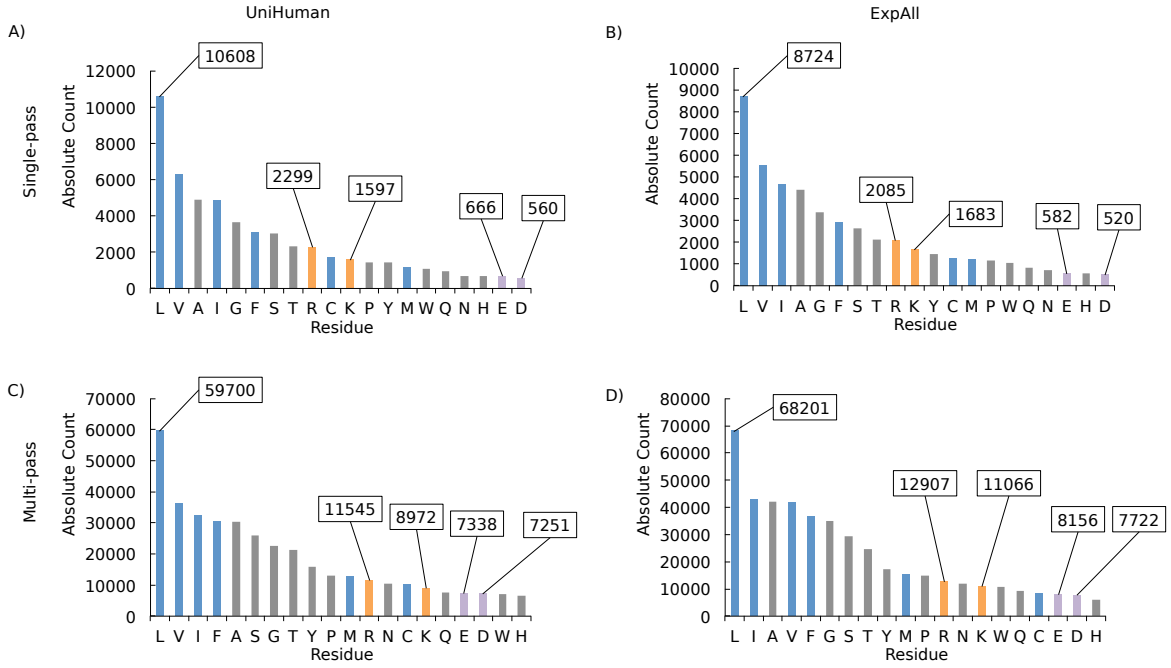

Figure 2.1: Negatively charged amino acids are amongst the rarest residues in TMHs and ±5 flanking residues. Bar charts of the abundance of each amino acid type in the TMHs with flank lengths of the accompanying ±5 residues from the (a) UniHuman single-pass proteins, (b) ExpAll single-pass proteins, (c) UniHuman multi-pass proteins, and (d) ExpAll multi-pass proteins. Amino acid types on the horizontal axis are listed in descending count. The bars were coloured according to categorisations of hydrophobic, neutral and hydrophilic types according to the free energy of insertion biological scale [**Hessa2005**]. Grey represents hydrophilic amino acids that were found to have a positive ΔG app, and blue represents hydrophobic residues with a negative ΔG app, purple denotes negative residues and positive residues are coloured in orange. The abundances of key residues are labelled.

The effect is most pronounced in single-pass TMPs (Figure 2.1). There are only 666 glutamates (just 1.24% of all residues) and 560 aspartates (1.05% respectively) among the total set of 53238 residues comprised in 1705 TMHs and their flanks. Within just the TMH regions, there are 71 glutamates (0.20% of all residues in TMHs and flanks) and 58 aspartates (0.16% respectively). This cannot be an artefact of

UniProt TMH assignments since this feature is repeated in ExpAll. There are only 582 glutamates (1.22%) and 520 aspartates (1.09%) among the 47568 residues involved. Within the TMH itself, there are 64 glutamates (0.19%) and 69 aspartates (0.21%). In both cases, the negatively charged residues represent the ultimate end of the distribution. To note, acidic residues are rare even compared to positively charged residues which are about 3–4 times more frequent. On a much smaller dataset of single-spanning TMP, Nakashima *et al.* [**Nakashima1992**] made similar compositional studies. To compare, they found 0.94% glutamate and 0.94% aspartate within just the TMH region (values very similar to ours from TMHs with small flanks; apparently, they used more outwardly defined TMH boundaries) but the content of each glutamate and aspartate within the extracellular or cytoplasmic domains is larger by an order of magnitude, between 5.26% and 9.34%. These latter values tend to be even higher than the average glutamate and aspartate composition throughout the protein database (5–6% [**Nakashima1992**]).

In the case of multi-pass TMPs (Figure 2.1), glutamates and aspartates are still very rare in TMHs and their ±5 residue flanks (1.94% and 1.92% from the total of 377207 in the case of UniHuman respectively, 1.79% and 1.70% from the total of 454700 in the case of ExpAll). Yet, their occurrence is similar to those of histidine and tryptophan and, notably, acidic residues are only about ∼1.5 times less frequent than positively charged residues. The observation that acidic residues are more suppressed in single-pass TMHs compared with the case of multi-pass TMHs is statistically significant. In Table 2.1, the acidic residues are counted in the helices (excluding flanking regions) belonging to either multi-pass or single-pass helices. Indeed, single-pass helices appear to tolerate negative charge to a far lesser extent than multi-pass helices as the data in the top two rows of Table 2.1 indicates (for datasets UniHuman and ExpAll). The trend is strictly observed throughout sub-cellular localisations (rows 3–5 in Table 2.1) and taxa (rows 6–10). Statistical significance (P0.001) is found in all but six cases. These are UniEcoli (D+E, D, E), UniArch (D+E, E) and UniFungi (E). The problem is, most likely, that the respective datasets are quite small. Notably, the difference between single- and multi-pass TMHs is greatest in UniPM; here, TMHs from multi-pass proteins have on average 0.400 negative residues per helix, whereas single-pass TMHs contained just 0.039 (P=3.86e-28).

2.4.2 Amino acid residue distribution analysis reveals a "negative-not-inside/negative-outside" signal in single-pass TMH segments

The rarity of negatively charged residues is a complicating issue when studying their distribution along the sequence positions of TMHs and their flanks. For UniHuman and ExpAll , we plotted absolute abundance of aspartic acid, glutamic acid, lysine, arginine, and leucine at each position (i.e., it scales as the equivalent fraction in the total composition of the alignment column) (Figure 2.2). To note, the known preference of positively charged residues towards the cytoplasmic side is nevertheless evident. Yet, it becomes apparent that any bias in the occurrence of the much rarer acidic residues is overshadowed by fluctuations in the highly abundant residues such as leucine.

The trends become clearer if the occurrence of specific residues is normalised with the total number of residues of the given amino acid type in the dataset observed in the sequence region studied as shown for UniHuman and for ExpAll in Figure 2.2. For comparison, we indicated background residue occurrences (dashed lines calculated as averages for positions -25 to -30 and 25 to 30). The respective average occurrences in the inside and outside flanks (calculated from an average of the values at positions -20 to -10 and 10 to 20 respectively) are shown with wide lines.

The "positive-inside rule" becomes even more evident in this normalisation: Whereas the occurrence of positively charged residues is about the background level at the outside flank, it is about two to three times higher both for the UniHuman and the ExpAll datasets at the inside flank. To note, the background level was found to be 1.7% (lysine) and 1.6% (arginine) in UniHuman and 1.4% (lysine and arginine) in ExpAll. The inside flank average is 4.3% (lysine) and 4.6% (arginine) in UniHuman and 4.2% (lysine) and 4.6% (arginine) in ExpAll. The outside flank is similar to the background noise levels: about 1.4% (lysine) and 1.5% (arginine) in UniHuman and about 1.5% (lysine) and 1.4% (arginine) in ExpAll.

Most interestingly, a "negativeinside depletion" trend for the negatively charged residues is apparent from the distribution bias. The inside flank averages for glutamic acid were 1.1% and 1.4% in UniHuman and ExpAll respectively; for aspartic acid, 1.2% and 1.4% in UniHuman and ExpAll respectively. Meanwhile, the outside flanks

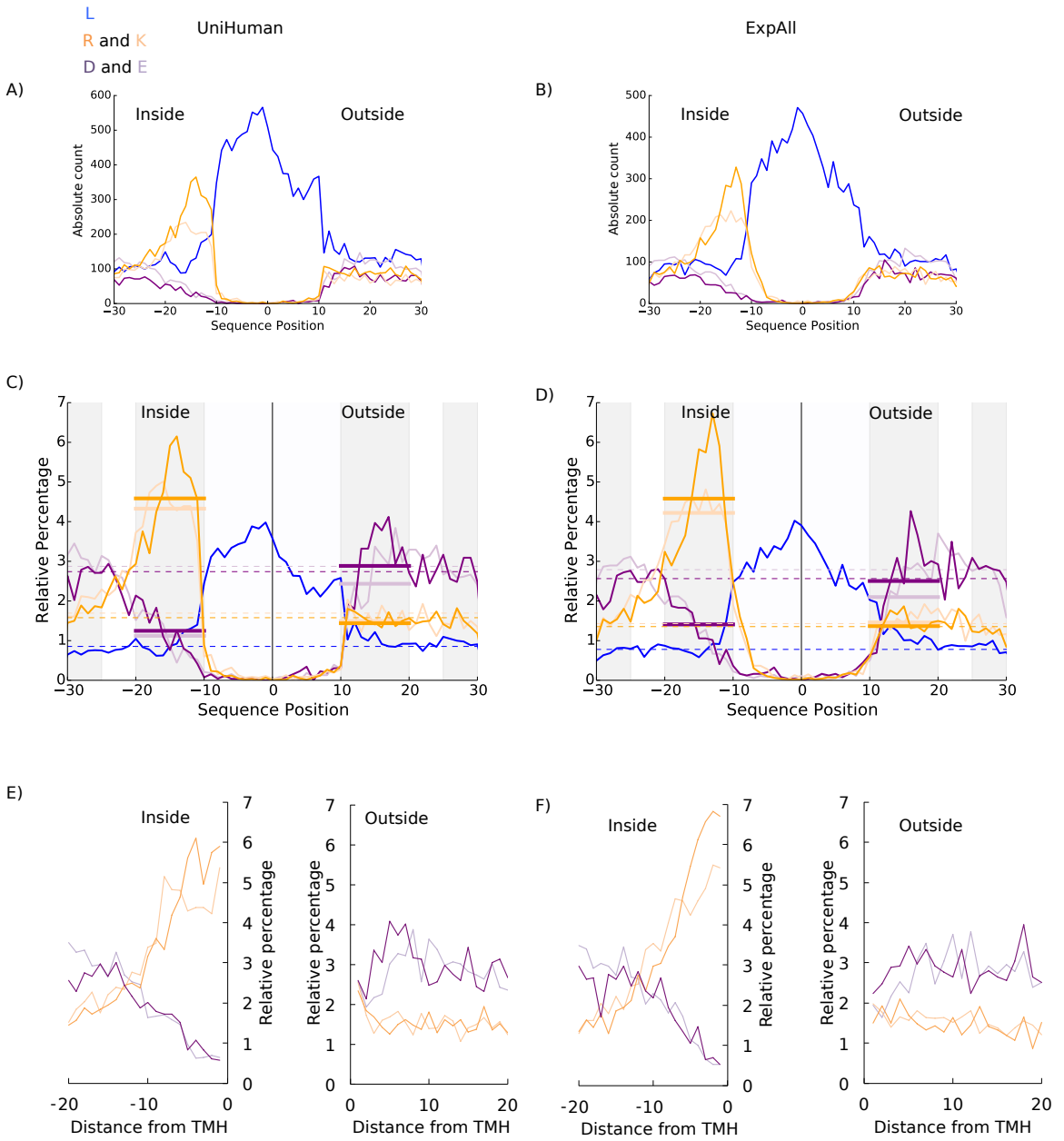

Figure 2.2: Relative percentage normalisation reveals a negative-outside bias in TMHs from single-pass protein datasets. All flank sizes were set at up to ±20 residues. We acknowledge that all values, besides the averaged values, are discrete, and connecting lines are illustrative only. On the horizontal axes (ad) are the distances in residues from the centre of the TMH, with the negative numbers extending towards the cytoplasmic space. For (e) and (f), the horizontal axis represents the residue count from the membrane boundary with negative counts into the cytoplasmic space. Leucine, the most abundant non-polar residue in TMHs, is in blue. Arginine and lysine are shown in dark and light orange respectively. Aspartic and glutamic acid are showing in dark and light purple respectively. (a) and (b) On the vertical axis is the absolute abundance of residues in TMHs from single-pass proteins from (a) UniHuman and (b) ExpAll. Note that no clear trend can be seen in the negative residue distribution compared to the positive-inside signal and the leucine abundance throughout the TMH. c and d On the vertical axis is the relative percentage at each position for TMHs from single-pass proteins from (c) UniHuman and (d) ExpAll. The dashed lines show the estimation of the background level of residues with respect to the colour; an average of the relative percentage values between positions 25 to 30 and 30 to 25. The thick bars show the averages on the inner (positions 20 to 10) and outer (positions 10 to 20) flanks coloured to the respective amino acid type. Note a visible suppression of acidic residues on the inside flank when compared to the outside flank in single-pass proteins when normalising according to the relative percentage. (e) and (f) The relative distribution of flanks defined by the databases with the distance from the TMH boundary on the horizontal axis. The inside and outside flanks are shown in separate subplots. The colouring is the same as in (a) and (b).

for aspartic acid and glutamic acid occurrences were measured at 2.9% and 2.4% respectively in UniHuman and, in ExpAll, these values for aspartic acid and glutamic acid were found to be 2.5% and 2.1% respectively. Against the background level of aspartic acid (2.8% and 2.9% in UniHuman) and glutamic acid (2.6% and 2.9% in ExpAll), the inside flank averages were found to be about 2–3 times lower than the background level while the outside flank averages were comparable to the background level (Figure 2.2). Taken together, this indicates a clear suppression of negatively charged residues at the inside flank of single-pass TMHs and a possible trend for negatively charged residues occurring preferentially at the outside flank. This is not an effect of the flank definition selection since the trend remains the same when using the database-defined flanks without the context of the TMH (Figure 2.2). For UniHuman, the negative charge expectancy on the inside flank doesnt reach above 2% until position -10 (D) and position -11 (E), whereas, on the outside flank, both D and E start >2%. The same can be seen in ExpAll where negative residues reach above 2% only as far from the membrane boundary as at position -9 (D) and position -7 (E) on the inside but exceed 2% beginning with position 1 (D) and 3 (E) on the outside (Figure 2.2).

The observation of negative charge suppression at the inside flank, herein the "negative-inside depletion" rule, is statistically significant throughout most datasets in this study. The inside-outside bias was counted using the Kruskal-Wallis (KW) test comparing the occurrence of acidic residues within 10 residues of each TMH inside and outside the TMH (Table 2.2). We studied both the database-reported flanks as well as those obtained from central alignment of TMHs (see Methods). The null hypothesis (no difference between the two flanks) could be confidently rejected in all cases (P-value<0.001 except for UniBacilli), the sign of the H-statistic (KW) indicating suppression at the inside and/or preference for the outside flank (except for UniArch). Most importantly, acidic residues were found to be distributed with bias in ExpAll (P-value<3.47e-58) and in UniHuman (P-value=1.13e-93). Whereas with UniBacilli, the problem is most likely the dataset size, the exception of UniArch, for which we observe a strong negative inside rule, is more puzzling and indicates biophysical differences of their plasma-membrane.

Table 2.2: Statistical significances for negative charge distribution skew on either side of the membrane in single-pass TMHs The Helices column refers to the total TMHs contained in each dataset (ExpALL, TMHs from TOPDB [**Dobson2015**]; UniHuman, human representative proteome; UniER, human endoplasmic reticulum representative proteome; UniGolgi, human Golgi representative proteome; UniPM, human plasma membrane representative proteome; UniCress, Arabidopsis thaliana (mouse-ear cress) representative proteome; UniFungi, fungal representative proteome; UniBacilli, Bacilli class representative proteome; UniEcoli, Escherichia coli representative proteome; UniArch, Archaea representative proteome; see Methods for details). In the "Database-defined flanks" column, the "Negative residues" column refers to the total number of negative residues found in the ±10 flanking residues on either side of the TMH and does not include residues found in the helix itself. In the "Flanks after central alignment" column, the "Negative residues" column refers to the total number of negative residues found in the 20 to 10 residues and the +10 to +20 residues from the centrally aligned residues of the TMH. Unlike the other tables, the global averages are derived from the ±20 datasets. The KW scores were calculated for negative residues by comparing the number of negatively charged residues that were within the 10 inside residues and the 10 outside residues in either case

Single-pass		**Database-defined flanks**				**Flanks after central alignment**			
Data-set	**Helices**	**Negative residues**		**H statistic**	**P value**	**Negative residues**		**H statistic**	**P value**
		Inside	**Outside**			**Inside**	**Outside**		
ExpAll	1544	848	1648	258.59	3.47E-58	735	1541	262.29	5.44E-59
UniHuman	1705	780	1922	421.53	1.13E-93	652	1865	501.86	3.74E-111
UniER	132	78	156	23.76	1.09E-06	76	150	21.62	3.33E-06
UniGolgi	206	60	240	104.45	1.61E-24	54	239	107.18	4.06E-25
UniPM	493	197	578	177.68	1.56E-40	161	569	215.18	1.02E-48
UniCress	632	314	450	18.23	1.96E-05	231	444	55.8	8.01E-14
UniFungi	729	449	631	28.15	1.12E-07	413	627	38.08	6.79E-10
UniBacilli	124	90	113	3.73	5.35E-02	86	106	2.53	1.12E-01
UniEcoli	54	32	77	17.24	3.30E-05	30	74	14.74	1.24E-04
UniArch	48	113	8	49.66	1.83E-12	96	7	45.62	1.43E-11

2.4.3 Amino acid residue distribution analysis reveals a general negative charge bias signal in outside flank of multi-pass TMH segments — the negative outside enrichment rule

As a result of the rarity of negatively charged residues, any distribution bias is difficult to be recognised in the plot showing the total abundance (or alignment column composition) of residues in multi-pass TMHs and their flanks from UniHuman and ExpAll (Figure 2.3). Yet, as with single-pass helices, the dominant general leucine enrichment, as well as positive inside signal, can be identified with certainty. When the residue occurrence is normalised by the total occurrence of this residue type in the sequence regions studied (shown as a relative percentage of at each position for multi-pass helices from UniHuman and ExpAll in Figure 2.3), the bias in the distribution of any type of charged residues becomes visible.

With regard to the positive-inside preference, positively charged residues have a background value of 2.0% for arginine and 2.2% for lysine in UniHuman, and 1.7% for arginine and 1.9% for lysine in ExpAll. At the inside flank, this rises to 4.6% for arginine and 4.1% for lysine in UniHuman and 4.6% for arginine and 4.2% for lysine in ExpAll. The mean net charge at each position was calculated for multi-pass and single-pass datasets from UniHuman and ExpAll (Figure 2.4). The positive inside rule clearly becomes visible as the net charge has a positive skew approximately between residues -10 and -25. What is noteworthy is that the peaks found for single-pass helices were almost three times greater than those of multi-pass helices. For single-pass TMHs, the peak is +0.30 at position -15 in UniHuman and +0.31 at position -14 in ExpAll, whereas TMHs from multi-pass proteins had lower peaks of +0.15 at position -13 in UniHuman and +0.10 at position -14 in ExpAll. Thus, there is a positive charge bias towards the cytoplasmic side; yet, it is much weaker for multi-pass than for single-pass TMHs.

Notably, a "negative outside enrichment" trend also can be seen from the distribution of the negatively charged residues, though with some effort (Table 3) as the effect is also weaker than in the case of single-pass TMHs. We studied the flanks under four conditions: (i) database-defined flanks without overlap between neighbouring TMHs,

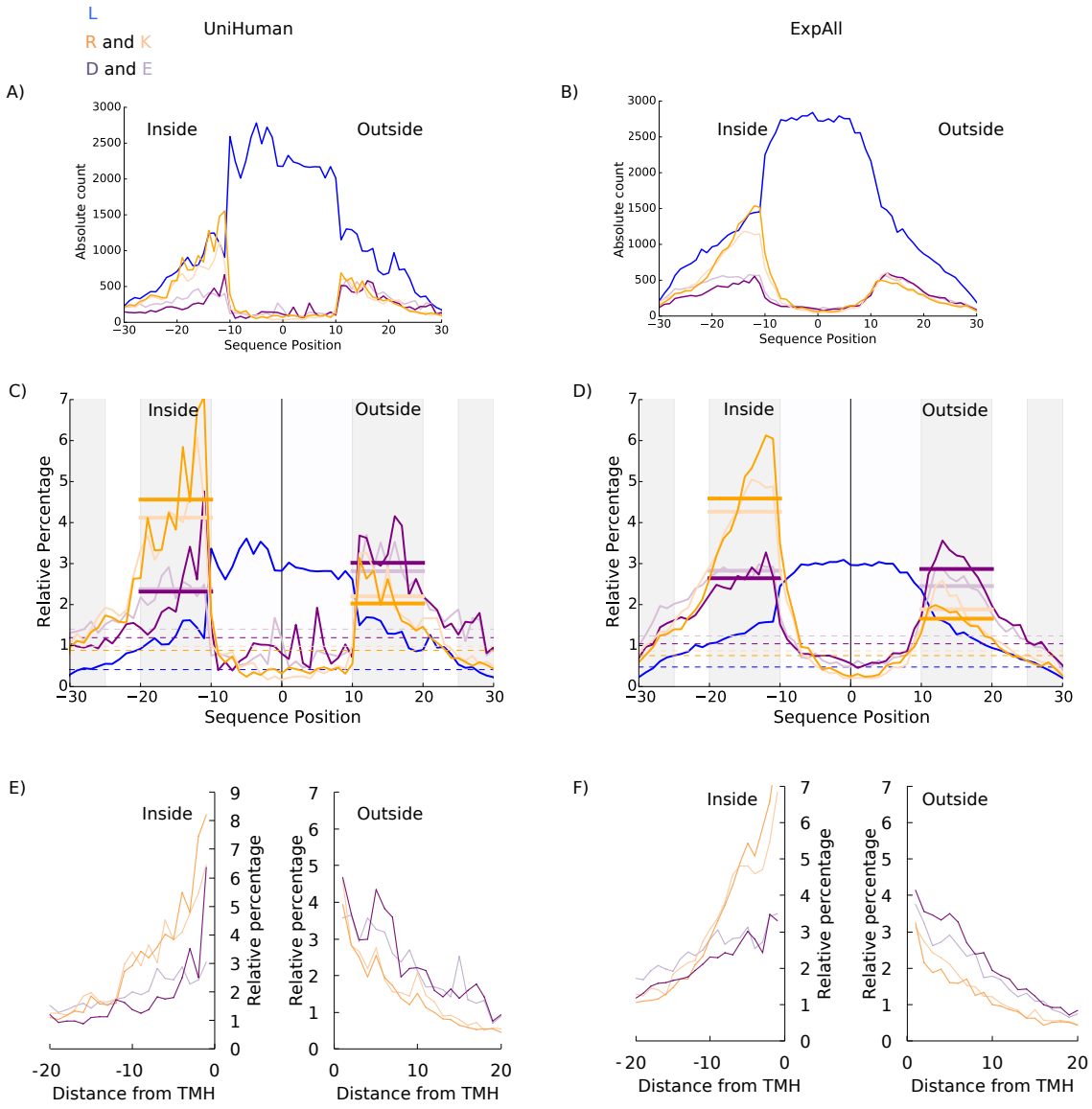

Figure 2.3: Negative-outside bias is very subtle in TMHs from multi-pass proteins. The meaning for the horizontal axis is the same as in Figure 2.2, with the negative sequence position numbers extending towards the cytoplasmic space. Leucine is in blue. Arginine and lysine are shown in dark and light orange respectively. Aspartic and glutamic acid are shown in dark and light purple respectively. All flank sizes were set at up to ±20 residues. (a) and (b) On the vertical axes are the absolute abundances of residues from TMHs of multi-pass proteins from (a) UniHuman and (b) ExpAll. c and d On the vertical axes are the relative percentages at each position for TMHs from multi-pass proteins from (c) UniHuman and (d) ExpAll. As in Figure 2.2(c) and (d), the dashed lines show the estimation of the background level of residues with respect to the colour, and the thick bars show the averages on the inner and outer flanks coloured to the respective amino acid type. e and f The relative distribution of flanks defined by the databases with the distance from the TMH boundary on the horizontal axis for both the inside and outside flanks. The colouring is the same as in (a) and (b).

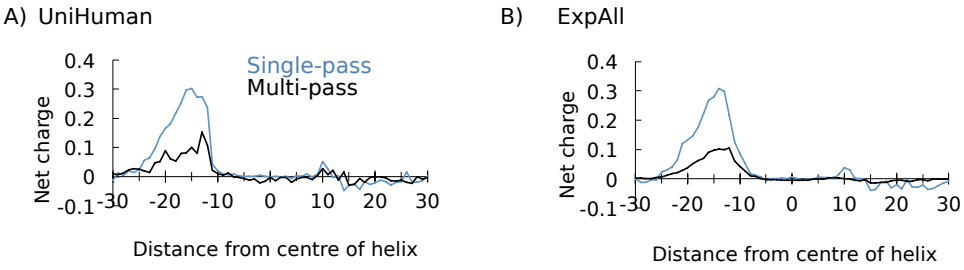

Figure 2.4: The net charge across multi-pass and single-pass TMHs shows a stronger positive inside charge in single-pass TMHs than multi-pass TMHs. The net charge per TMH plotted at each position; the positive-inside rule is stronger in TMHs from single-pass proteins than TMHs from multi-pass proteins. The net charge was calculated at each position as described in the Methods section for the (A) UniHuman and (B) ExpAll datasets. Net charge for TMHs from multi-pass proteins is shown in black, and the profile of TMHs from single-pass proteins is drawn in blue.

(ii) flanks after central alignment of TMHs without flank overlap, (iii) database-defined flanks but allowing overlap of flanks shared among neighbouring TMHs, (iv) same as condition (ii) but only the subset of cases where there is at least half of the required flank length at either side of the TMH. In UniHuman as calculated under condition (i), aspartic acid is lower on the inside flank (2.3%) than on the outside flank (3.0%). Glutamic acid is also lower at the inside flank (2.4%) than the 2.8% on the outside flank (Figure 2.3C). Slight variations in defining the membrane boundary point do not influence the trend (compare figures 2.3C and 2.3E). We find that, in all studied conditions, the UniHuman dataset delivers statistical significances (P-values: (i) 6.10e-34, (ii) 5.43e-41, (iii) 3.00e-57, (iv) 5.60e-41) strongly supporting negative charge bias (inside suppression/outside preference; see Table 2.3).

Surprisingly, the result could not straightforwardly be repeated with the considerably smaller ExpAll. Under condition (i), we find with ExpAll that aspartic acid has a background level of 1.0%, an average of 2.6% on the inside flank, and of 2.9% on the outside flank but glutamic acids background is 1.2% but 2.8% on the inside flank and 2.5% on the outside flank. Statistical tests do not support finding a negative charge bias in conditions (i) and (ii). Apparently, the problem is TMHs having no or almost no flanks at one of the sides. Statistical significance for the negative charge bias is detected as soon as this problem is dealt with either by allowing extension of flanks overlap among neighbouring TMHs as in condition (iii) or by kicking out examples without proper flank lengths from the dataset as in condition (iv). The respective P-values are 2.05e-6 and 9.81e-15 respectively.

