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
May 15, 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 makep around a quarter of all proteins, as well as two-thirds of drug targets, and contain some of the most critical proteins required for life as we know it. Yet they are fundamentally difficult to study experimentally. This is in part due to the very features that make them so biologically influential: their non–polar trans-membrane helix regions. What is missing in the current literature is a nuanced understanding of the complexities of the helix composition beyond a hydrophobic region of around 20 residues. Currently, it is known that the properties of trans-membrane protein α helices underpin membrane protein insertion mechanisms.

By leveraging large datasets of trans-membrane proteins, this thesis is focused on 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 a cellular barrier to separate themselves from the surrounding environment. This cellular skin can be thought of as the bag that contains all the important machinery required for normal cell function. The barrier works by being chemically very different to 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 form 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 with in 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 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 optmisation 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 Trans-membrane Helix Sequence Composition

Measurements of the TMH regions have found that they are roughly 20 residues in length; 17.3±3.1 from 160 TMHs [**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**].

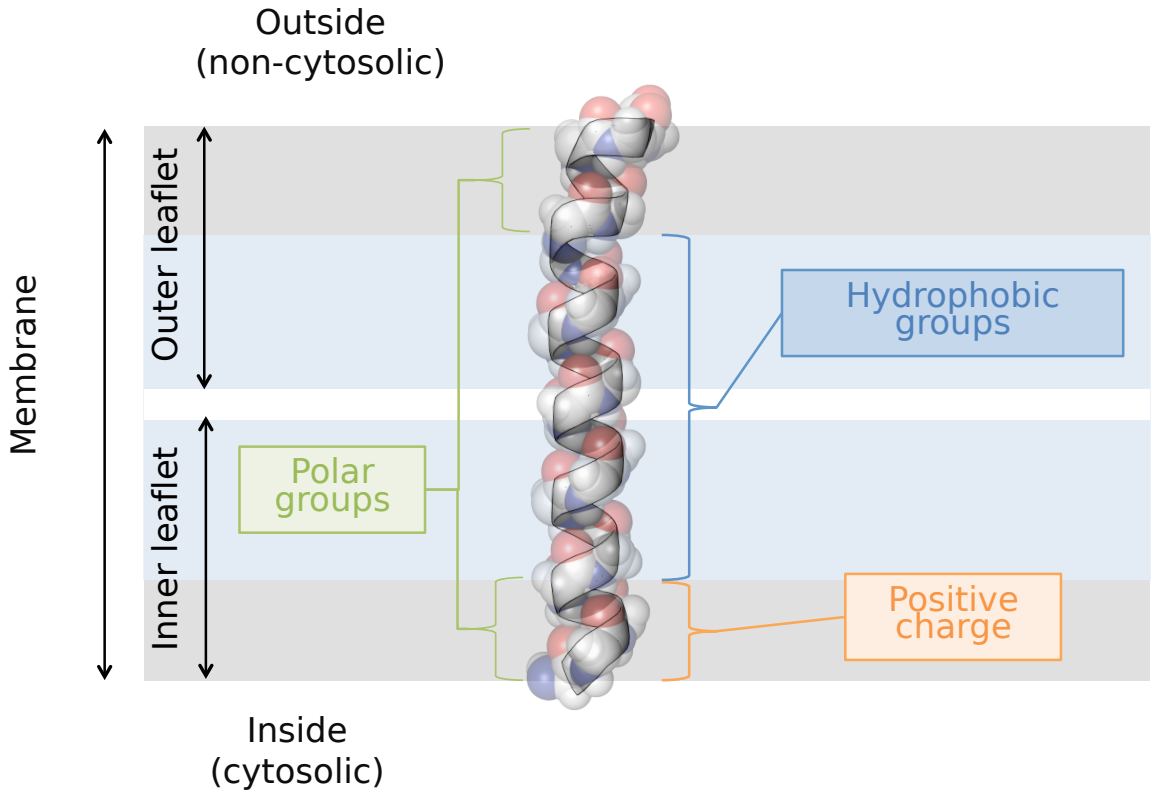

Figure 1.1: A cartoon showing the general components of the membrane and a typical TMH. The example used here for illustrative purposes is the trans-membrane region of therein (Protein Data Bank (PDB) 2LK9) [**Skasko2012**]. Dark grey areas denote the area of lipid head groups. The residues found in 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 ransmembrane (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 evolutionarily 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 trans-membrane 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 seek a compromise. 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.