Table 2.3: Statistical significances for negative charge distribution skew on either side of the membrane in multi-pass TMHs The "Helices" column refers to the total TMHs contained in each dataset (ExpALL, TMH from TOPDB [**Dobson2015**]; UniHuman, human representative proteome; UniER, human endoplasmic reticulum representative proteome; UniGolgi, human Golgi representative proteome; UniPM, human plasma membrane representative proteome; UniCress, Arabidopsis thaliana (mouse-ear cress) representative proteome, UniFungi, fungal representative proteome; UniBacilli, Bacilli class representative proteome; UniEcoli, Escherichia coli representative proteome; UniArch, Archaea representative proteome; see Methods for details). In (A) the "Database-defined flanks" and in (B) the "Database-defined viable* flanks" and the "Overlapping flanks" columns, the "Negative residues" column refers to the total number of negative residues found in the ±10 flanking residues on either side of the TMH and does not include residues found in the TMH itself. (A) In the "Flanks after central alignment" column, the "Negative residues" column refers to the total number of negative residues found in the 20 to 10 residues and the +10 to +20 residues from the centrally aligned residues with a maximum database defined flank length of 20 residues. The total number of proteins is given in the IDs column. The "Helices" column contains the total number of TMHs in the dataset (n), the average number of TMHs per protein in that population (μ) and the standard deviation of that average (σ). The KW scores were calculated for negative residues by comparing the number of negatively charged residues that were within 10 residues inside and 10 residues outside the TMH.
*Here, "viable" indicates that in each TMH used for both flanks either side of the TMH has a flank length of at least half the maximum allowed flank length, in this case 10 (the viable length is 5)

(A)

Multi-pass					Database-defined flanks				Flanks after central alignment			
Data-set	IDs	Helices			Negative residues		H statistic	P value	Negative residues		H statistic	P value
		n	μ	σ	Inside	Outside			Inside	Outside		
ExpAll	2205	15,563	7.07	3.95	9709	9598	0.04	8.43E-01	9648	9659	0.35	5.56E-01
UniHuman	1789	12,353	6.93	3.2	7196	9164	147.5	6.10E-34	6740	8968	179.77	5.43E-41
UniER	155	898	5.85	3.2	630	584	0.44	5.08E-01	578	576	0.03	8.58E-01
UniGolgi	61	383	6.28	2.97	274	261	0.02	8.75E-01	266	259	0.09	7.65E-01
UniPM	427	3079	7.22	3.3	1945	2499	47.98	4.30E-12	1791	2440	64.42	1.01E-15
UniCress	507	3823	7.55	3.32	2567	2426	0.73	3.93E-01	2398	2433	1.11	2.93E-01
UniFungi	1338	8685	6.5	3.75	5560	5266	5.83	1.57E-02	5140	5214	0	9.62E-01
UniBacilli	140	822	5.94	3.98	470	468	0.07	7.92E-01	450	471	0.92	3.38E-01
UniEcoli	529	3888	7.39	3.76	1990	1902	0.26	6.07E-01	1875	1887	0.18	6.71E-01
UniArch	59	327	5.97	2.73	245	175	7.98	4.72E-03	235	181	7.08	7.81E-03

(B)

Multi-pass	Overlapping flanks				Database-defined viable* flanks				
Data-set	Negative residues		H statistic	P value	N	Negative residues		H statistic	P value
	Inside	Outside				Inside	Outside		
ExpAll	11,969	12,615	22.54	2.05E-06	8808	6082	6916	59.93	9.81E-15
UniHuman	8645	11,181	254.3	3.00E-57	8183	5169	6915	179.71	5.60E-41
UniER	750	763	1.16	2.81E-01	516	398	441	3.16	7.55E-02
UniGolgi	333	369	7.12	7.64E-03	195	162	186	3	8.30E-02
UniPM	2319	3107	99.68	1.79E-23	1977	1343	1960	98.63	3.05E-23
UniCress	3142	3298	9.21	2.41E-03	2110	1626	1741	6.4	1.14E-02
UniFungi	6724	6814	0.46	4.96E-01	4581	3340	3411	0.41	5.22E-01
UniBacilli	585	636	2.65	1.04E-01	382	230	306	12.73	3.61E-04
UniEcoli	2574	2800	17.88	2.35E-05	1596	951	1114	16.57	4.69E-05
UniArch	342	248	14.67	1.28E-04	132	120	104	0.28	5.97E-01

The issues we had with ExpAll raised the question that, maybe, sequence redundancy in the UniHuman set could have played a role. Therefore, we repeated all calculations but with UniRef50 instead of UniRef90 for mapping into sequence clusters (see Methods section for detail). We were surprised to see that harsher sequence redundancy requirements do not affect the outcome of the statistical tests in any major way. For the conditions (i)- (iv), we computed the following P-values: (i) 1.31e-28 (5940 negatively residues inside versus 7492 outside), (ii) 1.38e-36 (5516 versus 7320), (iii) 5.60e-53 (7089 versus 9233) and (iv) 4.18e-41 (4232 versus 5730).

So, the amplifying effect of some subsets in the overall dataset on the statistical test that might be caused by allowing overlapping flanks (condition (iii)) is not the major factor leading to the negative charge skew. Similarly, the trend is also not caused by sequence redundancy. Thus, we have learned that the negative charge bias does also exist in multi-pass TMPs but under the conditions that there are sufficiently long loops between TMHs. Bluntly said: no loops equals to no charge bias. As soon as the loops reach some critical length, there are differences between single-pass and multi-pass TMHs with regard to occurrence and distribution of negative charges and the inside-suppression/outside-enrichment negative charge bias appears. Not only are there more negative charges within the multi-pass TMH itself (in fact, negative charges are almost not tolerated in single-pass TMHs; see Table 2.1), but also, there is a much stronger negative outside skew in the TMHs of single-pass proteins than those of multi-pass proteins.

2.4.4 Further significant sequence differences between single-pass and multi-pass helices: distribution of tryptophan, tyrosine, proline and cysteine

Amino acid residue profiles along the TM segment and its flanks differ between single- and multi-pass TMHs also in other aspects. The relative percentages of all amino acid types (normalisation by the total amount of that residue type in the sequence segment) from single-pass helices of the UniHuman (Figure 2.5A; from 1705 TMHs with flanks having 68571 residues) and ExpAll (Figure 2.5B; from 1544 TMHs with flanks having 60200 residues) were plotted as a heat-map. The amino acid types were listed on the

Y axis according to Kyte & Doolittle hydrophobicity [**Kyte1982**] in descending order.

In accordance with expectations, enrichment for hydrophobic residues in the TMH, for the positively charged residues on the inside flank as well as a distribution the negative distribution bias was found in both datasets. Additionally, the inside interfacial region showed consistent enrichment hotspots for tryptophan (e.g., 7.1% at position -11 in ExpAll, 6.2% at position -10 in UniHuman with flanks after central TMH alignment) and tyrosine (6.4% at -11 in ExpAll, 7.1% at -11 in UniHuman), and some preference can also be seen for the outer interfacial region (*e.g.*, 5.2% at position 11 for tryptophan in ExpAll, and 5.8% at position 10 for tryptophan in UniHuman) albeit the "hot" cluster of the outer flank covers fewer positions than that of the inner flank. Further, there is an apparent bias of cysteine on the inner flank and interfacial region (e.g., 5.5% at position -10 in ExpAll, 5.9% at position -11 in UniHuman), and a depression in the outer interfacial region and flank (up to a minimum of 0.3% in both ExpAll and UniHuman). Proline appears to have a depression signal on the outer flank. Note that, in a similar way to Figures 2.2 and 2.3, the distributions of the flanks derived from centrally aligned TMHs are corroborated by the distributions from the database defined TMH boundary flanks (see outside bands in Figures 2.5A-D).

A similar heatmap was generated for UniHuman multi-pass (Figure 2.5C; from 12353 TMHs with flanks having 452708 residues) TMHs and ExpAll multi-pass (Figure 2.5D; from 15563 TMHs with flanks having 535599 residues). Whereas Figures 2.5A-C appear quite noisy, the plot for ExpAll multi-pass TMHs appears almost Gaussian-like smoothed, thus, indicating the quality of this dataset. Tyrosine and tryptophan in the multi-pass case do not appear as enriched in the interfacial regions of single-pass TMHs from both UniHuman and ExpAll. Prolines are only suppressed in the TMH itself and are not suppressed in the outer flank as in the single-pass case but, indeed, are tolerated if not slightly enriched in the flanks.

2.4.5 Hydrophobicity and leucine distribution in TMHs in single- and multi-pass proteins

Generally, we see in Figure 2.5 that compositional biases appear more extreme in the single-pass case, particularly when it comes to polar and non-polar residues being

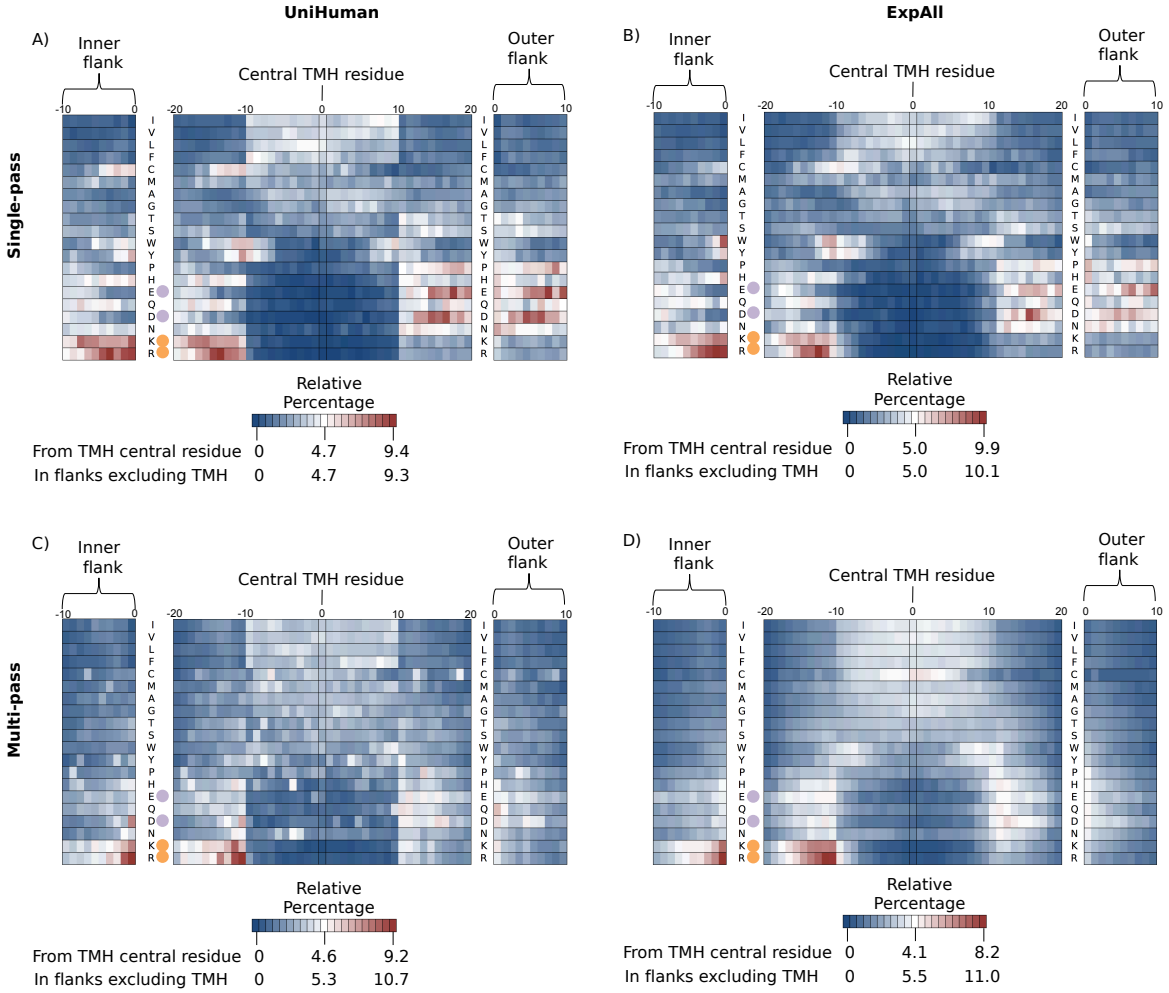

Figure 2.5: Relative percentage heat-maps from predictive and experimental datasets corroborate residue distribution differences between TMHs from single-pass and multi-pass proteins. The residue position aligned to the centre of the TMH is on the horizontal axis, and the residue type is on the vertical axis. Amino acid types are listed in order of decreasing hydrophobicity according to the Kyte and Doolittle scale [52]. The flank lengths in the TMH segments were restricted to up to ±10 residues. The scales for each heat-map are shown beneath the respective subfigure. The darkest blue represents 0% distribution, whilst the darkest red represents the maximum relative percentage distribution that is denoted by the keys in each subfigure, with white being 50% between "cold" and "hot". The central TMH subplots extend from the central TMH residue, whereas the inner and outer flank subplots use the database-defined TMH boundary and extend from that position. a TMHs from the single-pass UniHuman dataset. b Single-pass protein TMHs from the ExpAll dataset. c TMHs from the proteins of the multi-pass UniHuman dataset. d TMHs from ExpAll multi-pass proteins. The general consistency in relative distributions of every residue type between single-pass and multi-pass of either dataset including flank/TMH boundary selection allows us to infer biological conclusions from these distributions that are independent of methodological biases used to gather the sequences. The only residue that displays drastically differently between the datasets is cysteine in multi-pass TMHs only. The most striking differences in distributions between residues from TMHs of single-pass and multi-pass proteins include a more defined Y and W clustering at the flanks, a suppression of E and D on the inside flank, a suppression of P on the inside flank and a topological bias for C favouring the inside flank.

more heavily suppressed and enriched. To investigate this observation, we calculated the hydrophobicity at each sequence-position averaged over all TMHs considered (after having window-averaged over 3 residues for each TMH) using the Kyte & Doolittle hydrophobicity scale [**Kyte1982**] (Figure 2.6A) and validated using White and Wimley octanol-interface whole residue scale [**White1999**], Hessas biological hydrophobicity scale [**Hessa2005**], and the Eisenberg hydrophobic moment consensus scale [**Eisenberg1984**] (Supplementary Figure 2.7). The total set of TMHs was split into 15 sets of membrane-spanning proteins (1 set containing single-pass proteins, 13 sets each containing TMHs from 2-, 3-, 4-. . . 14-TMPs and another of TMHs from proteins with 15 or more TMHs). In Figure 2.6B, we show the P-value at each sequence position by comparing the respective values from multi-pass and single-pass TMHs using the 2-sample t-test (Figure 2.6B). Strikingly, the inside flank of the single-pass TMHs is much more hydrophilic (e.g., see the Kyte & Doolittle score=-1.3 at position -18) than that of multi-pass TMHs (P-value=5.64e-103 at position -14). Most likely, the positive inside rule, along with the interfacial clustering of tryptophan and tyrosine, contribute to a strong polar inside flank in single-pass helices that is not present in multi-pass helices en masse. Further, multi-pass TMHs cluster remarkably closely within the TM core; the respective hydrophobicity is apparently not dependent on the number of TMHs in a given multi-pass TMP. On average, single-pass TMHs are more hydrophobic in the core than multi-pass TMHs (P-value<1.e-72 within positions -55 and P-value=5.92e-190 at position 0). On the other hand, hydrophobicity differences between TMHs from single- and multi-pass proteins fade somewhat at the transition towards the flanks (P-value=1.85e-4 at position -10, and P-value=3.35e-31 at position 10).

Leucine is the most abundant residue in TMHs (Figure 2.1) and is considered one of the most hydrophobic residues by all hydrophobicity scales. Therefore, it plays a very influential role in TMH helix-helix and lipid-helix interactions in the membrane and recognition by the insertion machinery. When looking at the difference in the abundance of leucine between the inner and outer halves, we find that TMHs from single-pass proteins have a trend to contain more leucine residues at the cytoplasmic side of TMHs, particularly in the case of TMHs from single-pass proteins (see Figures 2.2 and 2.5).

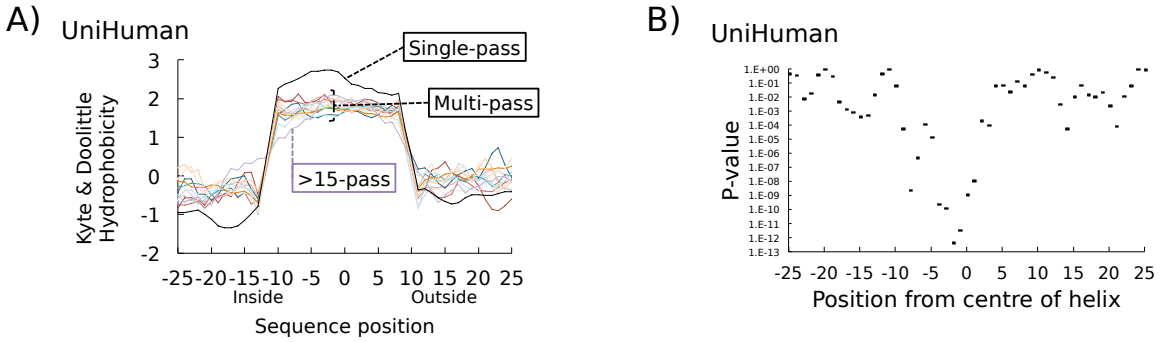

Figure 2.6: There is a difference in the hydrophobic profiles of TMHs from single-pass and multi-pass proteins. a The hydrophobicity of single-pass TMHs compared to multi-pass segments from the UniHuman dataset. The Kyte and Doolittle scale of hydrophobicity [**Kyte1982**] was used with a window length of 3 to compare TMHs from proteins with different numbers of TMHs. This scale is based on the water-vapour transfer of free energy and the interior-exterior distribution of individual amino acids. The same datasets also had different scales applied (Figure 2.7). The vertical axis is the hydrophobicity score, whilst the horizontal axis is the position of the residue relative to the centre of the TMH, with negative values extending into the cytoplasm. In black are the average hydrophobicity values of TMHs belonging to single-pass TMHs, whilst in other colours are the average hydrophobicity values of TMHs belonging to multi-pass proteins containing the same numbers of TMHs per protein. In purple are the TMHs from proteins with more than 15 TMHs per protein that do not share a typical multi-pass profile, perhaps due to their exceptional nature. b The Kruskal-Wallis test (H statistic) was used to compare single-pass windowed hydrophobicity values with the average windowed hydrophobicity value of every TMH from multi-pass proteins at the same position. The vertical axis is the logarithmic scale of the resultant P values. We can much more readily reject the hypothesis that hydrophobicity is the same between TMHs from single-pass and multi-pass proteins in the core of the helix and the flanks than the interfacial regions, particularly at the inner leaflet due to leucine asymmetry (Table 2.4)

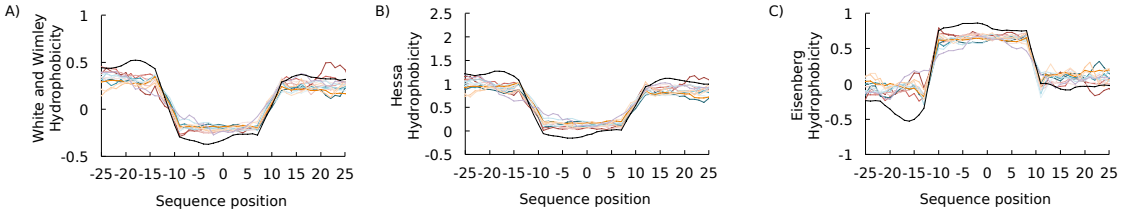

Figure 2.7: There is a difference in the hydrophobic profiles of TMHs from single-pass and multi-pass proteins. The difference in hydrophobicity between the single-pass and multi-pass datasets stratified by number of TMHs is not due to the choice of scale. As with Figure 2.6, UniHuman was stratified according to the number of TMHs in each protein. The mean amino acid hydrophobicity values of TMHs with a sliding unweighted window of 3 residues from UniHuman proteins at each position were plotted. To validate the findings presented in Figure 2.6A, several scales of hydrophobicity were used. (A) The White and Wimley whole residue scale [**White1999**] is based on the partitioning of peptides between water and octanol as well as water to POPC. A positive score indicates a more polar score. (B) The Hessa biological scale [**Hessa2005**]. The hydrophobicity values represent the free energy exchange during recognition of designed peptide TMHs by the endoplasmic reticulum Sec61 translocon and, therefore, negative values indicate an energetic preference for the interior of a lipid bilayer. (C) The Eisenberg consensus scale [**Eisenberg1984**] is a scale based on the earlier scales from Nozaki and Tanford [**Nozaki1971**], Wolfenden *et al.* [**Wolfenden1981**], Chothia [**Chothia1976**], Janin [**Janin1979**] and the von Heijne and Blomberg scale [**VonHeijne1979**]. The scales are normalised according to serine. A positive score indicates a generally more hydrophobic score.

This trend is statistically significant for TMHs in many biological membranes (Table 2.4, Figure 2.8). In the most extreme case of UniCress (single-pass), we see 49%

more leucine residues on the inside leaflet than the outside leaflet (P-value=5.41e-24). This contrasts with UniCress (multi-pass), in which the skew is far weaker, albeit yet statistically significant. There are 6% more leucine residues at the inside half (P-value=2.08e-4). The trend of having more leucine residues at the cytoplasmic half of the TMH is observed for all datasets (both single- and multi-pass) except for UniArch (single-pass). The phenomenon is statistically significant with P-value<1.e-3 for ExpAll, UniHuman, UniPM and UniCress (both single- and multi-pass). As with negative charge distribution, UniArch presents a reversed effect compared to other single-pass protein datasets with a 57% reduction in leucine on the inside leaflet compared to the outside leaflet (P-value=7.25e-6). However, leucine of TMHs from UniArch multi-pass proteins have no discernible preference for the inside leaflets (4% more on the inside leaflet, P-value=0.625).

Table 2.4: Leucines at the inner and outer leaflets of the membrane in TMHs The statistical results when comparing the number of leucine residues from the inner and outer leaflets in each protein in the dataset. The number of helices per dataset can be found in Table 2.1. The Kruskal-Wallis test scores (H statistics) were calculated for leucine residues by comparing the number of leucine residues that were in the inner half of the leaflet with those in the outer half of the leaflet of the database-defined TMH

Dataset	Single-pass					Multi-pass				
	Inside	Outside	Percentage	H statistic	P value	Inside	Outside	Percentage	H statistic	P value
ExpAll	4020	3403	118.13	40.07	2.44E-10	27,986	27,008	103.62	14.13	1.70E-04
UniHuman	4982	3697	134.76	193.02	6.99E-44	25,199	22,365	112.67	195.24	2.29E-44
UniER	359	297	120.88	8.41	3.72E-03	1863	1764	105.61	3.98	4.61E-02
UniGolgi	604	513	117.74	10.74	1.05E-03	753	677	111.23	5.61	1.79E-02
UniPM	1485	1006	147.61	98.9	2.65E-23	6221	5577	111.55	35.21	3.00E-09
UniCress	1495	1005	148.76	102.05	5.41E-24	6491	6099	106.43	13.76	2.08E-04
UniFungi	1389	1308	106.19	3.41	6.48E-02	14,505	14,099	102.88	6.74	9.41E-03
UniBacilli	260	251	103.59	0.03	8.72E-01	1488	1335	111.46	7.59	5.89E-03
UniEcoli	130	100	130	2.78	9.53E-02	7251	6975	103.96	5.92	1.50E-02
UniArch	51	118	43.22	20.13	7.25E-06	636	612	103.92	0.24	6.25E-01

2.4.6 A negative-outside (or negative-non-inside) signal is present across many membrane types

We explored the presence of amino acid residue compositional skews described above for human TMPs for those in other taxa and also specifically for human proteins with regard to membranes at various subcellular localisations. Acidic residues for TMHs from single-pass and multi-pass helices were plotted according to their relative percentage distributions (of the total amount of this residue type in the respective segment) for

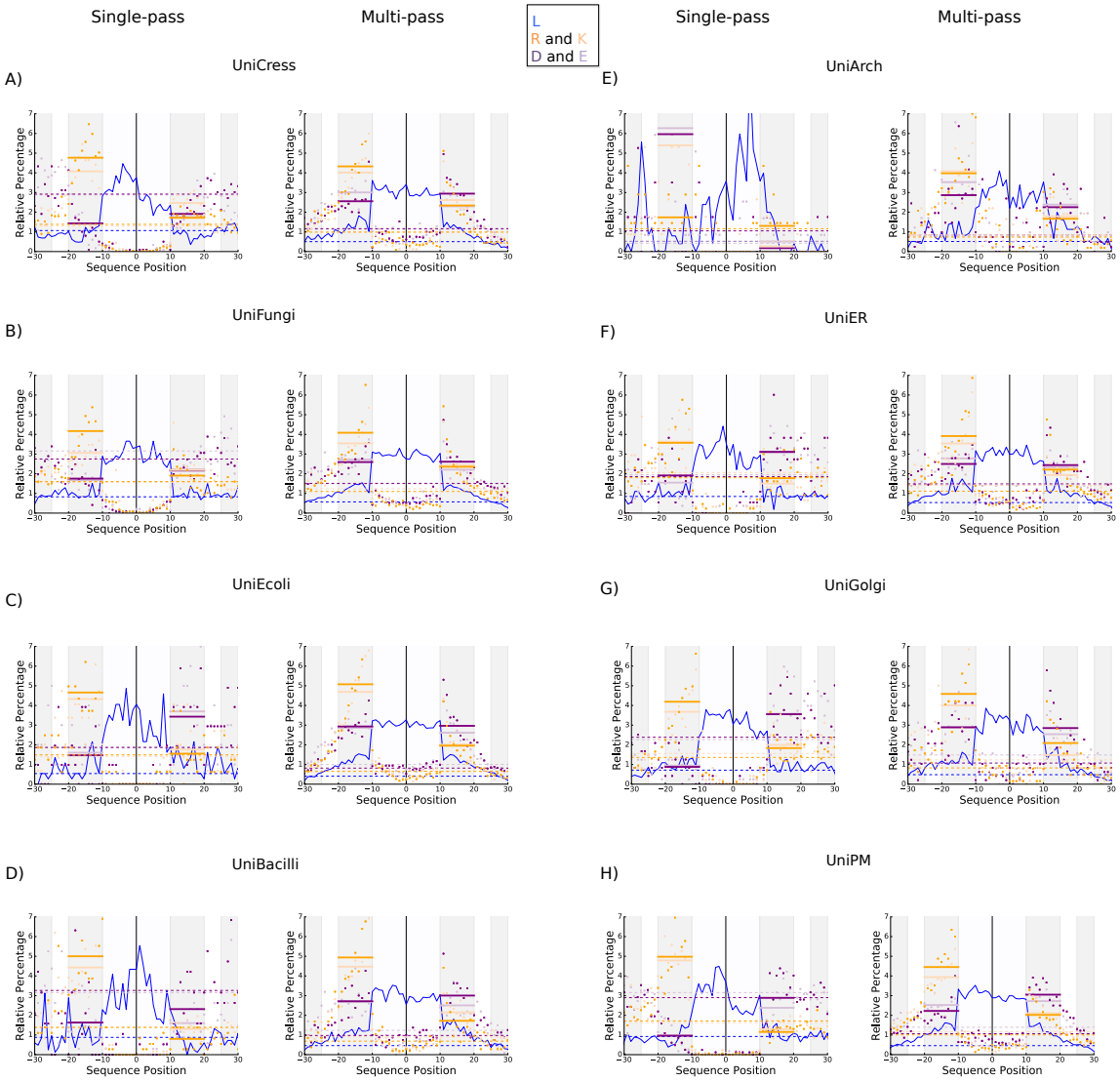

Figure 2.8: Comparing charged amino acid distributions in TMHs of multi-pass and single-pass proteins across different species and organelles. The relative percentage distribution of charged residues and leucine was calculated at each position in the TMH with flank lengths of ±20 in different datasets. The distributions are normalised according to relative percentage distribution. Aspartic acid and glutamic acid are shown in dark purple and light purple respectively. Leucine, the most abundant non-polar residue in TMHs, is in blue. Arginine and lysine are shown in orange. TMHs from single-pass proteins are on the left and TMHs from multi-pass proteins are on the right for different taxonomic datasets: a UniCress, b UniFungi, c UniEcoli, d UniBacilli, e UniArch, and different organelles: f UniER, g UniGolgi, h UniPM. As a trend, the negative-outside skew is more present in TMHs from single-pass proteins than multi-pass proteins (Tables 2 and 3). Another key observation is that in single-pass TMHs there is a propensity for leucine on the inner over the outer leaflet (Table 2.4)

five taxon-specific datasets UniCress (Figure 2.8A), UniFungi (Figure 2.8B), UniEcoli (Figure 2.8C), UniBacilli (Figure 2.8D), UniArch (Figure 2.8E) and for three organelle-specific datasets UniER (Figure 2.8F), UniGolgi (Figure 2.8G), UniPM (Figure 2.8H).

For single-pass proteins in all taxon-specific datasets (with the exception of UniArch), there are more negative residues at the outside than at the inside. The

skew is statistically significant (see Table 2.2, P<0.001) except for UniBacilli. Despite statistical significance found for UniFungi (P-value=1.12e-7 for database-defined and P-value=6.79e-10 for flanks after central alignment; Table 2.2), however, the trend is not very strong in this case (Figure 2.8B). Whereas the skew is just a suppression of negatively charged residues at the inside flank for ExpAll and UniHuman (as well as in UniCress), the bias observed for UniEcoli involves also a negative charge enrichment at the outside flank. In the case of UniArch (Figure 2.8E), we see a negative inside preference that is 6.0% in the case of aspartic acid, and 6.3% for glutamic acid (not shown), with much lower values close to 0% on the outside. Whilst the difference is statistically significant for both TMHs (Table 2.2) from single-pass proteins (P-value=1.83e-12 and P-value=1.43e-11 for two versions of flank determination) and multi-pass proteins (P-values 4.72e-3, 7.81e-3, 1.28e-4 for three versions of flank determination, see Tables 3A and 3B), the distribution along the position axis is heavily fluctuating, maybe as a result of the small size of the dataset. However, one can assuredly assign a "negative-inside" tendency to the flanking regions of Archaean TMHs.

In the human organelle datasets, we see trend shifts at different stages in the secretory pathway. In UniER, there is an enrichment of negative charge on the outside flank of 1–1.5% that is comparable to the magnitude of the positive inside signal. In UniGolgi, there is a suppression of negatively charged residues on the inside flank as well as an enrichment on the inside flank resulting in ∼2% distribution difference. For UniPM, there is a negative-inside suppression (but no outside enrichment) as well as a positive-inside signal. All observed trends are statistically significant (see Table 2.2, P<1.e-5).

For multi-pass TMH proteins, we see either the same trends but in a weaker form or no skews are observed at all as inspection of the graphs in Figure 2.8 shows. For datasets UniER, UniGolgi, UniCress, UniFungi, and UniBacilli, the hypothesis of equal distribution of negatively charged residues cannot be rejected (P-value>0.001, see Table 3); thus, a skew is statistically non-significant. Although UniPM has a statistically significant bias (P-value<4.30e-12, Table 3), the trends are more subtle and most present for aspartic acid of UniPM. We see many more negative and positive charges tolerated within the multi-pass TMHs themselves throughout all datasets (Table 2.1). To note, there is a positive-inside rule for all multi-pass datasets studied herein.

To conclude, we find that negative-charge bias distribution is a feature of single-pass protein TMHs that is present across many membrane types and it can have the form of a negative charge suppression at the inside flank or an enrichment of those charges at the outside flank.

2.4.7 Amino acid compositional skews in relation to TMH complexity and anchorage function

In previous work, we studied the relationship of TMH composition, sequence complexity and function [**Wong2010**, **Wong2011**, **Wong2012**] and concluded that simple TMHs are more probably responsible for simple membrane anchorage, whereas complex TMHs have a biological function beyond just anchorage. We wished to see how the skews observed in this work relate to that classification. Therefore, the single-pass TMHs from UniHuman and ExpAll were separated into subsets of simple, twilight, and complex TMHs using TMSOC [**Wong2011**, **Wong2012**]. The relative percentages of eight residue types (L, D, E, R, K, Y, W, C; normalisation with the total amount of residues of that amino acid type in all sequence segments considered) were plotted along the sequence position for simple and complex helices (Figure 2.9). Of UniHuman single-pass proteins, there were 889 records with simple TMHs and 570 with complex TMHs (Figure 2.9B). In ExpAll, 769 TMHs from single-pass proteins were simple TMHs and 570 were complex TMHs.

It is visually apparent (Figure 2.9) that there are (i) stronger skews and more inside-outside disparities in simple single-pass TMs than in complex single-pass TMs and (ii) greater similarities between single-pass complex TM regions and those from multi-pass proteins compared with simple single-pass TMs in comparison with either of the other two distributions. To examine the statistical significance of these observations, we compared the amino acid distributions (K, R, K+R, D, E, D+E, Y, W, L, C) across the range of TMHs with flank lengths ±10 residues using the Kolmogorov-Smirnov (KS), KW and the χ^2 statistical tests. To note, the KS test scrutinises for significant maximal absolute differences between distribution curves; the glskw test is after skews between distributions and the χ^2 statistical test checks the average difference between distributions. Calculations were carried out over single-pass complex,

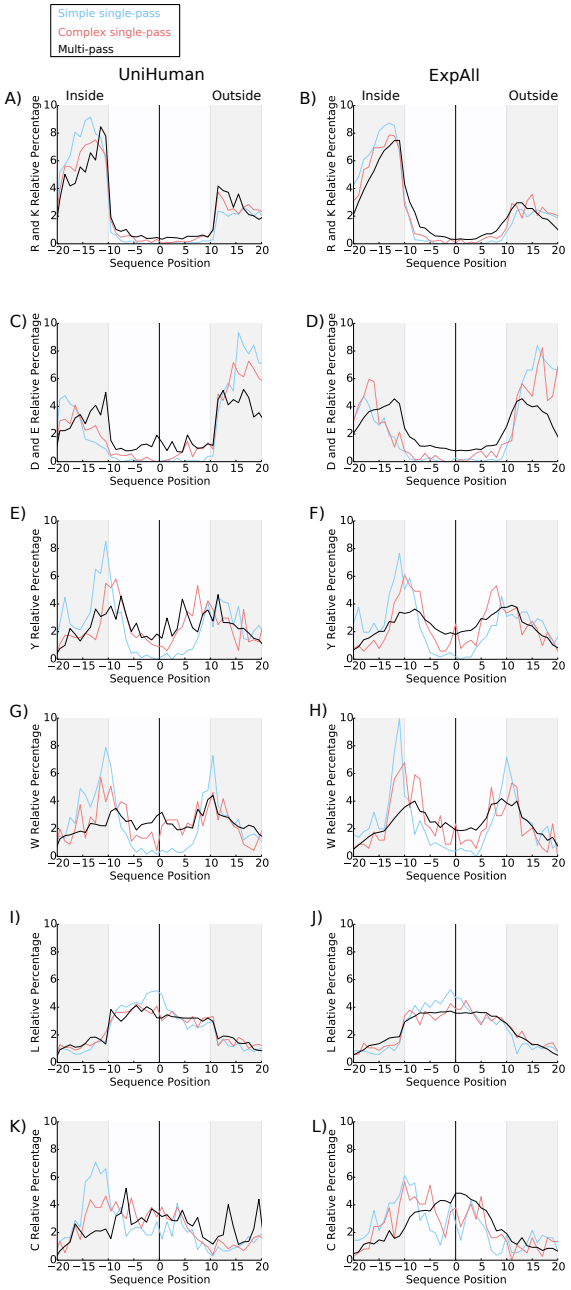

Figure 2.9: Comparing the amino acid relative percentage distributions of simple and complex TMHs from single-pass proteins and TMHs from multi-pass proteins. Comparing the amino acid relative percentage distributions of simple and complex TMHs from single-pass proteins and TMHs from multi-pass proteins. TMSOC was used to calculate which single-pass TMHs were complex and which were simple from ExpAll and UniHuman datasets. Simple TMHs are typically anchors without necessarily having other functions (Wong *et al.* [**Wong2010**]). The relative percentages from single-pass simple (shown in light blue), single-pass complex (red), and multi-pass protein TMHs (black) were plotted for (a, c, e, g, i and k) UniHuman and (b, d, f, h, j and l) ExpAll for (a and b) positive residues, (c and d) negative residues, (e and f) tyrosine, (g and h) tryptophan, (i and j) leucine and (k and l) cysteine. The slopes are statistically compared in Tables 5 and 6, and as a trend, the profiles of complex TMHs are more similar to multi-pass TMH profiles than simple TMHs are to multi-pass TMHs

single-pass simple and multi-pass TMH datasets from both ExpAll and UniHuman (for P-values and Bahadur slopes, Table 2.5 (dataset UniHuman) and Table 2.6 (dataset ExpAll)).

Many low P-values in Tables 2.5 and 2.6 indicate significant differences between the three distributions studied. For the UniHuman dataset (Table 2.5), we find most striking, significant differences between charged residue distributions (R, K, D, E) of simple and complex single-pass TMH+flank regions (χ^2 P-value<2.23e-3 for single amino acid types). Similarly, simple single-pass TMH+flank segments differ significantly from multi-pass TMH+flank segments (KW test P-values<3.e-2 for R, K, D, E, Y, W amino acid types as well as for K+R and D+E). The trends are the same for the ExpAll dataset (Table 2.6): simple and complex single-pass TMH+flank regions differ in charged amino acid type distributions (χ^2 P-value<4.21e-3 for all cases), as well as simple single-pass and multi-pass ones, do (KW test P-values<5.e-2 for R, D, E, Y, W amino acid types and D+E).

Whereas P-value tests for significant differences between distributions depend strongly on the amount of data, the more informative Bahadur slopes that measure the distance from the zero hypothesis are independent of the amount of data [**Bahadur1967**, **Bahadur1971**, **Sunyaev1998**]. As we can see in Tables 2.5 and 2.6, the absolute Bahadur slopes for the simple single-pass to multi-pass comparison are always larger (even by at least an order of magnitude): (ii) for all three statistical tests applied (χ^2, KS and KW), (ii) for all amino acid types, for K+R and E+D and (iii) for both datasets UniHuman and ExpAll. Thus, complex single-pass TMH+flanks have compositional properties that are indeed very similar to those of multi-pass ones (which are known to have a large fraction of complex TMHs [**Wong2011**, **Wong2012**]). This strong evidence implies that the actual issue is not so much about single- and multi-pass TMH segments but between simple and complex TMHs where the first are exclusively guided by the anchor requirements whereas the latter have more complex restraints to fulfil.

Several distribution features of simple TMHs from single-pass proteins when compared to complex TMHs from single-pass proteins and TMHs from multi-pass proteins that contribute to the statistical differences (Figure 2.9) are especially notable. There

Table 2.5: Simple TMHs are less similar than complex TMHs to TMHs from multi-pass proteins in UniHuman The statistical results were gathered by comparing complex single-pass TMHs, simple TMHs from single-pass proteins and TMHs from multi-pass proteins in UniHuman. The abundance of different residues at each position when using the centrally aligned TMH approach was compared with several statistical tests (the KS, KW and the χ^2 statistical tests) and the Bahadur slope values of those results

Residues	P values for χ^2			Bahadur slopes for χ^2		
	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi
R	3.20E-06	7.38E-02	1.24E-01	6.61E-03	2.20E-03	1.27E-04
K	2.23E-03	4.99E-02	2.14E-01	3.99E-03	3.70E-03	1.18E-04
D	1.67E-09	3.06E-01	3.02E-01	3.34E-02	3.24E-03	1.20E-04
E	3.80E-07	2.34E-01	2.31E-01	1.81E-02	3.05E-03	1.36E-04
Y	3.86E-01	3.97E-01	2.11E-01	1.06E-03	1.47E-03	8.25E-05
W	3.77E-03	2.97E-01	3.84E-01	8.52E-03	2.73E-03	1.13E-04
L	3.59E-01	2.88E-01	3.21E-01	1.52E-04	3.92E-04	1.69E-05
C	6.44E-01	3.97E-01	3.41E-01	4.29E-04	1.29E-03	8.57E-05
R+K	2.19E-02	2.83E-01	2.52E-01	1.11E-03	6.33E-04	4.68E-05
D+E	1.47E-03	2.86E-01	2.79E-01	4.59E-03	1.49E-03	6.15E-05
	P values for Kolmogorov-Smirnov			Bahadur slopes for Kolmogorov-Smirnov		
	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi
R	2.31E-01	3.57E-04	1.08E-02	7.66E-04	6.71E-03	2.76E-04
K	4.31E-02	2.18E-03	8.93E-01	2.06E-03	7.56E-03	8.68E-06
D	1.39E-01	5.02E-06	1.08E-02	3.26E-03	3.34E-02	4.52E-04
E	7.96E-02	1.58E-05	1.08E-02	3.10E-03	2.32E-02	4.20E-04
Y	7.96E-02	2.22E-02	2.31E-01	2.81E-03	6.07E-03	7.78E-05
W	2.31E-01	9.06E-04	4.31E-02	2.24E-03	1.58E-02	3.70E-04
L	2.31E-01	2.31E-01	5.31E-01	2.17E-04	4.61E-04	9.42E-06
C	1.39E-01	3.61E-01	3.61E-01	1.93E-03	1.42E-03	8.10E-05
R+K	7.96E-02	1.33E-04	7.96E-02	7.35E-04	4.48E-03	8.60E-05
D+E	4.31E-02	1.58E-05	4.98E-03	2.21E-03	1.31E-02	2.55E-04
	P values for Kruskal-Wallis			Bahadur slopes for Kruskal-Wallis		
	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi
R	2.19E-01	5.06E-02	2.37E-01	7.92E-04	2.52E-03	8.79E-05
K	2.90E-01	1.33E-01	7.00E-01	8.11E-04	2.49E-03	2.73E-05
D	3.50E-01	1.81E-02	2.81E-01	1.74E-03	1.10E-02	1.27E-04
E	2.59E-01	5.65E-02	1.78E-01	1.65E-03	6.04E-03	1.60E-04
Y	6.03E-01	4.53E-01	4.41E-01	5.62E-04	1.26E-03	4.34E-05
W	4.19E-01	1.84E-01	5.70E-01	1.33E-03	3.81E-03	6.62E-05
L	6.37E-01	4.88E-01	9.77E-01	6.68E-05	2.25E-04	3.47E-07
C	5.00E-01	2.22E-01	9.62E-01	6.76E-04	2.10E-03	3.11E-06
R+K	1.87E-01	8.67E-02	4.08E-01	4.86E-04	1.23E-03	3.05E-05
D+E	1.68E-01	4.52E-02	1.91E-01	1.25E-03	3.68E-03	7.97E-05

Table 2.6: Simple TMHs are less similar than complex TMHs to TMHs from multi-pass proteins in ExpAll As in Table 2.5, the statistical results were gathered by comparing complex single-pass TMHs, simple TMHs from single-pass proteins and TMHs from multi-pass proteins; however, in this case only ExpAll is used. The abundance of different residues at each position when using the centrally aligned TMH approach was compared with several statistical tests (the KS, KW and the χ^2 statistical tests) and the Bahadur slope values of those results

Residues	P values for χ^2			Bahadur slopes for χ^2		
	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi
R	5.10E-06	2.98E-01	5.10E-06	9.17E-03	1.61E-03	6.23E-05
K	2.35E-03	1.85E-01	2.35E-03	4.81E-03	3.88E-03	9.78E-05
D	2.61E-08	1.84E-01	2.61E-08	4.15E-02	7.90E-03	1.41E-04
E	2.38E-10	2.04E-01	2.38E-10	3.88E-02	7.08E-03	1.22E-04
Y	3.03E-01	3.11E-01	3.03E-01	2.01E-03	2.49E-03	5.51E-05
W	4.21E-03	4.29E-01	4.21E-03	1.11E-02	4.76E-03	6.46E-05
L	3.79E-01	3.04E-01	3.79E-01	2.28E-04	4.66E-04	1.50E-05
C	3.87E-01	2.52E-01	3.87E-01	1.75E-03	3.28E-03	1.48E-04
R+K	7.16E-04	2.52E-01	7.16E-04	2.80E-03	1.28E-03	3.76E-05
D+E	3.58E-05	2.94E-01	3.58E-05	1.03E-02	1.94E-03	4.90E-05
	P values for Kolmogorov-Smirnov			Bahadur slopes for Kolmogorov-Smirnov		
	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi
R	3.61E-01	4.31E-02	3.61E-01	7.66E-04	7.79E-03	1.62E-04
K	4.31E-02	8.93E-01	4.31E-02	2.49E-03	1.05E-02	6.57E-06
D	1.39E-01	2.18E-03	1.39E-01	4.68E-03	3.61E-02	5.10E-04
E	5.31E-01	1.33E-04	5.31E-01	1.11E-03	2.81E-02	6.87E-04
Y	2.31E-01	9.06E-04	2.31E-01	2.47E-03	6.26E-03	3.30E-04
W	5.31E-01	4.98E-03	5.31E-01	1.29E-03	1.13E-02	4.04E-04
L	2.31E-01	2.31E-01	2.31E-01	3.45E-04	2.12E-03	1.85E-05
C	5.31E-01	3.61E-01	5.31E-01	1.16E-03	8.91E-04	1.09E-04
R+K	1.39E-01	2.31E-01	1.39E-01	7.61E-04	4.82E-03	4.00E-05
D+E	1.39E-01	9.06E-04	1.39E-01	1.99E-03	1.41E-02	2.80E-04
	P values for Kruskal-Wallis			Bahadur slopes for Kruskal-Wallis		
	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi	Simple-vs-complex	Simple-vs-multi	Complex-vs-multi
R	4.37E-01	3.92E-01	4.37E-01	6.24E-04	2.52E-03	4.82E-05
K	3.83E-01	6.93E-01	3.83E-01	7.62E-04	2.88E-03	2.13E-05
D	4.49E-01	1.81E-01	4.49E-01	1.90E-03	1.06E-02	1.42E-04
E	7.64E-01	1.94E-01	7.64E-01	4.71E-04	9.05E-03	1.26E-04
Y	8.32E-01	3.36E-01	8.32E-01	3.09E-04	9.63E-04	5.15E-05
W	7.25E-01	1.36E-01	7.25E-01	6.53E-04	5.44E-03	1.52E-04
L	7.15E-01	7.95E-01	7.15E-01	7.90E-05	3.41E-04	2.90E-06
C	8.47E-01	9.54E-01	8.47E-01	3.05E-04	4.26E-05	5.06E-06
R + K	2.89E-01	5.13E-01	2.89E-01	4.79E-04	1.41E-03	1.82E-05
D+E	4.94E-01	2.07E-01	4.94E-01	7.11E-04	4.14E-03	6.29E-05

is a more pronounced trend for positively charged residues and tyrosine to be preferentially located on the inside flanks and for negatively charged residues to be on the outside flanks. The symmetrical peaks in the percentage distribution of tyrosine in complex single-pass TMHs are more akin to multi-pass TMHs, whereas in simple TMHs the distribution resembles a more typical single-pass helix (compare with Figure 2.2). Furthermore, the depression of charged residues within the TMH itself is strongest in simple single-pass TMHs.

To emphasise, tryptophan is essentially not tolerated within the simple TMHs and there are higher peaks of tryptophan occurrence at either flank. We also see a strong inside skew for leucine clustering within the core of simple TMHs which is not present in the "flatter" distributions of complex single-pass TMHs and TMHs from multi-pass proteins.

There is obviously a cysteine-inside preference for simple, single-pass TMHs but less in complex, multi-pass TMHs (Figure 2.9). This conclusion is contrary to a previous study [**Nakashima1992**] but that deduction was drawn from a much smaller dataset of 45 single-pass TMHs and 24 multi-pass TMPs.

2.5 Discussion

The "negative-outside/non-negative inside" skew in TMHs and their flanks is statistically significant We have seen that, consistently throughout the datasets, there is a trend for generally rare negatively charged residues to prefer the outside flank of a TMH rather than the inside (and to almost completely avoid the TMH itself); be it by suppression on the inside and/or enrichment on the outside. The trend is much stronger in single-pass protein datasets than in multi-pass protein datasets. However as we elaborated on further, the real crux of the bias appears to be associated with the TMH being simple or complex [**Wong2011**, **Wong2012**], thus, whether or not the TMH has a role beyond anchorage. The existence of this bias has implications for topology prediction of proteins with TMHs, engineering membrane proteins as well as for models of protein transport via membranes and protein-membrane stability considerations.

It should be noted that the controversy in the scientific community about the existence of a negative charge bias at TMHs was mainly with regard to multi-pass TMPs. Despite having access to much larger, better annotated sequence datasets and many more 3D structures than our predecessors, we also had our share of difficulties here (see Results section III and Table 3). The straightforward approach results in inconclusive statistical tests if datasets become small (for example, if selections are restricted to subcellular localisations, 3D structures or if very harsh sequence redundancy criteria are applied) and, especially, if TMHs with very short or no flanks are included. Therefore in the case of multi-pass proteins, we studied flanks as taken from the TM boundaries in the databases under several conditions: (i) without allowing flank overlap between neighbouring TMHs, (ii) as subset of (i) but with requiring some minimal flank length at either side, (iii) with overlapping flanks. We also studied flanks after central alignment of TMHs and assuming standardised TMH length. Multi-pass TMHs (without overlapping flanks) do not show statistically significant negative charge bias under condition (i) but, apparently, due to many TMHs without any or super-short flanks at least at one side. Significance appears as soon as subsets of TMHs with flanks at both sides are studied. Not surprisingly, there is no charge bias if there are no flanks in the first place. It is perhaps worth noting that the results from multi-pass TMHs with overlapping flanks may involve amplification of skews since it involves multiple counting of the same residues. Given the redundancy threshold of UniRef90, we cannot rule out that these statistical skews are the result of a trend from only a small sub-group of TMPs which is being amplified. Hence, we also needed to observe if these same observed biases were true in condition (ii), which is indeed the case.

As the "negative-outside/negative-not-inside" skew is widely observed among varying taxa and subcellular localisations with statistical significance, it appears to, at least to a certain extent, be caused by physical reasons and be associated with the background membrane potential. Several earlier considerations and observation support this thought: (i) Firstly, a concert between the negative and positive

charge on the TMH flanks drives anchorage and the direction of insertion of engineered TMHs [**Sipos1993**, **Hartmann1989**]. (ii) The inner leaflet of the plasmalemma tends to be more negatively charged [**Zachowski1993**]. Specifically, phosphatidylserine was found to distribute in the cytosolic leaflets of the plasma membrane and it was found to electrostatically interact with moderately positive-charged proteins enough to redirect the proteins into the endocytic pathway [**Yeung2008**]. The negative charge of proteins at the inside of the plasma-membrane would decrease the anchoring potency of the TMH via electrostatic repulsion. (iii) Thirdly in membranes that maintain a membrane potential, there are inevitably electrical forces acting on charged residues during chain translocation as this influences the translocon machinery when orienting the TMH. Therefore, it is no surprise that we see an inside-outside bias for negatively charged residues that is opposite to the one for positively charged residues. The negative charges in TMH residues have been shown to experience an electrical pulling force as they pass through the bacterial SecYEG translocon import [**Ismail2012**, **Ismail2015**]. Also, they are known to be involved in intra-membrane helix-helix interactions [**Meindl-Beinker2006**]. For example, aspartic acid and glutamic acid can drive efficient di- or trimerisation of TMHs in lipid bilayers and, furthermore, that aspartic acid interactions with neighbouring TMHs can directly increase insertion efficiency of marginally hydrophobic TMHs via the Sec61 translocon [**Meindl-Beinker2006**]. In support of this, less acidic residues are found in single-pass TMHs, among which only some will undergo intra-membrane helix-helix interactions. As the mutation studies have shown negative charge as a topological determinant [**Nilsson1990**], therefore, it is perhaps no surprise that we observe a skew in negatively charged residues in a similar manner to the skew in positively charged residues.

Whereas the "negative-outside/negative-not-inside" skew is observed for distantly related eukaryotic species and it is also present in Gram-negative bacteria such as *E. coli*, this sequence pattern was not observed for the Gram-positive bacteria in which there is no observable bias. In contrast, Archaea have a statistically significant "negative-inside" propensity both for single- and multi-pass TMPs. It is known that Archaea have remarkably different membranes compared to other kingdoms of life due to their extremophile adaptations to stress [**Oger2013**]. Whilst it is unclear

why negative charge is distributed so differently in UniArch to the other taxonomic datasets, one must appreciate that a much more nuanced approach would be needed to draw formal conclusions about Archaea, which current databases cannot provide due to the relatively limited information and annotation of Archaean proteomes.

Methodological issues made previous studies struggle to identify negatively charged skews with statistical significance

Whereas the influence of a negative charge bias in engineered proteins with TM regions on the direction of insertion into the membrane was solidly established [**Nilsson1990**, **Andersson1993**, **Kim1994**, **Andersson1992**, **Rutz1999**], the search for the negative charge distribution pattern in the statistics of sequences of TM proteins from databases failed to find significance for the expected negative charge skew [**Sharpe2010**, **Baeza-Delgado2013**, **Granseth2005**, **Pogozheva2013**, **Nilsson2005a**, **Andersson1992**].

Generally speaking, the datasets from previous studies have been considerably smaller compared with those in our work (only Sharpe *et al.* had a similar order of magnitude [**Sharpe2010**]), especially those with experimental information about 3D structure and membrane topology that we used for validation. And they might not have had the luxury of using UniProts improved TRANSMEM consensus annotation based on a multitude of TM prediction methods and experimental data, but this is also not the major issue. We found that there are other factors that are critical for observing sequence bias such as negative charge skew in the case of TMHs.

i Acidic residues are rare near and within TMH and biases in their distribution are easily blurred by minor fluctuations of much more frequent amino acid types, most notably leucine. Therefore, the method of normalisation is critical. We have shown that normalising by the total amount of residues of the amino acid type studied within the sequence region under consideration is appropriate to answer the question where to find a negatively charged residue if there is any at all (called "relative percentage" in this work).

ii The alignment of the TMHs is critical. It was common practice to align TMH according to the most cytosolic residue [**Sharpe2010**] although it is known that the membrane/cytosol boundary of the TMH is not well defined (and the exact

boundary is even less well understood at the non-cytosolic side). Aligning the TM regions and their flanks from the center of the TMH was first proposed by Baeza-Delgado *et al.* [**Baeza-Delgado2013**]. Since we know now that acidic residues are often suppressed in the cytosolic flank and within the TMH, this implies that the few acidic residues found in the cytosolic interface would appear more comparable to those in the poorly defined non-cytosolic interface as the respective residues are spread over more potential positions, diminishing any observable bias.

iii We find that separation into single- and multi-pass TM datasets (or, even better, simple and complex TMHs [**Wong2011**, **Wong2012**]) is critical to study the inside/outside bias. As many TMHs in multi-pass TMPs have essentially no flanks or very short flanks if the condition of non-overlap is applied to flanks of neighbouring TMHs, this might also obscure the observation of the negative charge bias. If there are no flanks, then there will be no residue distribution bias in these flanks. The problem can be alleviated by either studying only subsets with minimal flank lengths on both sides (although datasets might become too small for statistical analysis) or by allowing flank overlaps between neighbouring TMHs.

iv This classification is even more justified in the light of previous reports about the "missing hydrophobicity" in multi-pass TMHs [**Nilsson1990**, **Hedin2010**, **Hessa2007**, **Ojemalm2012**]. Otherwise, the distribution bias well observed among the exclusive anchors could be lost to noise. This addresses the more biologically contextualised issue that there are different evolutionary pressures on different types of TMHs. The negative charge skew is most pronounced for dedicated anchors frequently found with simple TMHs typically observed in single-pass TM proteins. These TMHs are pressured to exhibit residue biases that may aid anchorage in a topologically correct manner. Complex TMHs, typically within multi-pass membrane proteins that have a function beyond anchorage, comply with a multitude of restraints structural and functional constraints and the negative charge skew is just one of them.

The most representative precedent papers are those of Sharpe *et al.* [**Sharpe2010**] from 2010 (with 1192 human and 1119 yeast single-pass TMHs), Baeza-Delgado *et al.* [**Baeza-Delgado2013**] (with 792 TMHs mixed from single- and multi-pass TMPs)

and Pogozheva *et al.* [**Pogozheva2013**] (TMHs from 191 mixed from single- and multi-pass TMPs with structural information) both from 2013. Whereas the first analysis would have benefited from the central alignment approach and the first two studies from another normalisation as described above, the third study did come close to our findings. To note, their dataset mixed with single- and multi-pass proteins was too small for revealing the negative charge bias with significance; yet, they observed total charge differences at either sides of the membrane varying for both single- and multi-pass proteins. Membrane asymmetry due to positively charged residues occurring more frequently on the cytosolic side causes net charge unevenness at both sides of the membrane. This observation has been known to correlate with orientation for decades [**VonHeijne1989**, **Baeza-Delgado2013**, **Meindl-Beinker2006**]. Our data shows that the negative charge skew contributes to this asymmetry.

There are differences in charged amino acid residue biases in TMH flanks through each stage of the secretory pathway

Here, we observe differences throughout sub-cellular locations along the secretory pathway. We found that negative charges are enriched at the outside flank (in the ER), both enriched outside and suppressed inside for the Golgi membrane, and suppressed on the inside flank in the Plasma Membrane (PM). It has been suggested that the leaflets of different membranes have different lipid compositions throughout the secretory pathway [**VanMeer2008**] and this has led to general biochemical conservation in terms of TMH length and amino acid composition in different membranes [**Sharpe2010**, **Pogozheva2013**].

Lipid asymmetry in the Golgi and PM (in contrast to the ER) has been known about for over a decade [**Daleke2007**, **Devaux2004**]. To note, the Golgi and PM have lipid asymmetry with sphingomyelin and glycosphingolipids on the non-cytosolic leaflet, and phosphatidylserine and phosphatidylethanolamine enriched in the cytosolic leaflet. Although the ER is the main site for cholesterol synthesis, it has markedly low concentrations of sphingolipids [**Bell1981**]. Golgi synthesises sphingomyelin, a lipid not present in the ER, but present in both the Golgi [**Futerman2005**] and in the PM [**Li2007**, **Tafesse2007**]. The PM is also enriched with densely packed sphingolipids and sterols [**Paolo2006**]. Another factor influencing the sequence patterns

of TMHs and their along the secretory pathway appears to be the variation in membrane potentials [**Qin2011**, **Worley1994**, **Schapiro2000**].

Several sequence features can be assigned to anchor TMHs: Charged-residue flank biases, leucine intra-helix asymmetry, and the "aromatic belt".

We investigated the difference between TMHs from single-pass and multi-pass proteins and found significant differences in sequence composition that are reflective of the biologically different roles the TMHs play. To emphasise and validate these findings, we separated TMHs from single-pass proteins into simple and complex TMHs [**Wong2011**, **Wong2012**]; ones that likely contains mostly TMHs that act as exclusive anchors, and another that have roles beyond anchorage. This leaves us with "anchors" (simple TMHs from single-pass proteins) and "non-anchors" (complex TMHs from single-pass proteins, and TMHs from multi-pass proteins). If there are strong sequence feature differences between anchors and non-anchors, it is likely that the sequence feature has a role in satisfying membrane constraints to act as an energetically optimally stable anchor.

Future studies in the area would desirably directly include a comprehensive analyses of datasets oligomerised TMHs from single-pass proteins and ascertain if they appear to be more similar to simple anchors, multi-pass, or generally neither. Currently, no sufficiently complete set of intra-membrane oligomerised single-pass proteins exists that can be compared to a large set of known non-oligomerising proteins. The current work sidesteps this issue by comparing single-pass proteins with simple TMHs, which tend to be simple anchors (as shown in previous work [**Wong2011**, **Wong2012**]), against datasets that contain TMHs that will form intra-membrane bundles. Bluntly, the simple/complex status of a TMH can be easily computed from its sequence with TMSOC whereas the oligomerisation state of most membrane proteins still needs to be experimentally determined.

Unsurprisingly, both positively and negatively charged residues can be seen to be more strongly distributed with bias in anchors than non-anchors. Both the "positive-inside" rule as well as the "negative-outside/non-negative-inside" bias are mostly observable in simple single-pass TMHs (although they are statistically significant elsewhere). It is perhaps true that where a bias is clearly present in both non-anchors and

anchors alike, it is a strong topological determinant, whereas if the residue is only distributed with topological bias in exclusively anchoring TMHs, we can attribute these features more specifically to biophysical anchorage. This being said, we should not rule out that the same features aid topological determination since negative charge has been shown to be a weaker topological determinant than positively charged residues (35).

Tyrosine and tryptophan residues commonly are found at the interfacial boundaries of the TMH and this feature is called the "aromatic belt" [**Sharpe2010**, **Baeza-Delgado2013**, **Granseth2005**, **Nilsson2005a**, **Hessa2005**] and this was thought to be caused by their affinity to the carbonyl groups in the lipid bilayer [**Killian2000**]. Not all types of aromatic residues are found in the aromatic belt; phenylalanine has no particular preference for this region [**Granseth2005**, **Braun1999**]. It is still unclear if the aromatic belt has to do with anchorage or with translocon recognition [**Baeza-Delgado2013**]. Here, TMHs with exclusively anchorage functions showed stronger preferences for the W and Y in the aromatic belt region, otherwise known as the water-lipid interface region than TMHs with function beyond anchorage. This is strong evidence that the aromatic belt indeed assists with anchorage, and is less conserved where the TMH must conform to other restraints beyond membrane anchorage. Furthermore, we see that the tyrosine's preference for the inside interface region also appears to be to do with anchorage and this trend is somewhat true for tryptophan, too.

Finally, our findings corroborate earlier reports that many multi-pass TMHs are much less hydrophobic than typical single-pass TMH and about 30% of them fail the hydrophobicity requirements of ΔG TMH insertion prediction ("missing hydrophobicity") [**Hessa2005**, **Hedin2010**, **Hessa2007**, **Ojemalm2012**]. We also find that the leucine skew and the hydrophobic asymmetry towards the cytosolic leaflet of the membrane is more pronounced in simple, single-pass TMHs than in complex or multi-pass ones; thus, it appears to be another anchoring feature. It was found previously that the hydrophobic profiles of TMHs of multi-pass proteins share similar hydrophobicity profiles on average irrespective of the number of TMHs and TMHs from single-pass proteins have been found to be typically more hydrophobic than TMHs from multi-pass

proteins [**Wong2011**]. Sharpe *et al.* [**Sharpe2010**] report an asymmetric hydrophobic length for single-pass TMHs. Our study reiterates the hydrophobic asymmetry and attributes it mainly to the leucine distribution. The leucine asymmetry might be linked to the different lipid composition of either leaflet of biological membranes.

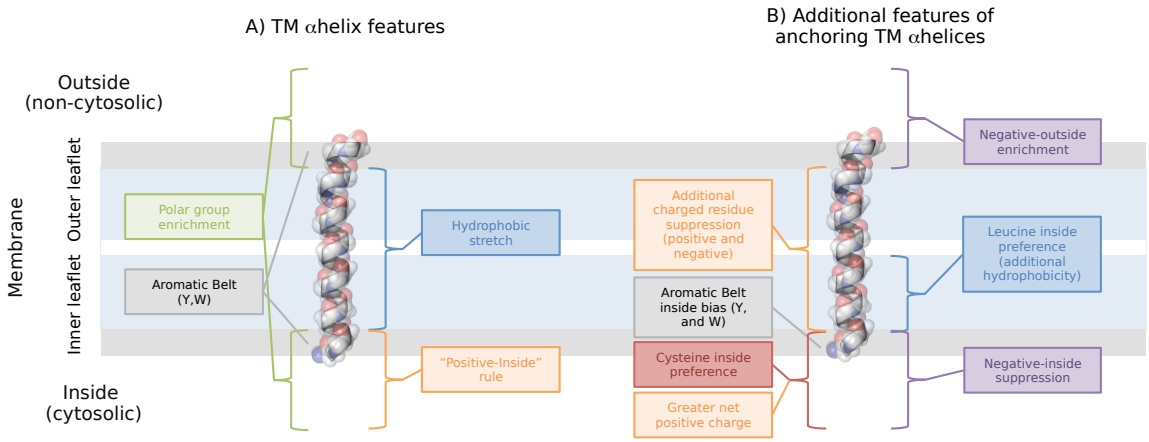

Figure 2.10: Residue distributions of transmembrane anchors. A view showing additional residue distribution features that TMHs with an anchorage function display.a The more classic model of a TMH showing the "positive-inside" rule [**VonHeijne1989**], the hydrophobic core [**Kyte1982**], the polar enrichment that flanks the hydrophobic stretch [**Baeza-Delgado2013**] and the aromatic belt [**Granseth2005**]. b Simple anchors may display additional features that conform to the membrane biophysical constraints: further suppression of charge in the hydrophobic core (Table 2.1), intra-membrane leucine asymmetry that likely causes hydrophobic skew [**Sharpe2010**] (Table 2.4, Figure 2.6), a higher preference for cysteine on the inside flanking region (Figure 2.9K and L), a higher net "positive-inside" charge (Figure 2.4), asymmetric skew of the hydrophobic belt favouring the inner leaflet interface (Figure 2.9E, F, G, and H) and a negative-outside bias via suppression on the inside flanking region or enrichment on the outside flanking region (Figure 2.9C and D, Tables 2 and 3)

In summary, three key features can be assigned to aiding TMH stability in the membrane (Figure 2.10): (i) charge, (ii) the aromatic belt, and (iii) leucine leaflet preference. What is most novel here is that each of these features are furthermore distributed with preference for a particular side of the bilayer in the case of anchoring TMHs. These differences in inside-outside topology that are most present in anchoring TMHs further supports the notion that there are broad lipid compositional differences between the inner and outer leaflets of the bilayers [**Sharpe2010**]. Furthermore, while some TMHs conform and complement to the properties of the bilayer, other TMHs with function beyond anchorage are less constrained to biophysically complement the bilayer. For these TMHs, any advantage gained by adhering to the membrane restrictions is outweighed by more complicated protein dynamics, topological frustration and protein functional requirements.

To conclude, the large fraction of functionally uncharacterised genomic sequences is the great bottleneck in life sciences at this moment that hinders many biomedical and biotechnological applications, some with tremendous societal need [**Eisenhaber2012**, **Kuznetsov2013**]. Among these uncharacterised genomic regions, there is ∼10000 protein-coding genes, especially many membrane-embedded proteins. It is hoped that the NNI/NO-rule as well as the other sequence properties of membrane anchoring TMHs described in this article will add new insights for membrane protein function discovery, design and engineering.

2.6 Methods

2.6.1 Datasets

Databases.

All datasets used for analysis are listed in Table 2.1. Transmembrane protein sequences and annotations were taken from TOPDB [**Dobson2015**] and UniProt [**TheUniProtConsortium2014**]. UniProt derived datasets are the most comprehensive datasets built with (i) robust transmembrane prediction methods providing the limit of todays achievable accuracy with regard to hydrophobic core localisation and (ii) subcellular location annotation that can be used for orientation determination. However, they mostly rely on predicted transmembrane regions. TOPDB has meticulous experimental verifications of the orientation from the literature that are independent of prediction algorithms [**Dobson2015**]. Unfortunately, this dataset is much smaller with too few entries to have it divided with regard to taxonomy or subcellular locations.

UniProt database files were downloaded by querying the server for different taxonomic groups as well as different subcellular membrane locations; UniHuman (human representative proteome), UniCress (Arabidopsis thaliana, otherwise known as mouse eared cress, representative proteome), UniER (human endoplasmic reticulum representative proteome), UniPM (human plasma membrane representative proteome), UniGolgi (human Golgi representative proteome). To enforce a level of quality control, the queries were restricted to manually reviewed

records and transmembrane proteins with manually asserted TRANSMEM annotation [**TheUniProtConsortium2014**]. Proteins were then sorted into multi-pass and single-pass groups according to having more than one or exactly one TRANSMEM region respectively. TRANSMEM regions are validated by either experimental evidence [**TheUniProtConsortium2014**], or according to a robust transmembrane consensus of the predictors TMHMM [**Krogh2001**], Memsat [**Jones2007**], Phobius [**Kall2004**, **Kall2007**] and the hydrophobic moment plot method of Eisenberg and co-workers [**Eisenberg1984**]. TMHs and flanking regions were oriented according to UniProt TOPO_DOM annotation according to the keyword "cytoplasmic". If a "cytoplasmic" TOPO_DOM was found in the previous TOPO_DOM relative to the TRANSMEM region then the sequence remained the same. If "cytoplasmic" was found in the next TOPO_DOM, relative to the TRANSMEM section then the sequence was reversed. Proteins without the "cytoplasmic" keyword in their TOPO_DOM annotation were omitted from further analysis.

The TOPDB database [**Dobson2015**] is a manually curated database composed of experimental records from the literature that allow determination of the protein topology. Experiments include fusion proteins, posttranslational modifications, protease experiments, immunolocalization, chemical modifications as well as revertants, sequence motifs with known mandatory membrane-embedded topologies, and tailoring mutants (Table 2.7).

Length cut-offs for the TMH were set at 16 as the shortest length and 38 as the longest.

To note, we are aware that proteome datasets are a moving target that have dramatically changed over the years and, probably, will continue to do so to some extent in the future[83]. Yet, we think that currently available protein sequence sets are sufficiently good for the purpose as we search for statistical properties in the TMH context only.

The following datasets are used throughout this work:

Table 2.7: The experimental evidences of TOPDB. The total number of experimental evidences that contribute to ExpAll according to the TOPDB database (More information at `http://topdb.enzim.hu/?m=exptype&mid=14`). "*" refers to the total number of a subsection being larger than the total of the subcategories, likely due to lack of annotation where ambiguous literature evidence is counted toward the total, but cannot be categorised further.

Experiment		Bitopic (Single-pass)	Polytopic (Multi-pass)
Fusion	PhoA	97	2332
	PhoAS	0	90
	LacZ	20	433
	PhoALacZ	0	224
	BlaM	162	570
	BAD	0	2
	PL	0	47
	GFP	18	591
	HIS	4	2
	SplitUbiquitin	0	11
	Suc2	0	96
	Other	1	137
	Total Fusion	316*	4600*
PostTransMod	NGlyc	4634	1130
	Cman	0	6
	Phosphorylation	4	1
	Ubiquitination	47	102
	Total Post-TransMod	4685	1239
Protease	Partial Proteolysis	51	264
	Signal Peptidase	1	0
	TID	13	15
	Total Protease	64	279
	Epitope In	33	213

ExpAll

TOPDB contained 4190 manually annotated transmembrane proteins at the time of download [**Dobson2015**]. CD-HIT [**Huang2010**] identified 3857 representative sequences using sequence clusters of >90% sequence identity. This choice of similarity threshold was chosen since CD-HIT ultimately underlies the clustering behind UniRef. Unlike the other datasets, which by definition contain reasonably typical TMHs, many of the transmembrane segments annotated in TOPDB are extremely short or long and this would cause severe unrealistic hydrophobic mismatches. Especially, the short segments could be the result of miss-annotation, TMHs broken into pieces due to kinks or segments that peripherally insert only into the interface of the membrane bilayer. To remove the atypical lengths, cut-offs were set at 16 as the lower cut-off and 38 as the upper cut-off after inspecting the length histogram. We found that, for the single-pass TMHs in TOPDB, 1215 out of 1544 are within the length limits (78.7%). Among the 17141 multi-pass TMHs, we find 15563 within our global length limits (from 2205 TOPDB records corresponding to 2281 UniProt entries). This removed 1578 very short TMHs and none of the long TMHs. Our cut-off selection is very similar to the one by Baeza-Delgado *et al.* [**Baeza-Delgado2013**].

To get an idea of the taxonomical breakdown in the ExpAll dataset, the UniProt ID tags were extracted and mapped to UniProtKB. The combined dataset of multi-pass (single-pass) proteins was mapped to 1288 (1343) eukaryotic records, 404 (776) of which were human records, 926 (191) bacterial records, 46 (5) archaea records, and 14 (22) viral records.

UniHuman

This is a set of mostly human TMH-containing proteins or their close mammalian homologues. UniProtKB contains 5187 human protein records that are manually annotated with TRANSMEM regions (query = "annotation:(type:transmem) AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]" AND proteome:up000005640". To reduce sequence redundancy, these sequences were submitted to UniRef90 [**Suzek2015**]. To note, UniRef90 was chosen over UniRef50 to maintain a viable size of datasets for statistical analysis of occurrence of negatively charged

residue, which are very rare in the vicinity of TMHs. 5015 UniRef90 clusters represented the 5187 sequences. A list of sequences representing those clusters was submitted back to UniProtKB resulting and 5014 representative entries were recovered. There is a small issue in that the list of representatives from UniRef includes non-canonical isoforms, while the batch retrieve query of UniProtKB only supports complete entries, i.e. canonical isoforms. This resulted in the loss of one record at this point is due to two splice isoforms acting as representative identifiers. Of those 5014 records, 4714 were records from human entries, 197 were from mice, 94 from rats, 5 from bovine, 2 from chimps, 1 from Chinese hamsters, and 1 from pigs. Although the TMH length variations within the UniHuman dataset are much smaller than for ExpAll, we applied the same length cut-offs for the sake of comparability. Out of the 1709 single-pass cases, 1705 entered the final dataset. Of those, 1596 were from human records, 87 were from mouse, 19 were from rat, and 2 were from chimpanzee. Among the 12390 multi-pass TMHs, 12353 were included into UniHuman. The other, multi-pass record identifiers were mapped to 1789 UniProtKB entries. 1660 of these were human entries, 63 from rat, 61 from mouse, 4 from bovine, and 1 from Chinese hamster. This clustered human dataset was then queried for subcellular locations to make the UniER, UniGolgi, and UniPM datasets (detailed below).

UniER

The clustered UniHuman dataset was queried using UniProtKB for endoplasmic reticulum subcellular location (locations:(location:"Endoplasmic reticulum [SL-0095]" evidence:manual)). This returned 487 protein entries, 457 of which belonged to human, 24 to mouse and 6 to rat. 287 of these records contained sufficient annotation for orientation determination. 132 were single-pass entries of which 120 records were from humans, 11 from mouse, and 1 from rat. 155 were multi-pass entries containing 898 transmembrane helices. 144 were records from human, 8 were from mouse and 3 were from rat.

UniGolgi

The clustered human dataset was queried using UniProtKB for Golgi subcellular location (locations:(location:"Golgi apparatus [SL-0132]" evidence:manual)). This returned 323 protein entries, 301 of which belonged to human, 19 to mice, 2 to rat and 1 to pig. 269 of these records contained sufficient annotation for orientation determination. 206 were single-pass entries of which 195 records were from human, 9 from mouse, and 1 from rat. 61 were multi-pass entries containing 383 transmembrane regions. 54 were records from human, 6 were from mouse and 1 was from rat.

UniPM

The clustered human dataset was queried using UniProtKB for the cell membrane subcellular location (locations:(location:"Cell membrane [SL-0039]" evidence:manual)). This returned 1036 protein entries, 948 of which belonged to humans, 62 to mice, and 26 to rats. 920 of these records contained sufficient annotation for orientation determination. 493 were single-pass entries of which 451 records were from human, 37 from mouse, and 5 from rat. 427 were multi-pass entries containing 3079 transmembrane regions. 394 were records from human, 17 were from mouse and 16 were from rat.

UniCress

For the mouse ear cress, a representative proteome dataset was acquired with the query annotation:proteomes:(reference:yes) AND reviewed:yes AND organism:"Arabidopsis thaliana (Mouse-ear cress) [3702]" AND proteome:up000006548. This returned 3174 records in UniProtKB. UniRef90 identified 3111 clusters. 3110 of the representative sequences were mapped back to UniProtKB. Of those, 3090 were from Arabidopsis thaliana, 2 from Hornwort, 1 from cucumber, 1 from tall dodder, 1 from soybean (Glycine max), 2 from Indian wild rice, 2 from rice, 2 from garden pea, 1 from potato, 4 from spinach, 1 from Thermosynechococcus elongatus (thermophilic cyanobacteria), 1 from wheat, and 2 from maize. Of those there were 1146 with suitable TOPO_DOM annotation for topological orientation determination. 632 of those records were identified as single-pass, all of which were from Arabidopsis thaliana. 507 protein records were from multi-pass records, which contained 3823 transmembrane helices. 506 of

those records were from Arabidopsis thaliana, whilst 1 was from Thermosynechococcus elongatus.

UniFungi

For the Fungi dataset, the query "annotation:(type:transmem) taxonomy:"Fungi [4751]" AND reviewed:yes" was used. This returned 5628 records that were submitted to UniRef90. UniRef90 identified 4934 representative records, all of which were successfully mapped back to UniProtKB. Of those, 2070 had suitable annotation for orientation. 1990 records belonged to Ascomycota including 1243 Saccharomycetales. 73 were Basidomycota, and 6 were Apansporoblastina. 729 records contained a single TMH region, 702 of which belonged to Ascomycota, 26 to Basidomycota and one to Encephalitozoon cuniculi, a Microsporidium parasite. 8698 helices were contained in 1338 records of multi-pass proteins. Of these records 1285 were Ascomycota, 47 were Basidomycota, and 5 were Apansporoblastina. One TMH from UniFungi was discounted from P32897 due to an unknown position.

UniEcoli

This dataset was generated by querying UniProt with "reviewed:yes AND organism:"Escherichia coli (strain K12)[83333]"" which returned 941 hits. The hits were submitted to UniRef90, which returned 935 clusters. The representative IDs were then resubmitted to UniProtKB, all of which returned successfully. 934 were from Bacteria, whilst one were from lambdalike viruses. Of the bacterial records, 862 were from various Escherichia species of which 565 were from E. coli strain K12, 28 were from Salmonella choleraesuis, 25 were from Shigella and the rest all also fell under Gammaproteobacteria class. This dataset contains 54 single-pass proteins and 3888 helices from 529 multi-pass proteins with sufficient annotation for topological determination.

UniBacilli

The Bacilli dataset was constructed by querying UniProt for "reviewed:yes AND taxonomy:"Bacilli"". This returned 5044 records, which were submitted to UniRef90. 2,591 clusters were found in UniRef from these records. The representative IDs were

successfully resubmitted to UniProtKB. 2031 of these were of the genus Bacillales whilst 560 were also of the genus Lactobacillales. This dataset contains 124 single-pass proteins and 822 helices from 140 multi-pass proteins.

UniArch

The Archaea dataset was constructed by querying UniProt for "reviewed:yes AND taxonomy:"Archaea [2157]"". This returned 1,152 records, which were submitted to UniRef90. 1,054 clusters were found in UniRef from these records. The representative IDs were successfully resubmitted to UniProtKB. 946 records belonged to the Euyarchaeota, 101 to Thermoprotei, 4 to Thaumarchaeota, and 3 to Korarchaeum cryptofilum. This dataset contains 48 single-pass proteins and 59 multi-pass proteins containing 327 helices from 59 proteins.

2.6.2 On the determination of flanking regions for TMHs and the TMH alignment

The determination of the boundary point at the sequence between the TMH in a membrane and the sequence immersed in the cytoplasm, extracellular space, vesicular lumen, etc. is not that trivial as it initially appears. There is a lot of dynamics in the TMH positioning and the actual boundary point will be represented by various residues at different time points. Whilst the TMH core region detection from a sequence is trivial with modern software, the exact determination of TMH boundaries remains difficult since it is unclear exactly how far in or out of the membrane a given helix extends [**Ojemalm2013**]. Previous studies have dealt with this issue in various ways [**Sharpe2010**, **Baeza-Delgado2013**, **Pogozheva2013**, **White2008**].

Here in this work, we explore two boundary definitions. First, we assign TMH boundary locations as described in the respective databases. These flanks are the ones that are reported in our TMH data files that are available at the WWW-site associated with this paper. We studied flank lengths of ±5, ±10, and ±20 residues preceding and following the inside and outside TMH boundaries. In these cases, the flanks are aligned relative to the residue closest to the TMH.

In cases where the loops before and after the TMH are shorter than the predefined

flank lengths, further precautions are necessary. In the multi-pass datasets particularly (Figure 2.11 & Figure 2.4), the flanks overlap with other membrane region flanks. We explore several variants. On the one hand, we work with data files where the flank residue stretches are equally truncated so that no overlap occurs. If the loop length was uneven, the central odd residue was not included into any flank. We find surprisingly, that a large number of TMH has no or just a super-short flank, a circumstance that should disturb any statistical analysis due to the absence of objects. Therefore, we also work with alternative datasets (i) with flanks overlapping between consecutive TMH (e.g., in Table 3B; yet, it leads to some residues being counted more than one time) as well as (ii) with subsets of the data where the flanks at both sides have a defined minimal length (50% or 100% of the required flanks; unfortunately, some of them become too small for analysis).

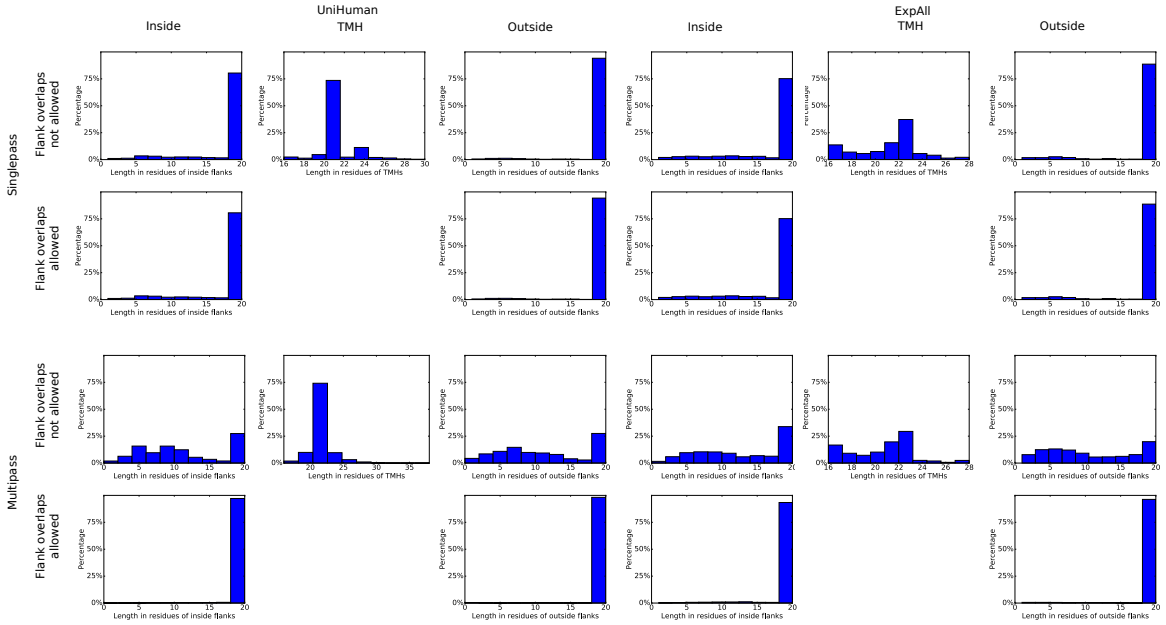

Figure 2.11: The lengths of flanks and TMHs in multi-pass and single-pass proteins in the UniHuman and ExpAll dataset. On the horizontal axis are the lengths of the TM segment regions in residues. On the vertical axis are the percentages of the population. There are three regions: the inside flank, the TMH and the outside flank. These regions are acquired according to the TMH boundary of the respective database. Where no overlap is permitted, if the flank encroaches the flank of another TMH, the flank length becomes half the number of residues in the loop region between the two features. Where they are allowed to overlap, flanking residues may include other flanks, or indeed other TMHs.

The problem of flanks overlapping does affect also some single-pass and multi-pass TMH proteins with INTRAMEM regions as described in some UniProt entries. We do not include INTRAMEM regions in the datasets as TMHs but, sometimes, the

flanking regions of TMHs were truncated to avoid overlap with INTRAMEM flanking regions (Supplementary Table S2). The identifiers affected for single-pass TMH proteins are Q01628, P13164, Q01629, Q5JRA8, A2ANU3 (UniHuman), P13164, Q01629, A2ANU3 (UniPM) and Q5JRA8 (UniER).

Table 2.8: Records with INTRAMEM and TRANSMEM flanking region overlap. The total number of TMHs from UniProt datasets with flanking region overlap between INTRAMEM and TRANSMEM regions. The number of multi-pass records that the TMHs belong to are shown in brackets.

Dataset	Flank length					
	5		10		20	
	Single-pass	Multi-pass	Single-pass	Multi-pass	Single-pass	Multi-pass
UniHuman	0	96 (80)	1	151 (90)	5	204 (96)
UniER	0	6 (6)	1	13 (8)	1	16 (8)
UniGolgi	0	1 (1)	0	2 (2)	0	4 (2)
UniPM	0	57 (46)	0	93 (51)	3	113 (52)
UniCress	0	17 (17)	0	24 (18)	0	46 (18)
UniFungi	0	0	0	0	0	0
UniBacilli	0	11 (3)	0	12 (3)	0	13 (3)
UniEcoli	0	22 (8)	0	25 (9)	0	31 (9)
UniArch	0	0	0	8 (8)	0	17 (9)

The second form of boundary point definition for flank determination was achieved with gaplessly aligning all TMHs relative to their central residue at the position equal to half the length of the TMHs at either side. Though there is some length variation among TMHs, most of them are centred around a length of 20-22 residues. In this case, flanks are the sequence extensions beyond the standardised-length 21-residues TMHs. We define the inside flanking segments as the positions -20 to -10 and the outside flanking regions to be +10 to +20 from the central TMH residue (with the label "0"). Instead of emphasising some artificially selected boundary residue, this definition allows the average TMH boundary transition to become apparent.

2.6.3 Separating simple and complex single-pass helices.

Single-pass helices from ExpAll and UniHuman datasets helices were split into two groups: simple and complex following a previously described classification [**Wong2011**, **Wong2012**] to roughly distinguish simple hydrophobic anchors and TMHs with additional structural/functional roles. Simple and complex helices were determined using TMSOC [**Wong2012**]. The complexity class is determined by calculating the hydrophobicity and sequence entropy. The resulting coordinates cluster with anchors being more hydrophobic and less complex whilst more complex and more polar TMHs are associated with non-anchorage functions. In UniHuman there were 889 simple helices and 570 complex TMHs. In ExpAll there were 769 simple helices and 570 complex helices.

2.6.4 Distribution normalisation

In this work, we have used normalisation techniques described in previous investigations as well as new approaches designed to more sensitively identify biases of rare residues. Baeza-Delgado and co-workers used LogOdds normalisation column-wise in TMH alignments. Critically, this is based on their definition of probability, which takes into account the total number of amino acids in the dataset as a denominator [**Baeza-Delgado2013**]. Since aliphatic residues such as leucine and other highly abundant slightly polar residues dominate the denominator, the distribution of the rare acidic residues will be easily lost in the "background noise" of those highly abundant residues. Pogozheva and co-workers used two approaches, (i) the total accessible surface area (ASAtotal) and (ii) total number of charged residues (N_{total}) as a denominator in their distribution normalisation [**Pogozheva2013**].

In this work, two methods for measuring residue occurrence in the TMH and its flanks were used. Similarly to previous work, we compute the occurrence of an amino acid type at a certain sequence position in a set of aligned sequences TMHs and their flanks. Following [**Sharpe2010**], the absolute relative occurrence of this amino acid type at the sequence position is then given by Equation 2.1 as:

$$p_{i,r} = \frac{a_{i,r}}{\max\limits_{r}(a_r)} \tag{2.1}$$

Here, the denominator is the maximal number of all residues in any alignment column (i.e., the number of sequences in the alignment) and, to emphasise, this will make mostly dependent on the most abundant residue types. This type of normalisation reveals the most preferred residue types at given sequence positions.

Our second normalisation method is independent of the abundance of any amino acid types other than the studied one; it answers the question: "If there is a residue of type in the TMH-containing segment, where would it most likely be?" This relative occurrence calculated in Equation 2.2 as:

$$q_{i,r} = \frac{100 \cdot a_{i,r}}{a_i} \tag{2.2}$$

The value a_i is the total abundance of residues of just amino acid type i in a given alignment of TMH-containing segments (i.e., in the TMH together with its two adjoining flanks summed over all cases of TMHs in the given dataset). Peaks in $q_{i,r}$ as function of r reveal the preferred positions of residues of type i. The difference in $q_{i,r}$ and $p_{i,r}$ normalisation is visualised in Figure 2.12.

2.6.5 Hydrophobicity calculations

Hydrophobicity profiles were calculated using the Kyte & Doolittle hydrophobicity scale [**Kyte1982**] and validated with the Eisenberg scale [**Eisenberg1984**], the Hessa biological scale [**Hessa2005**], and the White and Wimley whole residue scale [**White1999**](Figure 2.7). The hydrophobicity profile uses un-weighted windowing of the residue hydrophobicity scores from end to end of the TMD slice. Three residues were used as full window lengths and partial windows were permitted.

2.6.6 Normalised net charge calculations

Charge was calculated at each position by scanning through each position of the transmembrane helices and flanking regions and subtracting one from the position if an acidic residue (D or E) was present, or adding one if a positively charged residue (K or R) was present. The accumulative net-charge was then divided by the total number of transmembrane helices that were used in calculating the accumulative net-charge. Thus, the charge distribution is calculated by:

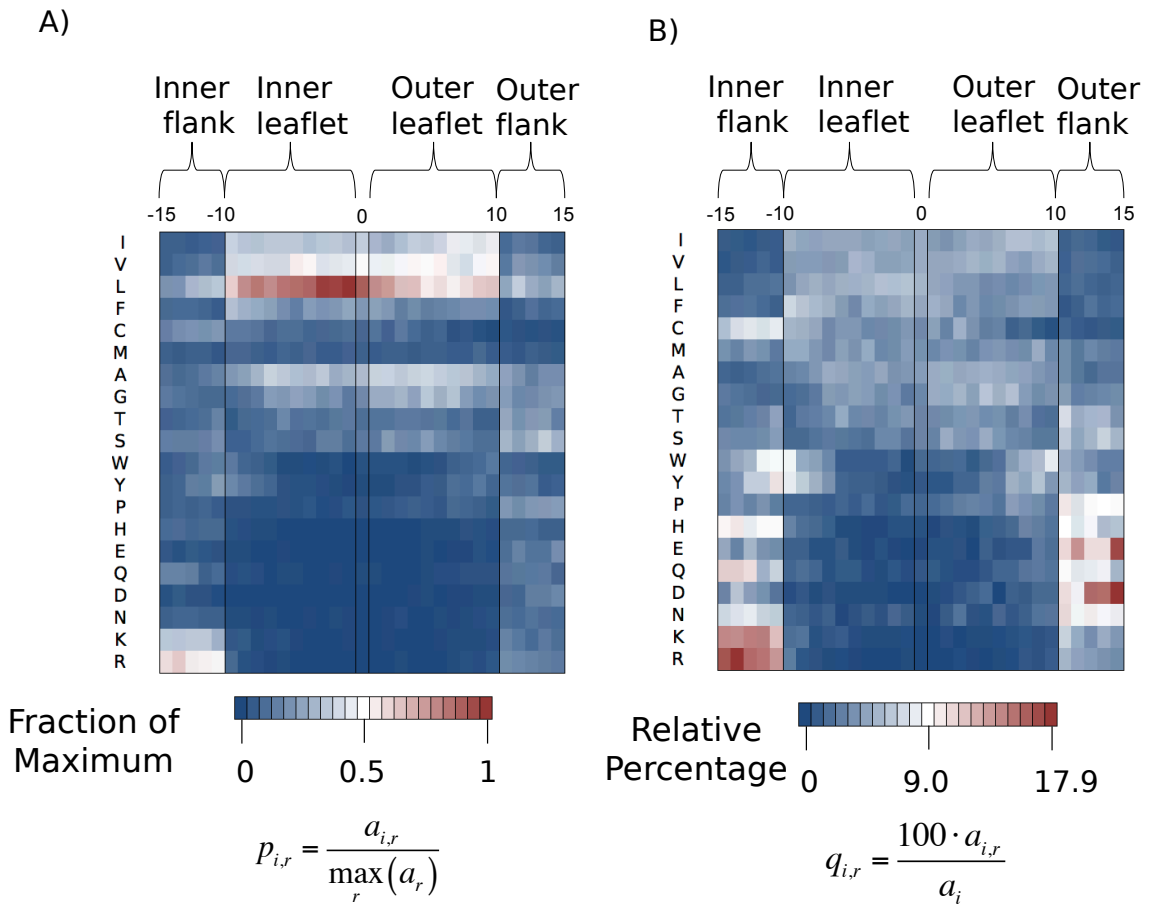

Figure 2.12: Relative percentage heatmaps from the predictive datasets calculated by fractions of the absolute maximum and by the relative percentage of a given amino acid type. The residue position aligned to the centre of the TMH is on the horizontal axis, and the residue type is on the vertical axis. Amino acid types are listed in order of decreasing hydrophobicity according to the Kyte and Doolittle scale [**Kyte1982**]. The flank lengths in the TMH segments were restricted to up to ±5 residues. The scales for each heatmap are shown beneath the respective subfigure. All TMHs and flank lengths are from the UniHuman dataset. (A) The heatmap has been coloured according to a scale that uses column-wise normalisations used in previous studies [**Sharpe2010**]. See Equation 2.1. As an illustrative example, we show how the value for E at position ±12 is obtained. There are in total 91/22 Es at these positions in 1705 sequences; thus, the represented value is 0.013 at 12 and 0.053 at 12. Note that L is clearly a hotspot as well as trends for other hydrophobic residues, I and V, as is to be expected. A positive inside effect can also be seen. (B) The heatmap has been coloured according to the relative percentage of each amino acid type (Equation 2.2). Here, 91/22 Es at position ±12 are compared with 615 Es seen within the flanks and the TMH section itself amongst all sequences in the alignment. So, the expectation of an E at position ±12 if there is any E in the TMH + flanks region at all is 0.036 at 12 and 0.148 at position 12. With this type of normalisation, not surprisingly, we see the positive-inside rule is hotter than in subfigure A. There are also hotspots in the flanks for the negatively charged residues on the outside flank. The leucine hotspot is no longer very pronounced, as the leucines are quite evenly spread over many positions.

$$c_r = \frac{(a_{K,r} + a_{R,r}) - (a_{D,r} + a_{E,r})}{N} \tag{2.3}$$

2.6.7 Statistics

The inside/outside bias of negative residues was quantified by computing the independent KW and the 2-sample t-test statistical method from the Python scipy stat package v0.15 python package [**VanderWalt2011**]. This test answers the question whether two means are actually different in the statistical sense. For the leucine residues, each TMH region was divided into two sections, representing the inner and outer leaflets (Table 2.4). For the hydrophobicity plot, 3 window values of hydrophobicity were taken for each TMH at each position. The statistical analyses were separately performed for single-pass and multi-pass transmembrane proteins. At each position, the two groups were compared using the KW test.

The zero hypothesis of homogeneity of two distributions was examined with the KS, the KW and the χ^2 statistical tests. To note, the KS test scrutinises for significant maximal absolute differences between distribution curves; the KW test is after skews between distributions and the χ^2 statistical test checks the average difference between distributions. As the statistical significance value ("Pvalue") is a strong function of N, the total amount of data used in the statistical test, we rely on the (absolute) Bahadur slope (B) as a measure of distance between two distributions [**Bahadur1967**, **Bahadur1971**]:

$$B = \frac{\ln(P\ value)}{N} \tag{2.4}$$

The larger the absolute Bahadur slope, the greater the difference between the two distributions.

Chapter 3

Tail-Anchored Protein Datasets

3.1 Abstract

Tail Anchor (TA) proteins are a functionally diverse group of post translationally inserted membrane proteins. The TMH and flanking regions of the TA protein contain sufficient information for subcellular targetting. Here, we built datasets based on sequence definitions of plausible TA proteins using the Uniprot database. We show that any statistical differences between the hydrophobicity of the TMH between TA proteins belonging to mammalian, plant, or yeast organsisms are unobservably small. Yet, hydrophobicity of glstmhs from different subcellular locations is, on average, different between the mitochondria and other membranes along the secretory pathway. Notably, in the case of mitochondria, this appears to be a difference in the hydrophobic residue preference of alanine instead of leucine or isoleucine, and does not seem to be an increase of intramembrane polar residues. Whereas in the ER, Golgi, and PM there are positively charged residues inside, and negatively charged residues outside the cytoplasm, we identify a charge skew reversal of the positive-inside and negative-outside rules in the mitochondrial TA proteins to negative-inside and positive-outside. Whether these adaptaions are the result of membrane environment adaptations or biological features useful for the biogenesis and accurate localisation of the protein remains unclear. Furthermore, structural homology modelling of the spontaneously inserting TMHs of PTP1b and cytochrome b5 reveal that TA proteins may gain their ability to integrate into the membrane unaided due to a strip of conserved relatively

polar residues and strong flanking charge that would allow effective membrane coupling.

3.2 Introduction

TA proteins are defined by their single carboxy-terminal TMH with a cytosolic facing amino-terminus and are a topologically distinct class of intracellular proteins. The integration of TA proteins into the membrane is post-translational rather than co-translational; the ribosome is not in complex with the membrane insertion machinery. TA proteins are involved in a range of key cellular functions such as translocation [**Osborne2005**] such as Sec61β and Sec61γ as well as the Bcl-2 apoptotic protein family [**Hockenbery1990**]. Additionally, within the TA class of proteins are a set of vesicle fusion proteins called Soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) proteins [**Ungar2003**], which contain typically hydrophobic TMHs [**Kalbfleisch2007**]. 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 [**Allen2013**, **Nordlund2014**].

The TA protein's TMH is unusual in that it is both the anchor and the targeting factor for the ER [**Kutay1993**]. Furthermore, the hydrophobicity appears to be a determining factor in the delivery pathway that TA proteins use for insertion [**Rabu2008**, **Rabu2009**], for which there is evidence demonstrating that are several mechanisms of translocation [**Rabu2009**, **Johnson2013**](Figure 3.1).

TA proteins have several pathways for biogenesis in the ER membrane. TA proteins were originally thought to be inserted into the membrane via different machinery than the co-translational machinery, but unexpectedly SRP was found to be a factor for post-translational targeting confirmed by both cross-linking studies [**Abell2004**] and an *in vitro* pull-down experiment [**Leznicki2010**]. SRP would deliver the TA protein to the membrane-bound Signal Recognition Particle Receptor (SR) in association with a highly conserved Sec translocon. Further cross-linking experiments suggested Sec61 is also involved during TA protein membrane insertion [**Abell2003**]. Previous studies had shown the Sec61 translocon is not necessary for TA protein membrane integration

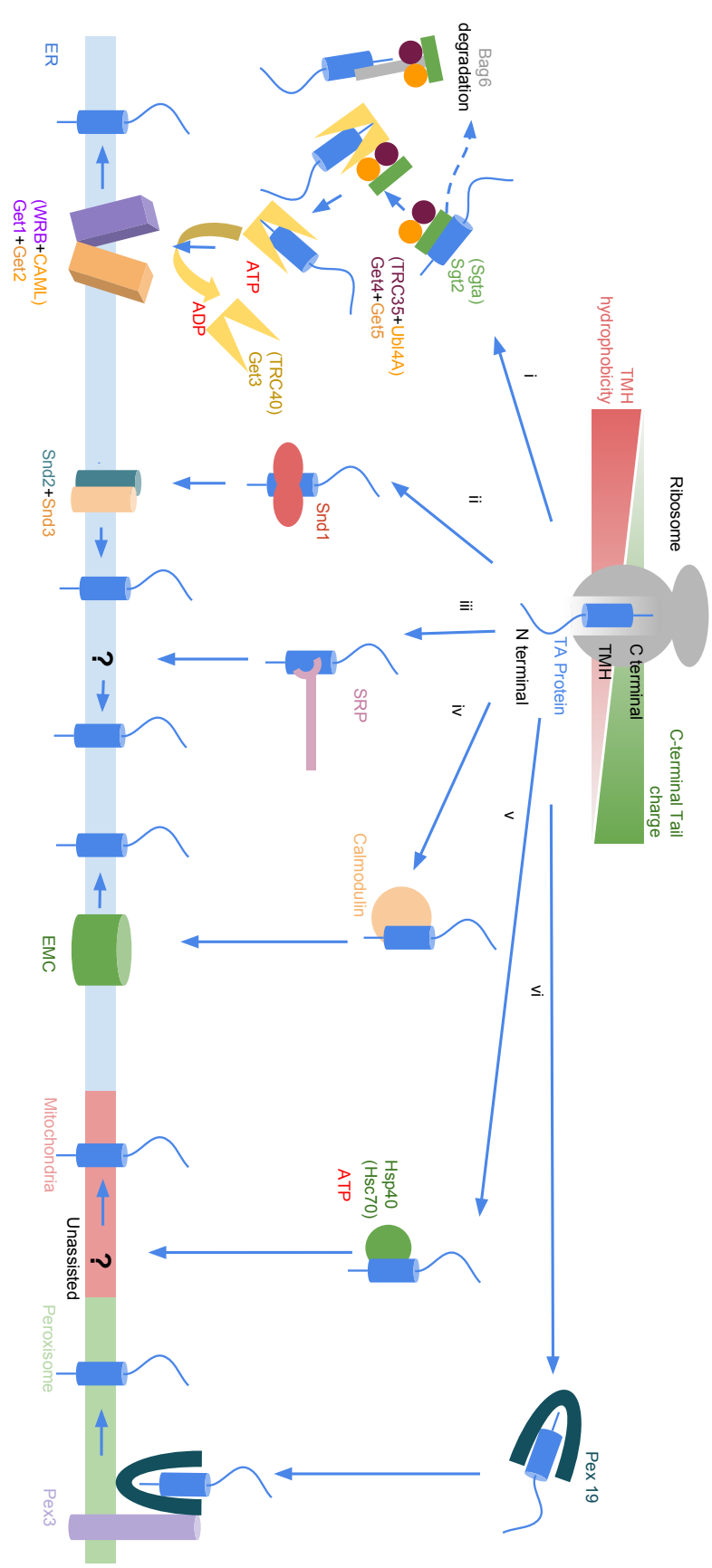

Figure 3.1: An overview of the biogenesis of tail-anchored proteins. (i) The intensively studied Get (yeast) or TRC40 (mammalian) pathway can target either for membrane integration or degradation of the TA protein. (ii) In yeast, a novel mechanism was identified in which Snd1 binds to the folded TA protein and delivers it to the membrane-bound Snd2 and Snd3 complex. (iii) Signal Recognition Particle (SRP) of the co-translational insertion mechanism has been shown to be able to integrate TA proteins. Sec61s role in this is disputed, and it is likely other proteins are involved in membrane partitioning instead. (iv) A recently identified insertase, the EMC, can integrate TA proteins with relatively polar TMHs into the ER. (v) A handful of TA proteins with relatively polar TMH regions have been observed spontaneously integrating into the membrane using Hsp40 and Hsc70 as chaperones. This system may be employed for mitochondrial localisation. (vi) Peroxisomal proteins with an abundance of charge in the tail region are chaperoned by Pex19 into association with the membrane-bound Pex3.

by biochemical reconstitution experiments [**Kutay1993**] and conditional mutants in yeast [**Steel2002**, **Yabal2003**]. So whilst it is hard to determine if Sec61 can be part of the post-translational pathway, we can conlude that it is certainly not essential, indeed almost no observable impact on biogenesis is had when it is removed [**Kutay1993**, **Steel2002**, **Yabal2003**]. Nevertheless, this suggests the possibility of at least one insertion mechanism that is related to the co-translational method of insertion. Most likely is that SRP binds to the TA protein after it is released from the ribosome and chaperones it until it is close enough to the established post translational machinery [**Casson2017**] (Figure 3.1(iii)).

A second redundant system is also known to be involved in TA protein biogenesis and is referred to as the TRC40 (also known as Asna1) pathway in mammals (Figure 3.1(i)). A conserved homologue was found in *Saccharomyces cerevisiae*, Get3 [**Schuldiner2008**], and in yeast, this mechanism is generally referred to as the Get pathway. Unlike co-translational insertion, the post-translational proteins do not couple with the ribosome, so the TA protein must be exposed to the cytosolic environment for at least some time [**Guna2018**]. At some point after the TA protein emerges from the ribosomal exit tunnel, the TA protein TMH associates with Sgt2. An *in vitro* assay revealed that Sgt2 associates with Get5 [**Wang2010**] as part of a dimerised Get4 and Get5 complex (two copies of each)[**Chang2010**, **Chang2012**, **Chartron2010**, **Chartron2012**]. At this point Sgt2 either associates with preferential Get3 which targets the TA protein for ER membrane biogenesis or if there are excess TA proteins Sgt2 also associates with Bag6 which targets the TA protein for degradation [**Shao2017**]. This "race" between Bag6 and Get3 ensures a level of quality control within the system. Assuming the TA protein is not targetted for degradation, Get3 associates first with this complex via an interaction with the N-terminal of Get4 [**Wang2010**]. A dimerised ADP-bound Get3 [**Mateja2009**, **Hu2009**, **Bozkurt2009**, **Suloway2009**, **Yamagata2010**] (TRC40) associates with and shields the C-terminal region of the TA protein [**Stefanovic2007**, **Schuldiner2008**, **Favaloro2008**]. This shielding may be especially important since Get3 is involved in the folding of any nascent TA proteins, which would be an unviable hydrophobic in the cytosol [**Jonikas2009**]. Fluorescence studies revealed that tagged Get3 appears at both the cytosol and the ER membrane so apparently shuttles the TA protein between the transmembrane complex of Get1

and Get 2 (WRB and CAML in mammalian cells), that contains cytosolic domains that receive Get3, and Get4 Get5 Sgt2 complex [**Huh2003**, **Zalisko2017**]. Yet it is an interesting note that a single molecule fluorescence study revealed that the minimum machinery required for TA protein insertion from this system is a Get1 and Get2 heterodimer [**Zalisko2017**]. The Get pathway exclusively delivers TA proteins to the ER membrane, and indeed has been recently shown to be responsible for some of the mislocalisation of mitochondrial TA proteins during overexpression or signal masking to the ER [**Vitali2018**]. The significance of this is that the Get machinery can recognise and tolerate integration of non-ER proteins. Yet there is also evidence that the deep groove of Get3 [**Mariappan2011**, **Stefer2011**] predisposes it do only effectively integrating the more hydrophobic TMHs of TA proteins [**Wang2010**, **Rao2016**]. As an example, increasing the hydrophobicity of the TA protein squaline synthase in a TRC40 inhibited system reduced the biogenesis of the protein, where the wild type was unaffected by TRC40 inhibition [**Guna2017**]. Around a half of TA proteins are estiamted to not use the TRC pathway.

Redundancy of the Get/TRC40 pathway and SRP pathway may be explained in part by a novel SRP and Get independent pathway. This pathway utilises the Snd protein pathway and was discovered in yeast [**Aviram2016**] (Figure 3.1 (ii)). Snd1 binds to the TA protein after it exits the ribosome and delivers it to the Snd2 and Snd3 membrane-bound complex which integrates the TA protein into the membrane. So far only the homologue of Snd2 has been identified (hSnd2) with relatively low sequence identity [**Hadenteufel2017**]. However, it is suspected that functional mammalian homologues exist for Snd1 and Snd3 also, albeit with low sequence similarity.

Even more recently identified as a TA protein insertase was the ER Membrane protein Complex (EMC) [**Guna2017**] (Figure 3.1iv). In the interaction study, squaline synthase did not effectively corsslink with any TRC pathway machinery. Calmodulin sufficiently prevented aggregation and acted as a chaperone in ER microsomes in a chaperone free *E. coli* translation system with purified translation factors. In this system calmodulin acted similarly to SGTA, and although in native cytosol calmodulin was prefered, SGTA could also be used by the protein. By contrast VAMP2, a TA protein known to interact with TRC40 [**Shao2017**], was unable to insert in this system. Abolition by knock outs of the EMC components greatly reduced the insertion

potential of squaline synthase and five mutants thereof and six other TRC40 independent proteins in ER membranes from semi-permeabilised cultured cells, but not that of VAMP2. Insertion of Sec61β was partially dependent on both systems, perhaps indicating the midway point between the two [**Guna2017**] (Figure 3.1).

In the absence of the mitochondrial TIM-TOM insertion machinery and Get machinery, Hsp40 and Hsc70 chaperones along with ATP are also sufficient for enough biogenesis of TA proteins for viable cell growth [**Rabu2008**, **Rabu2009**, **Ngosuwan2003**, **Colombo2009**, **Kemper2008**, **Meineke2008**, **Setoguchi2006**]. In biological systems, this is possibly used for mitochondrial delivery [**Kemper2008**] (Figure 3.1 (v)). Chimeric synaptobrevin, one of the first identified SNARE proteins, is capable of this spontaneous insertion if the tail anchor domain is replaced by the TM domains belonging to a protein of known spontaneously inserting domains [**Nordlund2014**]. Molecular dynamics simulations showed that direct insertion TMHs thermodynamically mimics the energies of TMHs integrated by the translocon [**Ulmschneider2014**] so in theory, no integration machinery is strictly necessary if the TMH can "correctly" interact with the membrane interface. Further, it was revealed that scrambling the TMH sequence, but maintaining hydrophobicity, reduced the insertion potential of spontaneously inserting TMHs [**Brambillasca2006**]. This phenomenon cannot, therefore, be explained entirely by the marginal hydrophobicity of the TMH.

The few peroxisomal TA proteins first associate with Pex19 which forms a complex with the membrane-anchored Pex3 protein from which the TA protein is integrated into the membrane [**Chen2014**, **Yagita2013**, **Costello2017**](Figure 3.1 (vi)).

The newly discovered EMC machinery also was shown to be involved in the integration of TA proteins with more hydrophobic TMHs [**Guna2017**].

Given a "choice", it is speculated that hydrophobicity determines the integration pathway since Sec61β has a hydrophobic TMH and is targetted via the SRP pathway, whereas marginally hydrophobic TA proteins like cytochrome b5 and PTP1b can spontaneously insert *in vitro* and biologically only rely on Hsp70 and Hsc40. Altering the hydrophobicity, at least in the case of the spontaneously inserting PTP1b, also determinants the localisation of the TA protein to either the mitochondrial membrane or the ER membrane, or rather a more hydrophobic TA protein TMH is less likely to

localise to the mitochondrial membrane [**Fueller2015**]. Broader analysis has shown that hydrophobicity [**White1999**] stratified by TM tendency score [**Zhao2006**] can distinguish between the ER and mitochondrial localised TA proteins [**Guna2018**]. However, the tremendous diversity and known biogenesis redundancy of these proteins may mean that no single factor applied en masse may be able to distinguish the TMH recognition factors and investigation into this area is becoming increasingly complex [**Guna2018**].

By regenerating a list of likely TMHs [**Kalbfleisch2007**] and using a manually curated list of TA proteins [**TheUniProtConsortium2014**], this investigation aims to find relationships between biochemical factors and a disposition to a certain insertion mechanism and terminal localisations. Here, we also present evidence for a conserved polar strip along the spontaneously inserting TA protein TMHs, which may be the key to the initial interaction of these TMHs with the membrane interface.

3.3 Methods

3.3.1 Building a List of Tail-Anchors

Steps carried out by Kalbfleisch *et al.* (Traffic 8: 16871694) to generate a list of all TA proteins in the human proteome [**Kalbfleisch2007**], were recreated using up to date tools and applied to other model representitive species. Whilst their study focused on the human proteome, here we take into account the entire TrEMBL and SwissProt database and then stratify the datasets by the organism at the end of the pipeline (Figure 3.2).

SwissProt Tail Anchored Dataset According to Filters

There were 557012 protein records downloaded from SwissProt via UniProt [**TheUniProtConsortium2014**] (downloaded 24–04–2018). 106149 TMHs (`TRANSMEM` annotation) were found between 76953 records (`annotation:(type:transmem)ANDreviewed:no`). This keyword is contained in a record according to either experimental evidence [**TheUniProtConsortium2014**] or a conservative meta-analysis of TMH prediction using TMHMM [**Krogh2001**], Memsat [**Jones2007**],

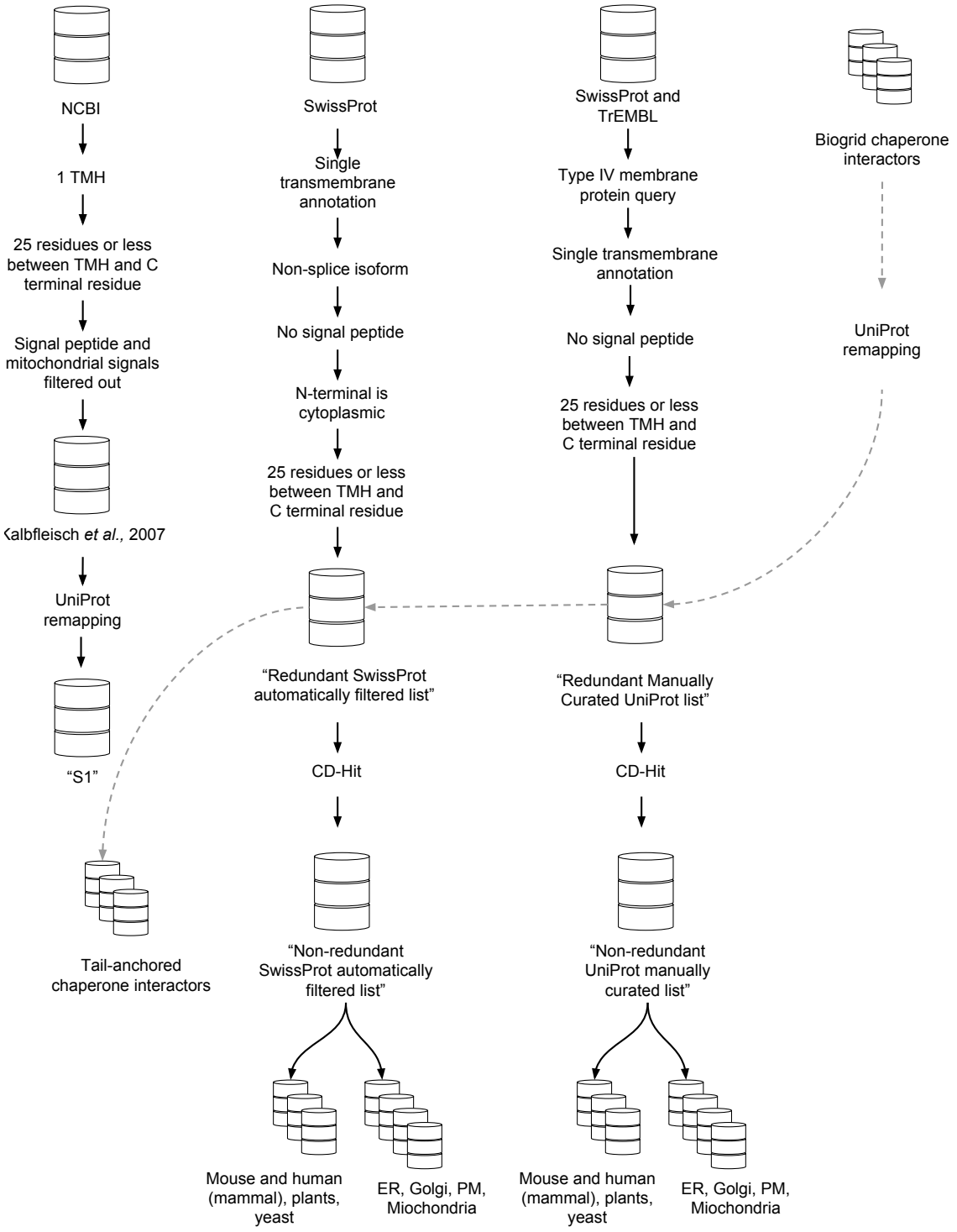

Figure 3.2: The sources, methods, and filters applied to the sequences in the datasets. From top to bottom are the sources of the sequences and the filters and methods applied to each of the datasets of sequences. The database symbol is used to denote when the dataset was used to capture results and is available as supplementary material. For the dataset size and more information, see the methods section text.

Phobius [**Kall2004**, **Kall2007**] and the hydrophobic moment plot method of Eisenberg and co-workers [**Eisenberg1984**]. 11141 of those records had only a single TMH. 11110 of those TMHs were within the length thresholds of 16 to 30

residues (None of those had the annotation for splice isoforms according to `NON_TER` annotation). 5548 of those had had no SP annotation (`SIGNAL`). 4332 of those had annotation (based on `TOPO_DOM` annotation) that the N terminal was cytoplasmic. 615 of those had the TMH within 25 residues of the C terminal, the same threshold used by Kalbfleisch *et al.,* [**Kalbfleisch2007**]. Running CD-Hit 4.5.3 on the WebMGA web-server [**Huang2010**, **Wu2011**] at 90% identical sequence at 90% coverage thresholds resulted in 443 representative proteins. This threshold was chosen as a compromise between avoiding over-representation of a certain protein and maintaining a viable sample size.

From this representative list, 46 were Archaeal, 66 were bacterial, and 320 were eukaryotic and 11 came from dsDNA viruses. When counting proteomes with greater than 20 records, 49 belonged to the *A. thaliana* proteome, 48 to Mouse, 46 to the human proteome, 24 to *S.cerevisiae*.

65 were annotated under the mitochondrion location (query `locations:(location:"Mitochondrion[SL-0173]")`), 157 in the PM (query `locations:(location:"Cellmembrane[SL-0039]")`, 82 in the Golgi (query `locations:(location:"Golgiapparatus[SL-0132]")`), and 98 in the ER (query `locations:(location:"Endoplasmicreticulum[SL-0095]"`). Only 16 records were found for the peroxisome (query `locations:(location:"Endoplasmicreticulum[SL-0204]"`) which is not high enough a sample size for accurate statistical analysis.

TrEMBL Tail Anchored Dataset According to Filters

111425234 records were stored in the TrEMBL database at the time of download (downloaded 25–04–2018). 22107826 of those contained `TRANSMEM` annotation (`annotation:(type:transmem)ANDreviewed:no`). 18053 of these were single-pass proteins. All of these were within the length restrictions of between 16 and 35 residues for the TMH region. 17973 of those did not contain a signal sequence when looking for `SIGNAL` annotation. 5157 of those contained a cytoplasmically located N terminal according to `TOPO_DOM` annotation. 155 records had a TMH within 25 residues of the C terminal residue. When considering which species these records come from, no more than 1 record belonged to any given species. To avoid representing a well-annotated SwissProt record that includes species annotation by a poorly annotated

TrEMBL record without species annotation, these TrEMBL records were omitted from the sequence redundancy protocol and further analysis.

UniProt Curated List

A query for `locations:(location:"Single\--passtypeIVmembraneprotein[SL-9908]")` was used in UniProt which returned 2633 UniProtKB IDs; 463 SwissProt results and 2170 TrEMBL results. Type IV anchors are sometimes split into two topological groups; A (A cytosolic facing N terminal domain) and B (The N terminal is targeted to the lumen), however, the UniProt nomenclature is strictly N terminal being cytosolic. This manually created list contained some TA proteins that didn't exactly fit the generally accepted definition of a TA protein and were excluded from further analysis. These could be examples of misannotation in the databases, or exceptional TA proteins that behave as post-translationally inserted TA proteins despite not matching the exact criteria. 101 exceeded the TA length restrictions of 25 residues between the TMH and the terminal residue. 8 contained annotation for `SIGNAL`, indicating an SP, inconsistent with the TA protein definition. 20 were multipass proteins. A full list of which records exceeded these limits and by how much is included in the supplementary files. Running these records through CD-HIT at 90% redundancy yielded 956 clusters; 269 SwissProt records and 687 TrEMBL records [**Huang2010**, **Wu2011**]. No further filters were applied to this list. Proteomes represented by more than 20 records include *A. thaliana* (53 records), Humans (30), Mouse (30), and *S. cerevisiae* (27).

426 were annotated under the mitochondrion location (query `locations:(location:"Mitochondrion[SL-0173]")`) 47 from SwissProt and 379 automatically assigned in TrEMBL. 397 in the ER (query `locations:(location:"Endoplasmicreticulum[SL-0095]")`), 88 from SwissProt and 308 automatically annotated in TrEMBL. 1 TrEMBL record (UniProt ID A0A1E5RT24) in the ER set contained an "X" residue in the C terminal flank and was omitted from the analyses. Two subcellular location datasets had no automatically ascribed records and only contained manually annotated SwissProt records; 31 in the PM (query `locations:(location:"Cellmembrane[SL-0039]")`, and 83 in the Golgi (query `locations:(location:"Golgiapparatus[SL-0132]")`). There were only 8 TA proteins located

in the peroxisome (`locations:(location:"Golgiapparatus[SL-0204]")`), making them an unsuitable dataset for statistcal analysis.

Remapping The Previous Dataset

189 of the 411 proteins from the previous Kalbfleisch *et al.,* 2007 study [**Kalbfleisch2007**] were successfully mapped to 222 UniProtKB IDs using the UniProt mapping tools with the RefSeq Protein to UniProtKB option [**TheUniProtConsortium2014**].

Tail Anchor Protein Chaperone Interactors

As discussed in the introduction text, there is evidence surrounding the TMH biochemical factors that determine which chaperone will interact with a given TA protein. To gain a more quantitative understanding of the relationship between the TA protein TMH and the potential chaperones, it would be ideal to have a large TA dataset stratified by known chaperone interactions.

Known interactor lists from BioGrid from the chaperones were checked against the SwissProt automatically filtered and the UniProt curated TA protein datasets. There were 91 interaction pairs for Hsp40 (Biogrid ID 119699) which mapped to 206 UniProt records. Hsc70 (Biogrid ID 109544) had 534 interaction pairs 61 of which were mapped to 91 UniProt records. Hsp40 and Hsc70 both returned 0 record hits after filtering the UniProt records IDs through the UniProt manually curated list and the automatically generated SwissProt list both before redundancy removal.

Snd1 (Biogrid ID 32240) had 237 interaction pairs which mapped to 239 records. Snd1 returned 15 hits when filtering it through the TA datasets.

SGT2 (BioGrid ID 34410) had 260 BioGrid interactor ids which mapped to 264 UniProt records. SGTA (BioGrid ID 112347) had 155 interactor ids from BioGrid, 153 of which were mapped to 274 records. SGT2 and SGTA returned 14 and 5 hits respectively.

50 BioGrid ids from TRC40 (Biogrid ID 106931) interaction pairs were mapped to 90 UniProt records. Get3 (BioGrid ID 31962) had 456 Biogrid interactor ids, which mapped to 465 UniProt records. After filtering those records through the TA anchor datasets TRC40 and Get3 returned 7 and 22 hits respectively.

In yeast, Pex19 (BioGrid ID 31994) contained 466 interactor pairs which mapped to 384 UniProt IDs. The human Pex19 (BioGrid ID 111782) contained 230 interaction pairs which successfully mapped to 218 UniProt records. When the TA list filters were applied, 2 human and 7 yeast records were found.

For SRP54, ideally the plant version of the protein was required, for which there is more precedent for post-translational protein interaction and biogenesis into the chloroplasts [**Abell2004**], however, of the 3 plant SRPs available on Biogrid, between them only 10 interactor pairs were available, none of which were common with our TA lists. The human SRP54 (Biogrid ID 112607) had 37 interactors which mapped to 85 UniProt IDs, but again, none of which were in our TA lists. On the other hand, the yeast SRP54 (BioGrid ID 36258) had 270 interactors which mapped to 273 UniProt records. 4 of those were found in our TA lists.

3.3.2 Calculating Hydrophobicity

Windowed hydrophobicity was calculated using a window length of 5 residues, and half windows were permitted. Average hydrophobicity takes the total of the raw amino acid hydrophobicity values and divides them by the number of amino acids in the slice. Unless explicitly stated, values reported in the results are based on the Kyte & Doolittle scale [**Kyte1982**] which is based on the water–vapour transfer free energy and the interior-exterior distribution of individual amino acids.

3.3.3 Calculating Sequence Information Entropy

Information entropy is essentially an estimate of the linguistic entropy of a string. In the context of biology, it can be thought of as an estimation of the non-randomness of a sequence. Sequence complexity can be used to analyse DNA sequences [**Pinho2013**, **Oliver1993**, **Troyanskaya2002**] and is a component of the TMSOC z-score which can predict function beyond anchoring of a TMH; an increase in complexity is associated with increased likelihood of function [**Wong2011**, **Wong2012**, **Baker2017**]. Here we focus on the analysis of the complexity of a string of characters in protein sequences.

Broadly speaking, the information theory entropy of a linguistic string can be defined as in equation 3.1, and we treat the protein sequence TMH as a string with or

without its flanking regions.

$$H(S) = -\sum_{i=1}^{n} p_i \log_s(p_i) \tag{3.1}$$

Where H is the entropy of a sequence S, and p_i is the probability (p) of a character i through each position (n) in S. This allows us to quantify the average relative information density held within a string of information [**Shannon1948**].

3.3.4 Statistics

The null hypothesis of homogeneity of two distributions was examined with the Kolmogorov Smirnov, the Kruskal-Wallis, and the 2-sampled Student's T-test statistical tests. These tests were all ran through the Python SciPy stat v0.17 package [**VanderWalt2011**]. To note, the KS test scrutinises for significant maximal absolute differences between distribution curves; the KW test is after skews between distributions and the student T-test statistical test checks the average difference between distributions.

Since the Pvalue is a product of a fraction of test statistics obtained from a permutated set of the samples, it exponentially increases as N increases; the P-value is a strong function of N. We rely on the Bahadur slope (B) as a measure of distance between two distributions [**Bahadur1967**, **Bahadur1971**, **Sunyaev1998**, **Baker2017**]. A larger Bahadur slope shows a greater difference between the two distributions.

$$B = \frac{|\ln(P\ value)|}{N} \tag{3.2}$$

In the heatmaps (Figure 3.6, Figure 3.7), the relative percentage normalisation was used rather than a fraction of the absolute value. This aims to answer the question of "if we have a certain amino acid, which position is it likely to be in?" and are able to sensitively identify clusters of skewed preference [**Baker2017**].

$$q_{i,r} = \frac{100 \cdot a_{i,r}}{a_i} \tag{3.3}$$

a_i is the total abundance of residues of a specific amino acid type (i) of an aligned set of TMH-containing segments. Peaks in $q_{i,r}$ as a function of r (the position index)

reveal the preferred positions of residues of type i.

3.3.5 Modelling Cytochrome b5 and PTP1b

The HHpred web server was used to query homologues of and model templates for Cytochrome b5 (UniProt accession code P00167) and PTP1b (UniProt accession code P18031) [**Soding2005**]. Homologues were queried using three iterations of HHblitscd against the sequence database version uniprot20_2016_02 to generate the query Hidden Markov Model. The choice of templates was driven by the quality and coverage of the alignments and of the quality of the models that resulted. For cytochrome b5, a multiple alignment was generated from PDB accession codes 2M33, 2KEO, 3X34, 1MJ4, 1MJ4, 2IBJ covering the globular domain, and PDB accession codes 5NAO, 5DOQ, 5NAM, and 2MMU covering the TMH. Modeller was run from within the HHPRED server to generate the homology model [**Eswar2007**, **Webb2016**]. The model was confirmed to be of high quality using ProSA (Z-Score: -4.61) [**Wiederstein2007**], Ramachandran plot on the RAMPAGE web server (98% allowed residues, including all TMH residues) [**Lovell2003**].

The coverage and alignment quality of PTP1b was, however, not as good quality. Although UniProt holds 145 associated PDB structures for PTP1b, these structures cover at least some part of the globular domains of the protein. There are no PDB structures for the TM domain or the nearby flanking regions. Instead of a global protein model, only the Trans-membrane Domain (TMD) was modelled using a homology model derived from a single sequence alignment of 5NAO based on the TMH±6 (the length of the C-terminal tail) residues of PTP1b.

Both these TMH regions were verified by a consensus of sequence TMH predictions (Scampi seq [**Bernsel2008**], Phobius [**Kall2004**], TMHMM [**Krogh2001**], MEMSAT3 [**Jones2007**], TMpred [**Hofmann1993**], HMMTOP [**Tusnady2001**], DAS–TMfilter [**Cserzo2004**], MINNOU [**Cao2006**], OCTOPUS [**Viklund2008**], PRODIV [**Viklund2004**], PRO–S [**Viklund2004**], S–TMHMM [**Viklund2004**], and proteus [**Montgomerie2008**]). However it should be noted that not all these predictions unanimously agreed. For PTP1b, several methods identified more than 1 TMH (HMMTOP, TMPred) whilst Memsat identified a short TMH in a completely different region for the TMH (35A to 45R). Besides that, only S–TMHMM and PRO agreed on the

exact start and stop positions, and it so happens that these are also the majority consensus positions [**Kurowski2003**] (409F to 429F).

APBS as a PyMol plugin was used to map the electrostatic surface of the model [**Baker2001**]. Consurf [**Ashkenazy2010**] was used to map the conservation scores based on 5 iterations of PSI-BLAST [**Altschul1997**] with an E-value cut-off of 0.0001. Hydrophobicity was mapped according to the Eisenberg aggregated hydrophicity scale [**Eisenberg1984**] using a script accessed at `https://pymolwiki.org/index.php/Color_h`.

3.4 Results And Discussion

3.4.1 A Comparison Of Up-To-Date Tail-Anchored Protein Datasets

Here, we use two sources for TA protein datasets. One dataset is based on a previous method [**Kalbfleisch2007**] to obtain TA datasets and consists of 9296 TMH residues (13279 including up to ±5 flanking residues) from 443 SwissProt entries with 90% redundancy removal. Another dataset contains the UniProt curated set of Type IV membrane proteins again with 90% redundancy removal. This dataset contains 20528 TMH residues (27950 including up to ±5 flanking residues) from 956 UniProt protein records.

In order to get an understanding of the consistency of the datasets, before removing redundant proteins, we compared these two datasets to a dataset remapped set of proteins from a previous 2007 method [**Kalbfleisch2007**]. The S1 dataset was built with an aim to gather TA proteins in the human genome from the NCBI. The greatest source of uncertainty here is that the original S1 list includes 411 records, however, only 222 of these were successfully mapped to the UniProt dataset. This figure is closer to the 202 proteins from the original S1 list that excluded proteins that were either hypothetical or splice isoforms. That being said, this mapping step prevents us from directly comparing the entire original S1 dataset. We compared the up-to-date datasets to S1 to see how many records are shared, how many are now obsolete, and how many are unique.

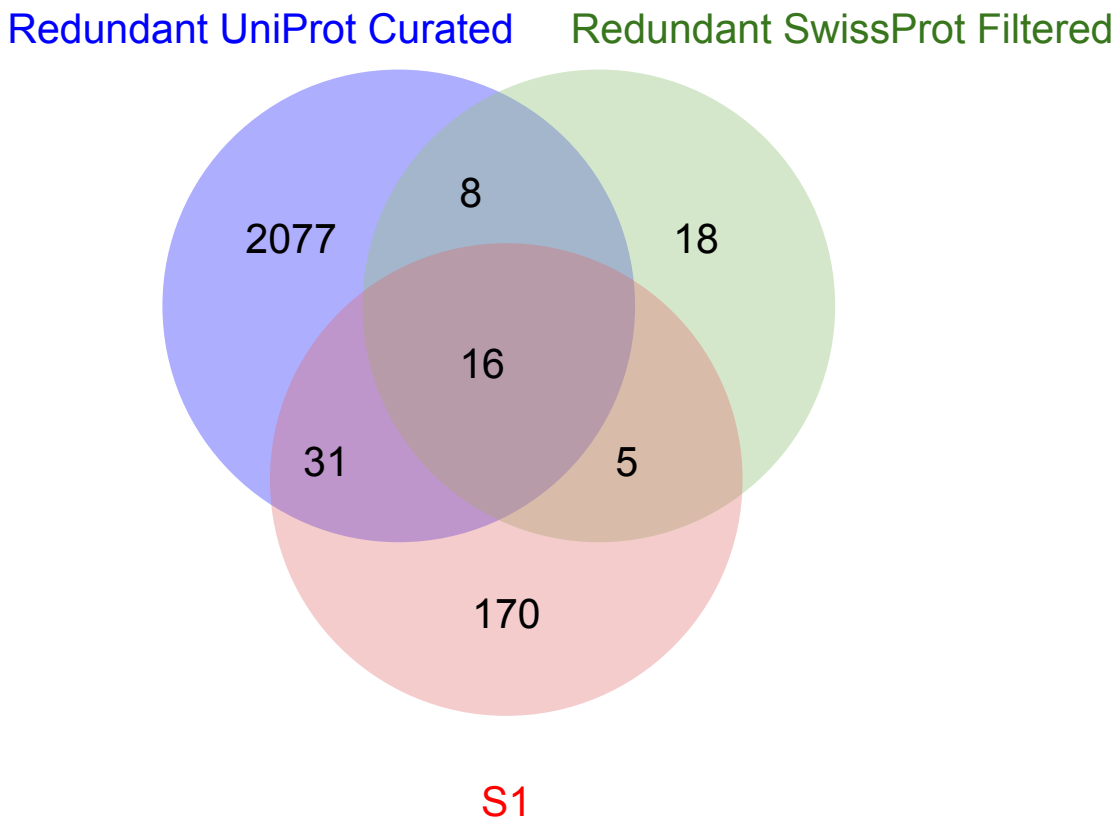

Figure 3.3: A Venn diagram showing tail-anchored protein UniProt ids present in each of the datasets as well as those present in multiple datasets. The number of ids present in redundant versions of i) the supplementary materials table of a previous study predicting the complete set of human tail-anchored proteins denote by S1 [**Kalbfleisch2007**] in red, ii) and in green is the SwissProt dataset filtered according to typical TA features limited to the human proteome [**TheUniProtConsortium2014**], and iii) in blue is the UniProt curated list of TA proteins [**TheUniProtConsortium2014**]. Note that to avoid losing IDs to redundancy reduction this diagram was generated without the use of CD-HIT [**Huang2010**, **Wu2011**], which is applied in the later statistical analysis.

Figure 3.3 shows that S1 has 175 record ids of 222 records (78.8%) which do not share overlap the up-to-date manually curated UniProt dataset [**TheUniProtConsortium2014**]. Of the 170 unique records of that S1 dataset, 4 were manually annotated as not belonging to the human proteome, 20 have the C terminal as annotated being cytoplasmic, only 125 had `TRANSMEM` annotation indicating a bona fide TMH. If we apply equivalent filters, only 42 have annotation verifying that they are TA proteins.

Equivalent criteria to the original Kalbfleisch *et al.,* 2007 [**Kalbfleisch2007**] study were applied to the entire SwissProt database and then restricted to the human proteome dataset. 24 of these 47 records (51.1%) are in the curated UniProt TA dataset. 21 of the 49 (44.7%) records from SwissProt filtered human dataset can be found in the original S1 list.

The same method applied to an up-to-date dataset overlaps more with a manually curated dataset. There is also a large degree of what we now believe to be mistakes that occurred in the older prediction tools and datasets, even when using similar methods. As a trend, this shows that up-to-date datasets improve the reliability of this automated predicted method. These automated criteria still do not fully align with the manually curated list. Of 2633 records in the manually curated list, only 2241 have the `TRANSMEM` annotation. Further is not only the transmembrane annotation itself, but also the type of transmembrane protein. Small integral membrane protein 1 is a blood group antigen (Uniprot ID B2RUZ4) that is just one example of a protein we know to be a post-translationally inserted TA protein, and yet in Uniprot it is annotated as a type II, not type IV, transmembrane protein. As a result of which it appears in the Swissprot automatically filtered list, and not the manually curated list. In an ideal database, where there are instances of discrepency, a note on post-translational or co-translational biogenesis would address this issue. Ultimately, this points to the idea that datasets are a moving target as they are constantly updated with more accurate information using evermore reliable tools and methods.

3.4.2 It Is Difficult To Observe Any Hydrophobic Variation Of TA Protein TMHs From Different Species

In single-pass proteins of eukaryotic species, there are typically various adaptations of the TMH to adhere to the membrane constraints of the specific membrane. For single-pass proteins, previous studies have observed differences in terms of TMH hydrophobicity between yeast and human TMPs [**Sharpe2010**], or in cress, yeast, bacteria, and human datasets [**Baker2017**]. We would expect to see a similar trend between the TMHs of TA proteins from different species. However, when assuming a zero-difference hypothesis, in these TMH TA protein datasets we cannot observe any species-level differences between the datasets at this sample size for TMH hydrophobicity.

When comparing the average Kyte & Doolittle [**Kyte1982**] hydrophobicity values for the TMHs from humans and mice, *A. thaliana*, and *S. cerevisiae*, we can see little difference between the mean values. All of the mean values lie between 2.3-2.6 when we only consider the TMH and at 1.3-1.6 when considering residues in close proximity to the TMH (±5 residues) (Figure 3.4).

Indeed, we see no strong observable statistical differences in hydrophobicity ($P > 3.35E-1$ in the SwissProt automatically filtered list Table 3.1, and $P > 2.40E-1$ in the UniProt curated list Table 3.2). There are also no consistent trends among the absolute Bahadur slopes; no datasets are greatly different from any other.

Table 3.1: Hydrophobicity statistical comparisons between mouse and human, yeast, and plants in the SwissProt Filtered Dataset. Here, we compare a mammalian set of TA proteins (Human N=46 and Mouse N=48) to *A. thaliana* (N=49) representing plants and *S. cerevisiae* (N=24) representing yeasts. The hydrophobicity was predicted as the mean average of the values of the sequences of the TMH, as well another group including up to ±5 flanking residues, since predicting the boundary of TMHs is difficult, according to the Kyte & Doolittle hydrophobicity scale [**Kyte1982**]. The Test column refers to the statistical score obtained from the test; H statistic for the Kruskal Wallis, the KS statistic for the Kolmogorov Smirnov test, and the t-statistic for the T-test. P is the P-value of that statistical score. B refers to the Bahadur slope, an interpretation of the P-value that accounts for the sample size powering the test [**Bahadur1967**, **Bahadur1971**].

		Mammal and Plant			Mammal and Yeast			Plant and Yeast		
		Test	P	B	Test	P	B	Test	P	B
TMH	KW	0.93	3.35E-1	7.64E-3	0.10	7.56E-1	2.37E-3	0.84	3.60E-1	1.40E-2
	KS	0.13	6.36E-1	3.17E-3	0.12	9.24E-1	6.69E-4	0.19	5.28E-1	8.76E-3
	T-test	-0.86	3.90E-1	6.58E-3	0.21	8.31E-1	1.57E-3	0.79	4.33E-1	1.15E-2
TMH and flanks	KW	0.04	8.52E-1	1.12E-3	0.12	7.28E-1	2.69E-3	0.04	8.33E-1	2.51E-3
	KS	0.11	7.72E-1	1.81E-3	0.13	8.79E-1	1.09E-3	0.11	9.80E-1	2.81E-4
	T-test	-0.22	8.23E-1	1.37E-3	-0.38	7.04E-1	2.97E-3	-0.19	8.50E-1	2.22E-3

Here, we are dealing with datasets at least an order of magnitude smaller than those broad studies [**Sharpe2010**, **Baker2017**] which could explain the absence of

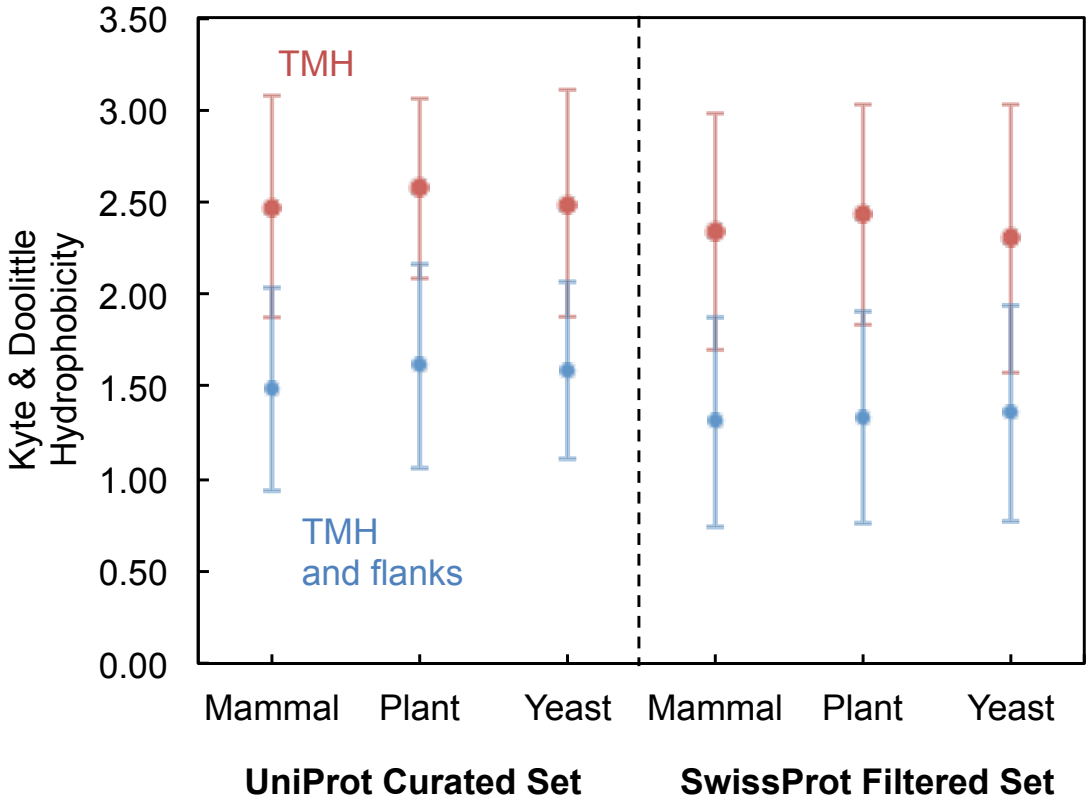

Figure 3.4: Average values of species datasets from UniProt manually curated set and SwissProt automatically filtered dataset.
The average hydrophobicity values from the Kyte & Doolittle scale [**Kyte1982**].for both the TMH and the TMH±5 residues. Values are shown for both the UniProt manually curated set and the SwissProt filtered set. In the UniProt manually curated set we compare the mammalian set of TA proteins (Human N=30 and Mouse N=30) to *A. thaliana* (N=57) representing plants and *S. cerevisiae* (N=27) representing yeasts. For the SwissProt filtered set we compare the mammalian set of TA proteins (Human N=46 and Mouse N=48) to *A. thaliana* (N=49) representing plants and *S. cerevisiae* (N=24) representing yeasts. Error bars are shown at $\pm 1\sigma$ from the mean of the respective dataset.

the effect. However, this only goes to show that if there is a biochemically distinct effect in TA proteins in terms of hydrophobicity between species, it is indeed weak.

3.4.3 There Are Biochemical Differences Between Tail-Anchored TMHs From Different Organelles

Although the species datasets appeared to have no significant differences between them in terms of hydrophobicity, we also investigated the subcellular membranes. We see clear differences in the biochemistry of the TMH (Figure 3.5).

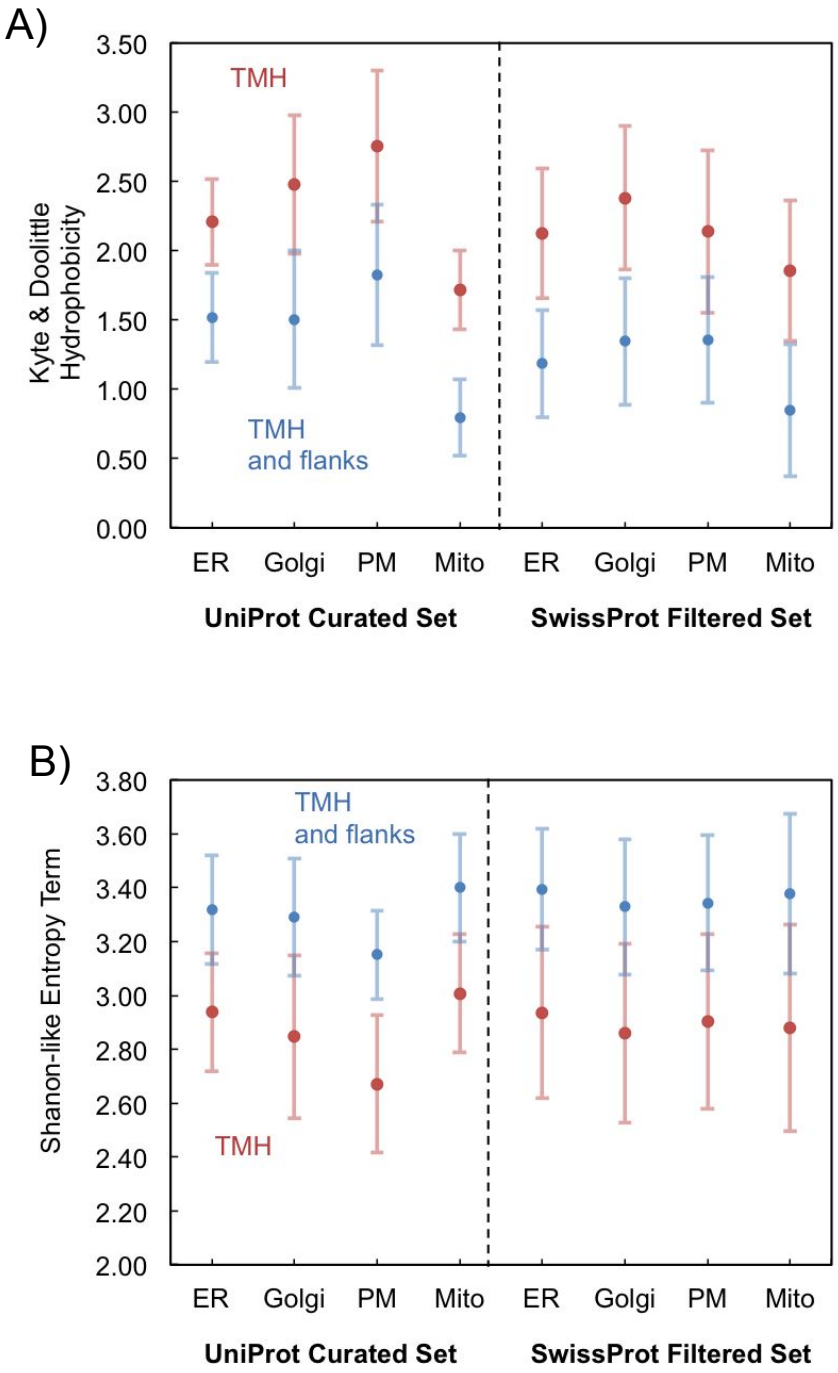

Figure 3.5: Average sequence-based biochemical values of organelle datasets from UniProt manually curated set and SwissProt automatically filtered dataset.
A) The average hydrophobicity values from the Kyte & Doolittle scale [**Kyte1982**], B) the average information entropy [**Shannon1948**] (see methods) for both the TMH and the TMH±5 residues. Values are shown for both the UniProt manually curated set and the SwissProt filtered set. In the UniProt manually curated set we compare TA proteins from the ER (N=400) to the Golgi (N=82), the PM (N=37), and the mitochondria (N=401). For the SwissProt filtered set we compare TA proteins from the ER (N=98) to the Golgi (N=82), the PM (N=157), and the mitochondria (N=65). Error bars are shown at $\pm 1\sigma$ from the mean of the respective dataset.

Table 3.2: Hydrophobicity statistical comparisons between mouse and human, yeast, and plants in the UniProt Curated Dataset. Here, we compare a mammalian set of TA proteins (Human N=30 and Mouse N=30) to *A. thaliana* (N=53) representing plants and *S. cerevisiae* (N=27) representing yeasts. The hydrophobicity was predicted as the mean average of the values of the sequences of the TMH, as well another group including up to ±5 flanking residues, since predicting the boundary of TMHs is difficult, according to the Kyte & Doolittle hydrophobicity scale [**Kyte1982**]. The Test column refers to the statistical score obtained from the test; H statistic for the Kruskal Wallis, the KS statistic for the Kolmogorov Smirnov test, and the t-statistic for the T-test. P is the P-value of that statistical score. B refers to the Bahadur slope, an interpretation of the P-value that accounts for the sample size powering the test [**Bahadur1967**, **Bahadur1971**].

		Mammal and Plant			Mammal and Yeast			Plant and Yeast		
		Test	P	B	Test	P	B	Test	P	B
TMH	KW	0.71	4.01E-01	8.09E-03	0.03	8.72E-01	1.57E-03	0.57	4.48E-01	1.00E-02
	KS	0.13	6.93E-01	3.24E-03	0.13	9.11E-01	1.08E-03	0.20	4.16E-01	1.10E-02
	T-test	-0.93	3.55E-01	9.15E-03	-0.11	9.13E-01	1.04E-03	0.64	5.22E-01	8.12E-03
TMH and flanks	KW	1.37	2.42E-01	1.26E-02	0.38	5.36E-01	7.17E-03	0.08	7.80E-01	3.11E-03
	KS	0.19	2.40E-01	1.26E-02	0.14	8.13E-01	2.38E-03	0.09	9.97E-01	3.21E-05
	T-test	-1.17	2.45E-01	1.24E-02	-0.79	4.35E-01	9.58E-03	0.20	8.43E-01	2.14E-03

In the UniProt manually curated dataset, the Kyte & Doolittle hydrophobicity scores range from 1.7 in mitochondria to 2.7 in the PM (Figure 3.5A).

Table 3.3: Statistical comparisons between TMH sequences from organelles in the UniProt Curated Dataset. Here, we compare an organelle subset from the UniProt curated dataset of TA proteins. We compare ER (N=397) to Golgi (N=83), PM (N=31), and the mitochondria (N=426). The hydrophobicity was predicted as the mean average of the values of the sequences of the TMH, as well another group including up to ±5 flanking residues, since predicting the boundary of TMHs is difficult, according to the Kyte & Doolittle hydrophobicity scale [**Kyte1982**]. The linguistic information entropy was calculated according to the methods section [**Shannon1948**]. The Test column refers to the statistical score obtained from the test; H statistic for the Kruskal Wallis (KW), the KS statistic for the Kolmogorov Smirnov test (KS), and the t-statistic for the student's T-test (T-test). P is the P-value of that statistical score. B refers to the Bahadur slope, an interpretation of the P-value that accounts for the sample size powering the test [**Bahadur1967**, **Bahadur1971**].

		ER and Golgi			ER and PM			ER and mito		
		Test	P	B	Test	P	B	Test	P	B
Hydrophobicity of TMH	KW	21.83	2.98E-06	2.66E-02	28.53	9.21E-08	3.80E-02	377.02	5.54E-84	2.34E-01
	KS	0.34	1.61E-07	3.27E-02	0.57	5.32E-09	4.47E-02	0.67	4.22E-82	2.28E-01
	T-test	-6.45	2.72E-10	4.61E-02	-8.86	2.30E-17	8.99E-02	23.53	6.58E-94	2.61E-01
... and flanks	KW	0.21	6.48E-01	9.07E-04	17.53	2.83E-05	2.46E-02	490.46	1.13E-108	3.03E-01
	KS	0.19	1.10E-02	9.44E-03	0.50	4.69E-07	3.42E-02	0.82	5.58E-123	3.43E-01
	T-test	0.32	7.48E-01	6.07E-04	-4.85	1.75E-06	3.11E-02	34.60	2.19E-162	4.53E-01
Sequence Entropy of TMH	KW	4.66	3.09E-02	7.28E-03	27.54	1.54E-07	3.68E-02	24.03	9.48E-07	1.69E-02
	KS	0.24	4.78E-04	1.60E-02	0.46	4.20E-06	2.91E-02	0.18	2.10E-06	1.59E-02
	T-test	3.22	1.37E-03	1.38E-02	6.42	3.71E-10	5.10E-02	-4.55	6.28E-06	1.46E-02
... and flanks	KW	0.52	4.70E-01	1.58E-03	19.50	1.01E-05	2.70E-02	40.11	2.40E-10	2.70E-02
	KS	0.13	2.06E-01	3.31E-03	0.41	7.97E-05	2.22E-02	0.23	5.53E-10	2.60E-02
	T-test	1.08	2.82E-01	2.65E-03	4.47	1.00E-05	2.70E-02	-5.84	7.51E-09	2.28E-02

In the UniProt curated list, there are clear hydrophobic differences between all the organelle TMH datasets excluding flanks ($P < 2.98E-6$) which as a trend becomes less clear when considering the TMH±5 flanking residues except for mitochondria which increases in significance when considering the flanks also (Table 3.3). The ER and mitochondrial tests are very significant ($P < 4.22E-82$). Consistently the Bahadur slope is at least an order of magnitude greater in the ER and mitochondrial comparison than for the other considerations, so these differences cannot be accounted for by the larger sample size. This gap in hydrophobicity appears to be due to a trend of the ER, PM, and Golgi using isoleucine, valine, and leucine as their most

common TMH residues, whereas in the case mitochondrial located TA proteins, the most common residue type is alanine in the UniProt manually curated dataset (16.3% of total residues) followed by valine (12% total residues)(Figure 3.6).

Similarly, alanine is the second most common residue in mitochondrial located TA proteins from the Swissprot automatically generated dataset at 11.9% of the total residues after leucine which is 13.4% of the total residues(Figure 3.7).

Analysis from 16 TA proteins with known subcellular locations showed that both the C-terminal tail charge and hydrophobicity are determinants of the terminal destination to the ER, mitochondria, and the peroxisome intracellular subcellular locations [**Costello2017**]. They found that less hydrophobicity and more charge in the "tail" determined the TA protein for the mitochondria rather than the ER. This corroborates what we see in terms of hydrophobicity (Figure 3.5A). When we consider charge difference between organelles on larger datasets, we see trends that reinforce this idea, however, rather than net charge, we see charge distribution along the TMH and the neighbouring flanks. In the Swissprot automatically filtered dataset, in the ER 9.4% of the residues are positively charged, and 2.5% are negatively charged. Most of the positively charged residues cluster following the "positive-inside" rule between positions -15 and -10 for R and K, but so do the negatively charged residues D and E, effectively reducing this local charge by 2.5%. In mitochondria, we find that the proportion of charge is similar (10.4% R and K, 3.3% D and E) however the negatively charged residues cluster on the N flank (-15 to -8) and the positively charged residues cluster more strongly on the outside flank (positions 9 to 15) (Figure 3.7).

In the UniProt manually curated ER set, 6.6% of residues are positively charged and 3.3% of residues are negatively charged. K clusters strongly on the inside flank as expected, yet R clusters strongly between positions 7 to 15 and rather weakly at the inside flank (positions -15 to -10) (Figure 3.6). Similarly to the SwissProt sets, D prefers the inside flank but is tolerated in the outside flank. The more abundant E residues behave very unusually and cluster at positions 5-10. Generally, charged residues are suppressed in the TMH core [**Sharpe2010**, **Baeza-Delgado2013**], especially in anchoring TMHs [**Baker2017**]. It is unclear why this is observed, yet, altogether the 313 glutamic acid residues and 397 arginine residues that appear unusually deep in

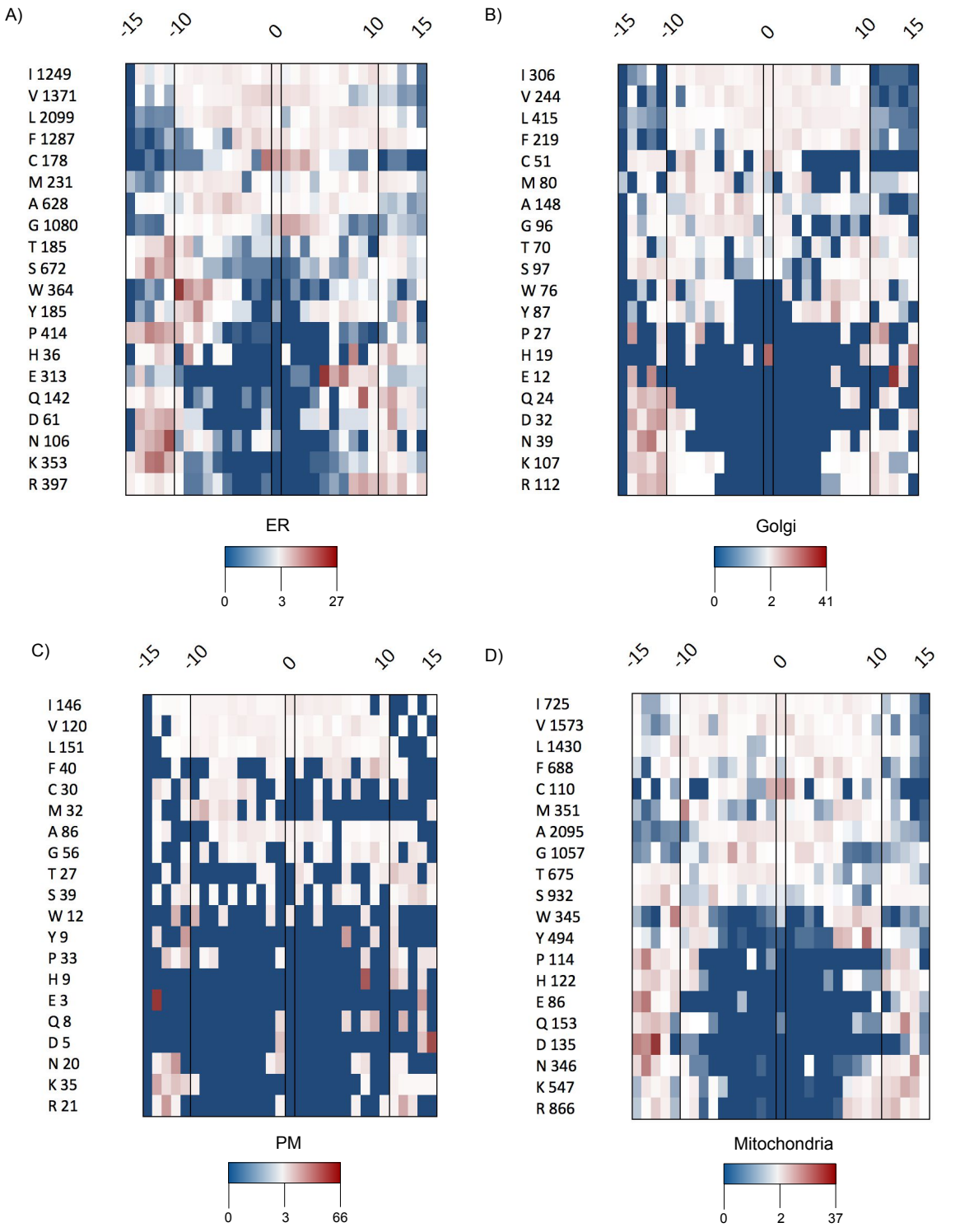

Figure 3.6: The normalised skews of each amino acids from TA proteins grouped by localisation from the UniProt manually curated dataset.
The residue position aligned with the centre of the TMH is on the horizontal axis, and the residue type is on the vertical axis. Amino acid types are listed in order of decreasing hydrophobicity according to the Kyte and Doolittle scale [**Kyte1982**]. Flank lengths were restricted to ±5 residues. The edge residues from proteins with flank lengths and TMH lengths that exceeded the plotted 31 residues were still included in the normalisation calculations despite not being plotted. The colour scale represents the relative percentage of a particular amino acid and is shown with dark blue as 0, white as the 50th percentile value of the entire heatmap, and dark red as the highest percentage on the heat map. The panels are constructed from TA proteins derived from the SwissProt automatic method with redundancy removal applied detailed in the methods section. The datasets were further separated by subcellular locations: (a) the ER, (b) the Golgi, (c) the cell membrane, (d) the mitochondria. These datasets are more thoroughly outlined in the methods section.

the TMH core may be to an extent neutralising one another in the folded TMH arrangement, but are ultimately not that abundant compared to the total number of residues in this organelle dataset (11351 total residues). In mitochondria, 1413 positively charged residues (11% of the total residues in the mitochondrial dataset) were preferentially located at the outside flank and somewhat into the core (positions 6 to 15) than the expected "inside" flank (positions -15 to -5). The 221 negatively charged residues (1.7%) unusually cluster at the inside. This results in a strong net positive-outside charge signal since there are more positively charged residues on the outer flank uncountered by the negatively charged residues, which are skewed with a preference for the inside flank.

Information entropy has been known to identify cryptic function in TMHs when considered along with hydrophobicity [**Wong2011**, **Wong2012**]. In terms of information entropy, there is a marked decrease in entropy in the PM subset (mean entropy = 3.15 in the TMH, 2.67 including ±5 flanking residues) from the UniProt curated dataset compared to the other organelle datasets (entropy > 3.29 and > 2.85 including the flanks). However, this stark difference between TMHs from PM bound TA proteins and the other organelle datasets cannot be observed in the SwissProt set (Figure 3.5).

No clear significant differences can be observed for the information entropy ($P > 6.33E-2$). This is unsurprising given that the hydrophobic nature of the TMHs demands that certain residues must be over-represented, which lowers the information entropy. In this case, we have a highly hydrophobic set, the PM UniProt set, which likely contains a higher proportion of the most hydrophobic residues. As a trend, the information entropy mirrors the hydrophobicity albeit with less range between dataset means (2.67-3.15 in the TMH for information entropy, 1.72-2.74 for hydrophobicity)(Figure 3.5).

Similarly, in the SwissProt filtered dataset, the mean TMH hydrophobicity for mitochondria is the lowest at 1.9, but it appears to be the Golgi apparatus that is the peak at 2.4. In the SwissProt dataset, when we compare each subset of only the TMH to the ER subset, we find significance between the ER and the Golgi ($P < 1.98E-3$), and the ER and the mitochondria ($P < 4.62E-3$), however, the ER and PM are more similar considering the Bahadur values are $< 6.44E-4$, two orders of magnitude smaller than the other sets (Bahadur values $> 3.3E-2$) (Table 3.4). When we take

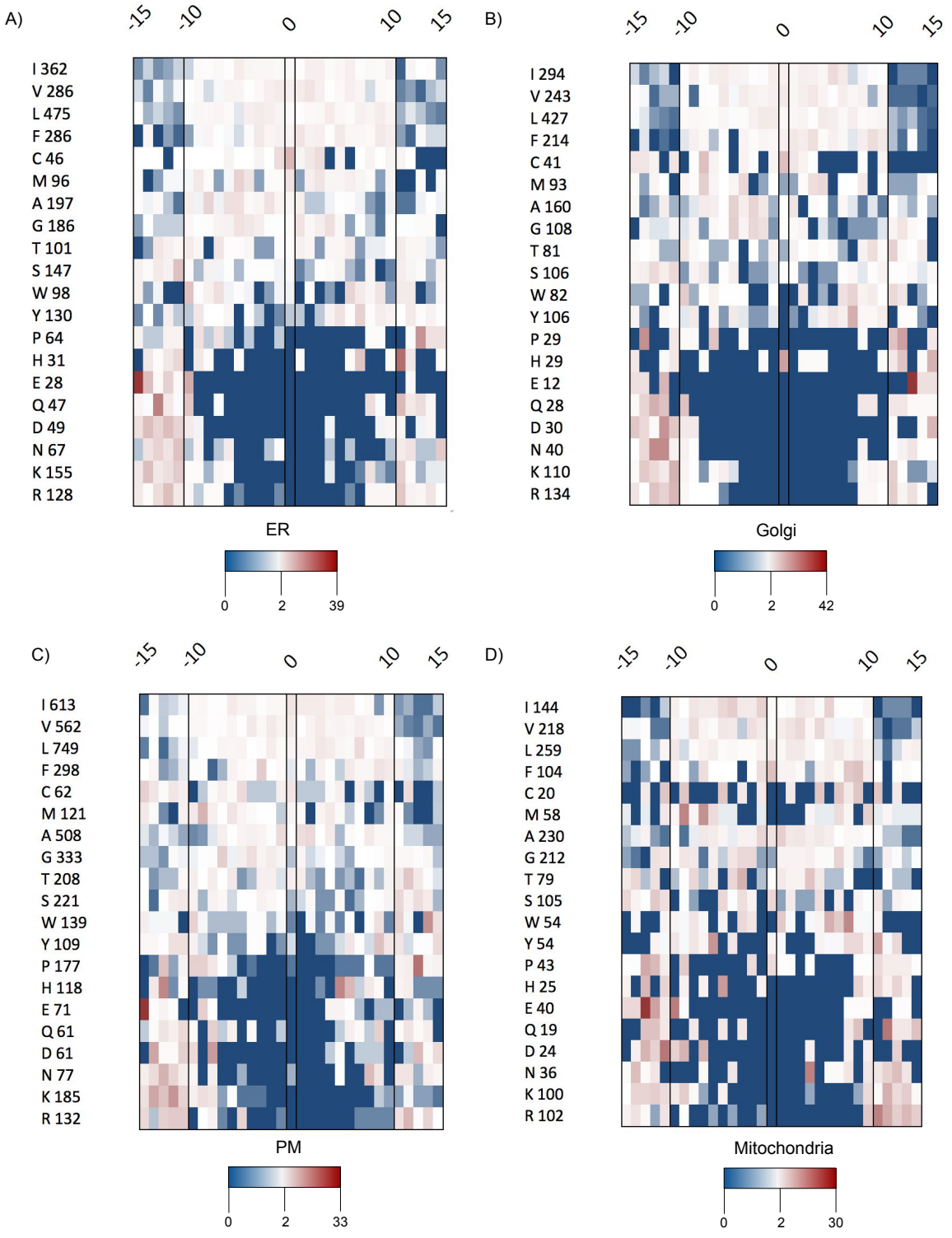

Figure 3.7: The normalised skews of each amino acids from TA proteins grouped by localisation from the SwissProt automatically filtered dataset
Similarly to figure 3.6, the residue position aligned with the centre of the TMH is on the horizontal axis, and the residue type is on the vertical axis. Amino acid types are listed in order of decreasing hydrophobicity according to the Kyte and Doolittle scale [**Kyte1982**]. Flank lengths were restricted to ±5 residues. The edge residues from proteins with flank lengths and TMH lengths that exceeded the plotted 31 residues were still included in the normalisation calculations despite not being plotted. The colour scale represents the relative percentage of a particular amino acid and is shown with dark blue as 0, white as the 50th percentile value of the entire heatmap, and dark red as the highest percentage on the heat map. The panels are constructed from TA proteins derived from the SwissProt automatic method with redundancy removal applied detailed in the methods section. The datasets were further separated by subcellular locations: (a) the ER, (b) the Golgi, (c) the cell membrane, (d) the mitochondria. These datasets are more thoroughly outlined in the methods section.

Table 3.4: Statistical comparisons between TMH sequences from organelles in the SwissProt Filtered Dataset. Here, we compare organelle subsets from the SwissProt automatically filtered dataset of TA proteins. We compare ER (N=98) to Golgi (N=82), PM (N=157), and the mitochondria referred to as "mito" (N=65). The hydrophobicity was predicted as the mean average of the values of the sequences of the TMH, as well another group including up to ±5 flanking residues, since predicting the boundary of TMHs is difficult, according to the Kyte & Doolittle hydrophobicity scale [**Kyte1982**]. The linguistic information entropy was calculated according to the methods section [**Shannon1948**]. The Test column refers to the statistical score obtained from the test; H statistic for the Kruskal Wallis (KW), the KS statistic for the Kolmogorov Smirnov test (KS), and the t-statistic for the student's T-test (T-test). P is the P-value of that statistical score. B refers to the Bahadur slope, an interpretation of the P-value that accounts for the sample size powering the test [**Bahadur1967**, **Bahadur1971**].

		ER and Golgi			ER and PM			ER and mito		
		Test	P	B	Test	P	B	Test	P	B
TMH Hydrophobicity	KW	11.96	5.43E-4	4.18E-2	0.02	8.77E-1	5.14E-4	8.46	3.64E-3	3.45E-2
	KS	0.27	1.98E-3	3.46E-2	0.08	8.48E-1	6.44E-4	0.27	4.62E-3	3.30E-2
	T-test	-3.47	6.50E-4	4.08E-2	-0.17	8.67E-1	5.60E-4	3.45	7.24E-4	4.44E-2
... including flanks	KW	5.92	1.50E-2	2.33E-2	9.14	2.50E-3	2.35E-2	26.42	2.75E-7	9.27E-2
	KS	0.21	2.85E-2	1.98E-2	0.26	4.88E-4	2.99E-2	0.43	4.93E-7	8.91E-2
	T-test	-2.52	1.25E-2	2.43E-2	-3.09	2.23E-3	2.40E-2	4.95	1.87E-6	8.09E-2
TMH entropy	KW	2.96	8.56E-2	1.37E-2	0.66	4.17E-1	3.43E-3	0.69	4.05E-1	5.54E-3
	KS	0.13	4.32E-1	4.66E-3	0.10	5.27E-1	2.51E-3	0.18	1.40E-1	1.20E-2
	T-test	1.58	1.15E-1	1.20E-2	0.79	4.32E-1	3.29E-3	1.03	3.06E-1	7.26E-3
... including flanks	KW	2.62	1.06E-1	1.25E-2	2.87	9.04E-2	9.42E-3	0.05	8.31E-1	1.14E-3
	KS	0.15	2.48E-1	7.75E-3	0.17	6.56E-2	1.07E-2	0.21	6.33E-2	1.69E-2
	T-test	1.84	6.75E-2	1.50E-2	1.66	9.84E-2	9.09E-3	0.42	6.72E-1	2.44E-3

into account the flanks, the ER and PM dataset can be distinguished ($P < 2.50E-3$), however, as a trend the other two comparisons, ER and Golgi become less significant, and ER and mitochondria become more significant.

The information entropy of the TMH string was also examined. No significance was observed in any consideration of the information entropy, but similarly to the UniProt subset, as a trend, the entropy mirrors the hydrophobicity (Figure 3.5).

The known lipid asymmetry caused by sphingomyelin and glycosphingolipids on the non-cytosolic leaflet and phosphatidylserine and phosphatidylethanolamine in the cytosolic leaflet in the Golgi and PM and lack of asymmetry in the ER [**Daleke2007**, **Devaux2004**], or that sphingomyelin is not present in the ER but is present in the Golgi [**Futerman2005**] and PM [**Li2007**, **Tafesse2007**]. Furthermore, the PM contains densely packed sphingolipids and sterols [**Paolo2006**]. Mitochondira have bacterial lipids in their membrane and uniquely contain cardiolipin [**Choi2005**], which is also present in the mitochondrial outer membrane [**Gebert2009**].

Hydrophobic differences in the have already been observed in the mitochondria localised TA protein TMHs [**Borgese2003**]. Here we observe that average biochemical features are evidently of significance. Furthermore we see that the typically positive-inside negative outside tandem in positively and negatively charged residues is reversed

in the mitochondria to positive-outside negative-inside. This variation in TMH hydrophobicity and charged residue skew reversal may be yet another nuance of the system which go some way to explaining how signals are maintained in local environments even when the average values are ambiguous. We also identify that alanine is a key reason behind the hydrophobic difference between subcellular organelles, with alanine being highly selected, if not favoured over other hydrophobic residues like leucine, in mitochondrial TA proteins compared to ER, the Golgi, and the PM. This could be an adaptation to the mitochondrial membrane, which contains a higher level of cardiolipins than other membranes [**Meer2008**, **Gebert2009**]. Regarding the charged residue distribution, there should also be a consideration of membrane potential. Although membrane potentials are in flux, typically, the PM has a potential of 70mV with the negativity being on the cytoplasmic side. It has been shown that the ER is again between 75-95mV with negativity on the lumenal side [**Quin2011**, **Worley 1994**]. There is no detectable potential across the Golgi [**Schapiro2000**], and the mitochondrial inner membrane has a potential of 150-180 with negativity on the matrix side [**Seth2011**]. However, whilst those numbers go some way to satisfy the flanking charge reversal we see, they do not apply to the mitochondrial outer membrane in which the TA proteins are localised; porins on the mitochondrial outer membrane effectively diminish the membrane potential across the membrane.

It has been known for some time that both the charge and the hydrophobic length of the TA protein TMH region determine subcellular targeting of the protein [**Borgese2003**]. The C-terminal tail charge is particularly important for determining subcellular localisation, and can even override the hydrophobic signal if strong enough [**Costello2017**]. Here we see that while hydrophobicity is statistically different between the subcellular membranes, there is overlap. Whilst there are differences in total average charge at the C-terminal flank, this is not an absolute rule.

Nevertheless, it is tempting to conclude that these biochemical differences between localised TA proteins is an adaptation to the membrane composition and environment. But this must be tempered by noting that the spontaneously inserting cytochrome b5 localises to the mitochondrial membrane in the absence of cytosol, and to the ER in the presence of cytosol [**Costa2018**]; there are also biological factors determining localisation.

In summary, the mitochondria located TA protein TMHs typically have a preference for alanine over leucine unlike their secrtetory counterparts and have a negative-inside positive-outside tendency counter to the overwhelming majority of TMPs. It is unclear if these features are a biophysical adaptation, or part of a biological sorting process.

3.4.4 More annotation is required to identify chaperone interaction factors of the TMH.

TA proteins known to interact with certain chaperones were acquired by filtering the interactor partner IDs for chaperones from BioGrid through the redundant versions of these UniProt manually curated lists and Swissprot automatically generated lists.

Hsp40, Hsc70, SRP54 (both plant and human) returned 0 hits, indicating a lack of annotation regarding TA proteins with these chaperons probably due to the relatively polar, and non-trivially predictable, TMHs of TA proteins that these chaperones interact with.

Snd1 has 15 records that were in our TA lists. The average Kyte & Doolittle hydrophobicity of these records was 2.60 in the TMH itself and 1.58 including $\pm$ 5 flanking residues. Sgt2, with 14 records, had a TMH hydrophobicity of 2.47 and 1.51 including the flanks. 5 records were captured for SGTA with a TMH hydrophobicity of 2.27 and 1.19 including the flanks. TRC40 had the highest TMH hydrophicity of 2.77 and 1.82 including flanks. However, TRC40 also only had 7 records. Get3 had 22 TA interactor records with an average TMH hydrophobicity of 2.36 and 1.48 including the flanks. The 2 records for human Pex19 had an average hydrophobicity of 1.33 for the TMH and 0.70 including the flanking residues. The yeast Pex19 had a TMH average hydrophobicity of 2.48 and 1.41 including the flanking residues. The 4 yeast SRP54 interactors had an average TMH of 2.43 and 1.98 including the flanking residues.

At the time of the investigation, these sample sizes are not statistically viable for analysis. Whilst it appears TRC40 interactors have notably hydrophobic TMHs, TRC40s yeast homologue Get3 has interactors with much more polar TMHs, yet SGTA was lower than SGT2 on average. So although these average values differ and overlap between various chaperone systems, we tried to identify clearer patterns from the TMH

hydrophobic profiles (Figure 3.8). Similarly, whilst a clear dip in hydrophobicity at position +5 in Get3 from 2-3 across the rest of the TMH core to 0.32, there is no such spike for TRC40 meaning this is probably not of any functional importance, but rather an artefact of overrepresented proteins in the Get3 dataset. Snd1 also lies among these values, reinforcing Snd1 as a biological redundancy system [**Rabu2009**, **Johnson2013**, **Schuldiner2008**].

As expected, the human Pex19 interactors are as a trend among the most polar throughout the TMH core, however, when we consider the yeast Pex19, this trend is less clear.

At least at a handful of locations (-3, 3, 4 and 5) the SRP54 interactors have the most hydrophobic TMH cores.

TRC40 has the highest TMH hydrophobicities at 3.9 at position -1, however the yeast homologue Get3 doesn't appear to have any preference for especially hydrophobic TMHs.

In order to remove redundant proteins and investigate this further, more records with greater levels of accurate annotation need to be available to both BioGrid and UniProt. We also observe a great deal of overlap between the profiles, indicating that as a trend this is more complex than hydrophobicity alone, or at least polarity is not the absolute determinant of subcellular targeting. Hydrophobicity alone would have resulted in stark contrasts between chaperone interactor pairs, even with low sample sizes.

However, this method demonstrates a potential way that this chaperone-interaction problem can be investigated to verify that indeed hydrophicity plays a deterministic role in chaperone selection.

3.4.5 Spontaneous Insertion May Be Achieved by Polar Strips in the TMH of Tail-Anchored Proteins

The TMHs of cytochrome b5 and PTP1b are among the least hydrophobic of the TA proteins and in theory misses the ΔG requirements of a TMH due to their relatively polar TMHs [**Rabu2008**, **Rabu2009**]. Indeed the TMH is so polar that it is not

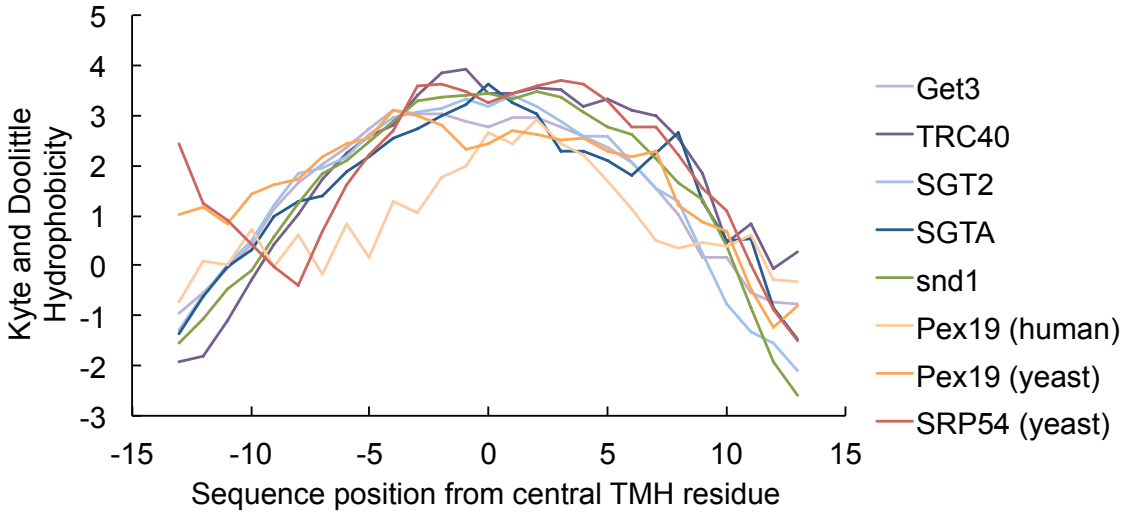

Figure 3.8: The profile of TMH and flanks hydrophobicity from TA protein groups stratified by chaperone interactors. On the horizontal axis is the position relative to the central TMH residue defined by UniProt. On the vertical axis is the Kyte & Doolittle hydrophobicity windowed across 5 residues allowing for half windows. The chaperone interactors are colour coded according to the key.

trivial to predict and is not found in either dataset prepared herein. Structural modelling and analysis thereof reveal features that may explain the "missing hydrophobicity" [**Hessa2005**, **Hedin2010**, **Hessa2007**, **Ojemalm2012**] of these particular TMHs.

The electrostatic surfaces are prototypical of a TMH anchor with large "positive-inside" patches [**VonHeijne1989**, **Andersson1992**, **Sharpe2010**, **Baeza-Delgado2013**, **Pogozheva2013**, **Baker2017**] and a strong "negative--outside" charge [**Baker2017**](Figure 3.9C). Once in the membrane, this may allow it to be an effective anchor despite such poor hydrophobicity since it satisfies electrostatic coupling to the membrane potential.

Furthermore there is the question of overcoming the unfavourable interaction most TMHs would face when coming into contact with the highly polar membrane interface. We observe a highly conserved strip of relatively polar / non-hydrophobic residues on one side of the TMH core (in cytochrome b5 these are N112, P116, A120, A124 Y127, and R128). Similarly, a polar face exists for the PTP1b TMH (R430, N434, Y426, T422, and T419). These polar faces would not be as repulsed by the interfacial environment as either a more hydrophobic TMH or an equally hydrophobic TMH with a different sequence and structure order (Figure 3.9 and Figure 3.10). Scrambling the cytochrome b5 TMH sequence whilst maintaining the same hydrophobicity (DSNSS

W W T N W V I P A I S A L I V A L M YR to DSNSS W W A S A I I A T M I P L L V N V W YR) reduces the insertion potential [**Brambillasca2006**]; there is more to it than hydrophobicity alone. It becomes apparent that the 3D arrangement of these relatively polar TMH residues is conserved and is probably the key to spontaneous insertion of TMHs.

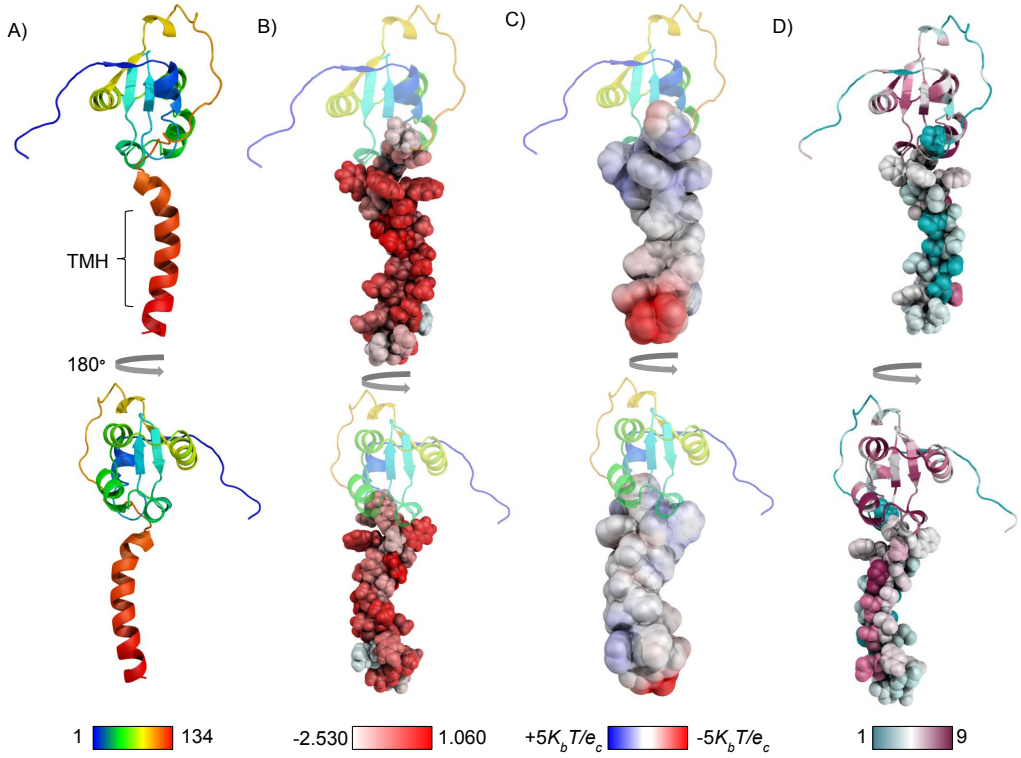

Figure 3.9: Structural biochemical analysis of a homology model of cytochrome b5. (A) The secondary structure of the protein coloured from the N terminus in blue to the C terminus in red coloured through the rainbow according to the residue number. (B) The hydrophobicity of the TMH from white representing relatively polar residues to red showing relatively hydrophobic residues [**Eisenberg1984**]. (C) The electrostatic surface with a threshold of ±5 KT/e calculated by APBS in PyMol [**Baker2001**]. Red patches are negatively charged whilst blue is positively charged. (D) The consurf scores on a scale of 1-9 (all residues had sufficient data) [**Ashkenazy2010**]. Purple represents the most conserved whilst blue is the least. Note the correlation between the highly and modestly conserved TMH residues and the relatively polar residues. Another observable feature is the very strong "positive inside" [**VonHeijne1989**, **Andersson1992**, **Sharpe2010**, **Baeza-Delgado2013**, **Pogozheva2013**] and "negative outside" features which are associated with anchoring [**Baker2017**].

We speculate that this polar face allows the unassisted approach to the membrane's polar phospholipid head groups, and once in sufficiently close proximity, the hydrophobic side of the helix is entropically driven by the water environment into the membrane (Figure 3.11). This close proximity is less likely to be achieved if there is no side of the TMH to favourably interact with the head groups, even though average

hydrophobicity could be similar. Once integrated, the TMH charged flanking regions help sustain integration in lieu of a more hydrophobic TMH core.

3.5 Summary

Here, we have observed a large biochemical distinction between TA proteins with different terminal destinations. Previously it was known that both hydrophobicity and charge are involved in targeting [**Costello2017**]. In this study, we find that the location of the charge along and around the TMH is different in different subcellular compartments. Crucially, there is a shift in the charged residue inside-outside tandem in the TMH flanking residues in different organelles which in the secretory pathway adheres to the membrane-potential electrostatic coupling, but in the mitochondrial outer membrane where very little potential exists, positively charged residues are skewed outside the cytoplasm, and the negatively charged residues are preferentially found inside the cytoplasm. Furthermore, the missing hydrophobicity of mitochondrial TA proteins can be in part attributed to the high abundance of alanine rather than leucine or isoleucine in the TMH. We expected to see evolutionary adaptations of TMH hydrophobicity to species-specific membranes, even within eukaryotes [**Baker2017**, **Sharpe2010**]. In this study using both a manually curated dataset from UniProt and an automatically filtered list using SwissProt annotation, we do not observe any strong differences. Since we could not scrutinise a difference in the species, the strong hydrophobic differences between organelle TMHs are indicative of a stronger adaptation pressure than between species as a whole. These differences are likely to be partially adaptations to the organelle location membrane type and also possible cryptic biological factors that play a role in their targeting via chaperone-binding affinity. This could be a functional similarity to the signal-anchored proteins. Signal anchored proteins contain a single hydrophobic segment that serves as both a mitochondrial targeting signal and a membrane anchor. Signal anchored proteins, along with some TA proteins, have been shown to be able to spontaneously insert into the membrane independently from the translocon [**Elisa2012**, **Lan2000**, **Colombo2009**].

We could not find any clear trends or perform statistical work on the interactor datasets due to the small sample sizes, however, as the databases are enriched, this

same method will be able to answer the questions about chaperone affinity with more accuracy in the future.

Furthermore, the spontaneously inserting TA proteins PTP1b and cytochrome b5 appear to share a polar face that emerges in structural models and a strong positive--inside negative-outside electrostatic surface. The polar face may be responsible for the promotion insertion potential in the absence of insertion proteins since when the sequence is scrambled, the insertion potential is reduced [**Brambillasca2006**]. The positively and negatively charged residues are distributed like an ideal anchoring TMH [**Baker2017**] which could allow the marginally hydrophobic TMH to perform as a suitable membrane anchoring feature.

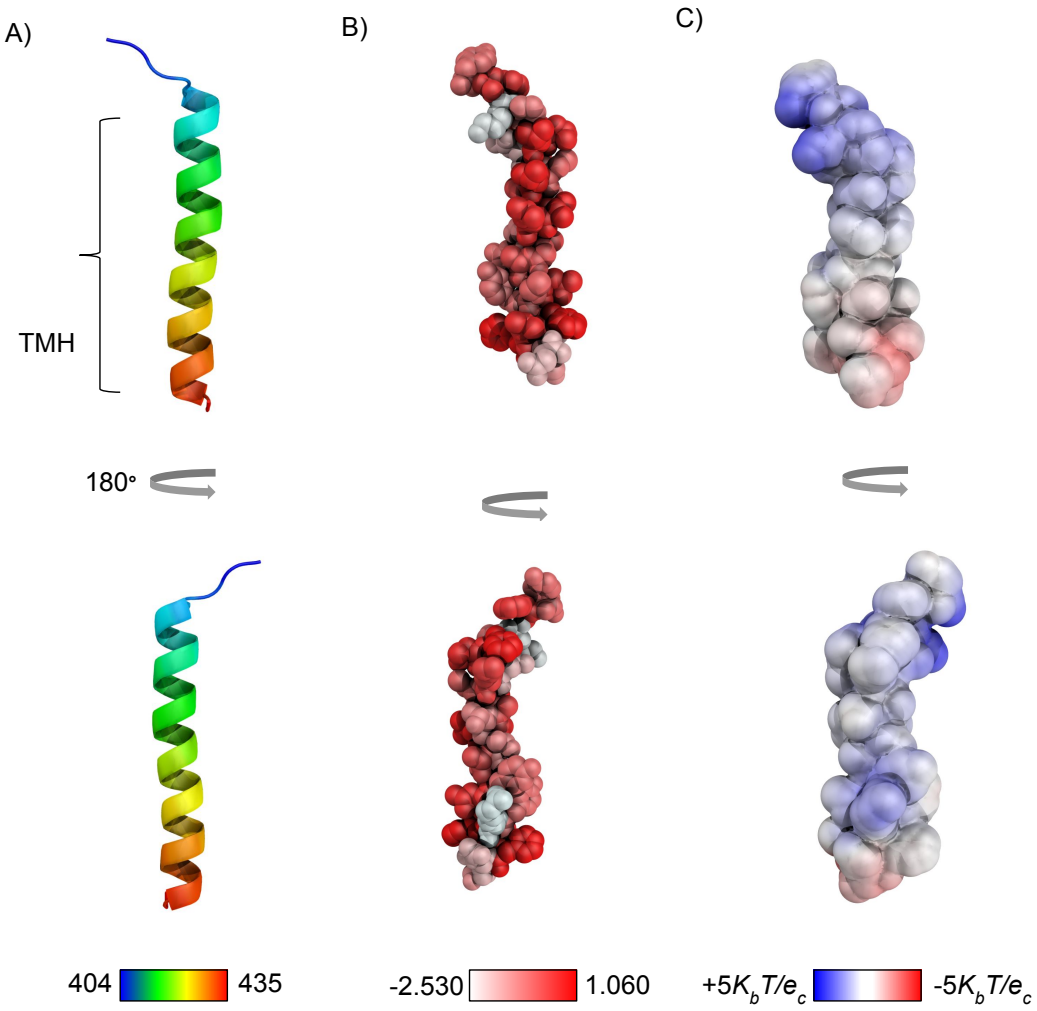

Figure 3.10: Structural biochemical analysis of a homology model of PTP1b. (A) The secondary structure of the protein coloured from the N terminus in blue to the C terminus in red coloured through the rainbow according to the residue number. (B) The hydrophobicity of the TMH from white representing relatively polar residues to red showing relatively hydrophobic residues [**Eisenberg1984**]. (C) The electrostatic surface with a threshold of ±5 KT/e calculated by APBS in PyMol [**Baker2001**]. Red patches are negatively charged whilst blue is positively charged. Note the one hydrophobic face of the TMH and the opposing relatively polar face. Another observable feature is the "positive inside" [**VonHeijne1989**, **Andersson1992**, **Sharpe2010**, **Baeza-Delgado2013**, **Pogozheva2013**] and "negative outside" features which are associated with anchoring [**Baker2017**].

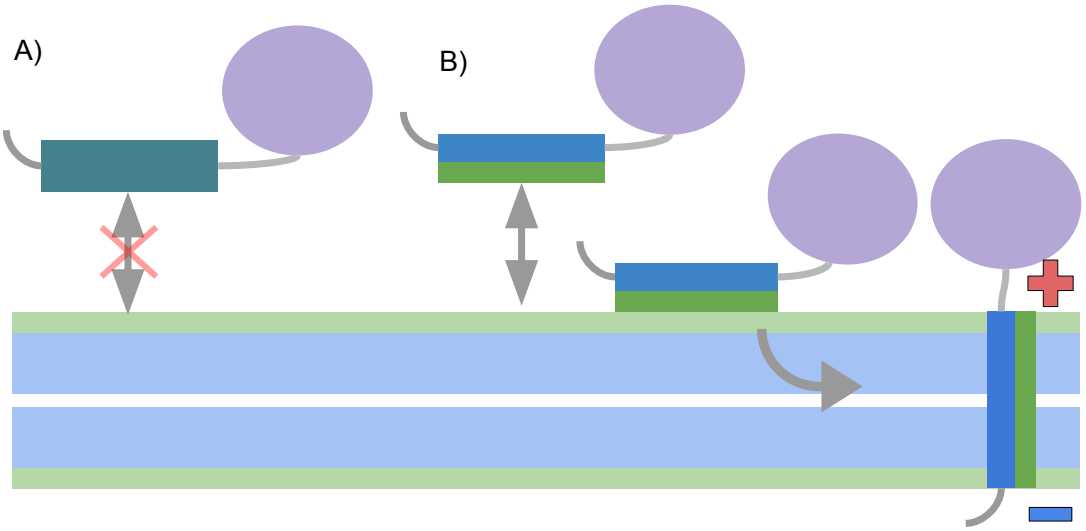

Figure 3.11: A cartoon of a potential method the cytochrome b5 TMH could integrate spontaneously into the membrane. A) The marginally hydrophobic TMH in teal cannot approach the membrane interface. B) Although the average hydrophobicity is the same as the teal TMH, by having a more hydrophobic side (blue), and a more polar side (green), the TA protein now has a TMH surface that may interact more favourably with the interfacial region. Once interacting with the membrane sufficiently close to the interfacial region, the hydrophobic face is still being entropically driven by water molecules from the cytosol, which would lead to partitioning into the membrane. Once integrated, the strong positive-inside negative-outside charges on the TMH flanks compensate for the lack of hydrophocitiy in the core of the TMH.

Chapter 4

Co-operative TMHs

4.1 Abstract

4.2 Introduction

Translocation is when a ribosome translates the Ribonucleic Acid (RNA) to a nascent peptide chain which is handed directly or indirectly to the translocon insertion machinery which threads the chain through and, in the case of TMHs, releases the TMH into the membrane environment.

The overwhelming majority of TMPs use the co-translational method of translocation. It has long been understood that this method is essentially the SRP recognising and attaching to the nascent peptide chain whilst it is still associated with the ribosome, and the SRP then targets the peptide and ribosome to a SR in association with some membrane insertion machinery on the ER membrane [**Pool2005**, **Hessa2005**].

Crystal structures showed the SRP targets the nascent peptide chain for membrane insertion via a GTPase in both the SRP and SR, that is initially associated with the translocon machinery, coming together to form a complex thus bringing the nascent peptide chain in proximity to the translocon [**Shan2005**]. Mutant studies of SRP revealed key discrete conformational stages [**Shan2005**]. These are the specific recognition of signal sequences on cargo proteins, the targeting of the package to the membrane, the handing over of the cargo to the translocation machinery all the while maintaining precise spatial and temporal coordination of each molecular event [**Saraogi2011**].

The prevailing idea about membrane insertion by the translocon is that the TMHs partition in the membrane one at a time as the translocon lateral gate opens, exposing the TMH to the membrane (Figure 4.1)[**Cymer2015**].

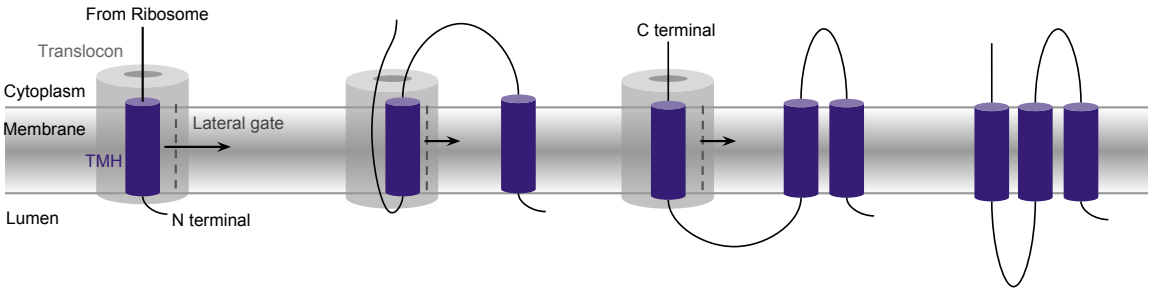

Figure 4.1: A cartoon showing the generally accepted schematic of sequential multipass TMH insertion into the membranes. The two key concepts are that, one at a time, the TMHs emerge from the ribosome into the translocon. This appearence of hydrophobicity triggers the lateral gate to open. As the nascent TMH is exposed to the membrane, it begins to partition. The downstream protein from the TMH is then threaded through the translocon until the next TMH is recognised. This implies that the TMHs ultimately have no meaningful interactions with one another until the protein has been threaded into the membrane and the multiple TMHs form a bundle.

4.2.1 Co-operative insertion

Multiple TMHs in a nascent protein can be associated with the eukaryotic translocon simultaneously. It was shown that TMHs can stay inassociation with the translocon in order to mediate integration of downstream TMHs demonstrated by crosslinking analysis [**Sadlish2005**, **Cross2009**]. Not only this, but it was shown that there is a direct interaction between the TMHs; more recently Arrest Peptide (AP)s were used to show pulling forces between a TMH and more C-terminally located TMH during the C-terminal TMH membrane partitioning from the translocon *in vivo* [**Cymer2013**]. This could be facilitated during the probing of a TMH from the translocon as the lateral gate "cracks" open in an intermediate stage before the TMH satisfies the full hydrophobic requirements to open the gate fully, an intermediate stage observed in a SecY crystal structure [**Egea2010**].

GPCRs

G protein-coupled receptor (GPCR)s are a diverse family of membrane surface receptors with 7 TMH segments. GPCRs have long be known to be overrepresented among genomes [**Remm2000**]. They have adapted to respond to a wide range of specific

signals ranging from macromolecules to photons. The specific signal triggers a conformational change of the GPCR that is translated across the membrane. GPCRs have been associated with tumerigenesis [**OHayre2013**], metastasis [**Singh2015**] and in cancers [**Bar-Shavit2016**] and are a potential target for therapies [**Arakaki2018**]. Their ubiquitous presence in cellular life and medical relevance makes them an important topic of study.

Opsins are a group of light sensitive GPCRs. It was shown by cross linking studies that opsin TMHs 5-7 are retained in the ER translocon and only parition once biosynthesis is complete [**Ismail2008**]. The timing of this partitioning is controlled by the hydrophobicity of the TMH, not protein length or the relative position of the TMH within the protein. Although artificially extending the C-terminal did not result release of the TMHs, by replacing native TMH 7 with a more hydrophobic TMH, the speed of insertion was decreased. TMHs 1-4 are inserted independently, and the 5-7 TMHs partition into the membrane at the same time.

4.2.2 Voltage gated ion channels

Another example of cooperative TMH insertion is that of the 3rd and 4th TMH of the potassium channel (shaker family) which was shown to insert either sequentially or cotranslationally [**Zhang2007**, **Cymer2015**]. This is especially notable in the case of KAT1, that is a plant K_v channel that is thought to mediate long-term potassium influx into guard cells causing the stomata to open. In the case of KAT1, N-glycosylation of various mutant fusion KAT1 constructs revealed that there is no choice of sequential insertion since TMH 3 and 4 have no insertion potential and no topogenic functions themselves [**Sato2002**, **Sato2003**]. In TMH 4 this is due to the charged residues making it relatively polar. However, previous experimentation in Kv1.3 had found that while TMH 4 did not initiate insertion, it did have insertion potential, and that when constgructs contained multiple TMHs, membrane insertion efficiency increased [**Tu2000**]. Without the ability to stop the translation through the translocon and form a TMH, it was suggested that a different means was needed than classic sequential insertion, and even that TMH 3 and 4 are integrated by the translocon at the same time post-translationally, i.e the TMHs are folded prior to insertion [**Sato2003**]. They achieve this in part becase the previous TMHs 1 and 2 form a firm “base” within the

membrane environment.

4.2.3 Ribsomes in the biogenesis of membrane proteins.

Ribosomes translate mRNA sequences to amino acid chains and are present in all living cells, and indeed the ribosomal complexes presence and activity is for many used to define whether something is alive. They are a highly conserved RNA-protein complex with a multitude of accessory proteins and targetting factors.

During translation of a TMP protein, the SRP binds to the ribosome after recognising the nascent protein as a TMP. This complex then binds to the SR in association with the membrane bound translocon. The nascent peptide is then fed into the translocon as it is being translated; hence "co-translational insertion".

APs are typically 10-15 residues long that bind to the upper end of the ribosomal exit tunnel. Once a specific mRNA codon is recognised, ribosomal stalling is induced [**Ito2010**] and translation is halted unless a strong enough pulling force from the downstream insertion is acting on the nascent chain at that time [**Butkus2013**]. Several "strengths" of AP have been identified. For example, SecM from *E. coli* is 17 residues long and relatively weak, whereas a mutated SecM from *Mannheimia succiniciproducens*(Ms-Sup1) is much stronger and 8 residues long ending in a proline which will halt translation [**Ismail2012**]. There are several other SecM proteins of other strengths from various bacterial species [**Yap2009**]. Therefore APs are a technique that can be used to measure precise forces acting on a specific part of the nascent chain during co-transltaional membrane protein integration allowing the study of TMP kinetics during insertion and folding. Indeed the force profile of a single residue can now be obtained *in vivo* [**Ismail2012**]. In an idealised TMH segment composed of alanine and leucine being inserted into *E. coli* membrane through SecM AP with SDS-PAGE, hydrophobicity is more able to overcome the arrest peptide when it is near the N-terminal (of an N-terminal-inside TMH) [**Ismail2012**]. This could be either the TMH finally coming into contact with the cytoplasmic face of the lipid bilayer, or an interaction between the N-terminal and the tip of the lateral gate as previosuly shown in Sec61; part of a pre-integration TMH interrogation [**MacKinnon2014**].

The journey of the TMH through this machinery has been studied using both crosslinking experiments and the relatively new technique of APs [**Cymer2015**].

Accessibility assays and an improved intramolecular crosslinking assay showed that the helical transmembrane S3bS4 hairpin (the paddle) of a voltage-gated potassium (Kv) forms in the ribosome tunnel [**Tu2014**]. Ribosomal folding of the TMHs in Kv1.3, a potassium channel, is maintained in the translocon [**Tu2010**]. Therefore, some of the final structural folded elements of the voltage sensor domain occurs within the ribosomal exit tunnel.

Furthermore, it has recently been suggested that larger structures fold as the ribosomal exit tunnel widens [**Kudva2018**]. This size dependent folding was observed by using the SecM translational AP. Two ribosome mutants were compared (uL23 that is close to the exit tunnel and uL24 deletions which is a hairpin loop that obstructs the tunnel exit.) zinc finger folds deeper in uL23 mutant than wildtype (but not uL24) and a 100 residue domain folds deeper than the uL24 mutant (but not the uL23) [**Kudva2018**].

The ribosomal tunnel also speeds up elongation of neutral and negatively-charged peptides. This is attributed to the sporadic negative patches within the ribosomal exit tunnel [**Lu2008**].

The ribosome clearly has the potential to prefold motifs and small domains before translocon insertion.

4.3 Methods

4.3.1 Datasets

4.3.2 Complexity

4.3.3 Statistics

4.4 Results

4.4.1 There are step changes in TMH complexity depending on the TMH number in GPCRs

GPCR distribution tables for complexity and hydrophobicity

Graphs of complexity and hydrophobicity distributions

Show there are step changes in GPCRs from Bahadur

Supplementary tables for additional stats tests and hydrophobicities

4.4.2 Complexity ascention repeats according to how many TM-bundles are in the protein.

GPCR distribution tables for complexity and hydrophobicity Graphs of complexity and hydrophobicity distributions Show there are step changes in GPCRs from Bahadur Supplementary tables for additional stats tests and hydrophobicities

4.4.3 The pattern is present for GPCR subfamilies

Figure of complexity distributions with Rhodopsin like, Secretin, metabotropic glutamate, Fungal mating, cyclic AMP, Frizzled and smooth. Bahadur tables also

4.4.4 The prevelance of this amongst all TMPs.

Mechano-sensitive (controlled vocabulary if no list available) distributions Voltage gated (controlled vocabulary if no list available) distributions

4.5 Discussion

Seen across a variety of 7TM families with varying functions with datasets built from all membrane types (hence variety). Suggests a pressure for simpler TMHs to precede more complex ones, repeating every 3-4 TMHs. The universality points toward translocon behaviour pressure, or thermodynamic stability in the membrane. Would we expect this behaviour if the translocon acted on only one TMH at a time?

Chapter 5

Conclusions

5.1 Outlook

5.1.1 The hydrophobicity–sequence complexity continuum

We hypothesise 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 uncharacterised 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.