Snorkeling

Broadly speaking, TMHs are non-polar. However, some contain polar and charged residues in the helix itself. Whilst this might seem thermodynamically unstable at first

glance, a molecular dynamic feature called the "snorkel" effect explains in part how this is possible [**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 headgroups 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, snorkeling, bilayer deformation, and peptide tilting combined to to be sufficient to lower the thermodynamic stability penalty of arginine insertion so that hydrophoic TMHs with a central arginine residue will readily insert into the membrane.

1.1.2 Hydrophobicity of Trans-membrane Segments

Perhaps the most prevalent and important feature of the trans-membrane 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 polypeptide 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 subsegments of the raw region which have the lowest probability. The algorithm searches for and concatenates sub-threshold segments for the Shannon entropy-like term in equation 1.3

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

The lowest probability subsegment can be defined as $K_1 = \log N/L$. By altering the window lengths, and the thresholds SEG can be 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 behavior 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 equilibriums 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 equilibria. Once inserted the protein doesn't leave the membrane as a result of the TMH being very hydrophobic. This hydrophobicity and the hydrophobicity of the lipid tails means that they self-associate and this association is entropically driven by water. Another way of describing it is that they fiercely dissociate from the water. The overall ΔG for a TMH in the membrane is -12kcalmol^{-1} [**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 compartmentalization of cellular biochemistry is arguably one of the most significant events to have occurred in evolution and is certainly one of the fundamental prerequisites for life [**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 trans-membrane α helices are inextricably linked, and often what we observe from the α helices reflect the properties of the much harder to study membranes. The lipid membranes influence the local structure, dynamics, and activity of proteins in the membrane in non-trivial ways [**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**].

Differenes in Membrane Compositions

It has been known for some time that the outer membranes of Gram-negative bacteria are asymmetric in terms of lipid composition. The outer membranes contain 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, proteinlipid interactions have been shown to be determinants of membrane curvature [**Jensen2004**], and undertake complex orientations and conformations to allow for hydrophobic mismatch [**Planque2003**]. It may need changing with every bib.tex update unless the permanent record is changed.

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

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

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 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 Trans-membrane 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.

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

1.4.2 Co-Translational Translocation

The overwhelming majority of TMPs use the co-translational method of translocation. It has long been understood that this method is essentially the Signal Recognition Particle (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 Signal Recognition Particle Receptor (SR) in association with some membrane insertion machinery on the ER [**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 ofSRP [**Shan2005**] revealed key discrete conformational stages. These are the specific recognition of signal sequences on cargo proteins, the targetting 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**].

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 analyze large sequence datasets that have rich topological annotation. By analyzing these sequences in the context of anchorage, we find that some TMHs are confined to biological constraints of the membrane, whereas others that likely contain function beyond anchorage, are less conforming to the membrane. Specifically, there is further elaboration of statistical definitions in the methods than in the published paper.

2.1 Abstract

2.2 Summary

As the idea of positive residues inside the cytoplasm emerged during the late 1980s, so did the idea of negative residues working in concert with TMH orientation. It was shown that removing a single lysine residue reversed the topology of a model *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 traveling through the SecYEG translocon indicating that negative charges are biologically relevant [**Ismail2015**]. In this chapter, we explore the literature surroudning 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 center 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 multi-pass 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 luminal 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 subcellular 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 localization. Datasets 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

Dataset	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.

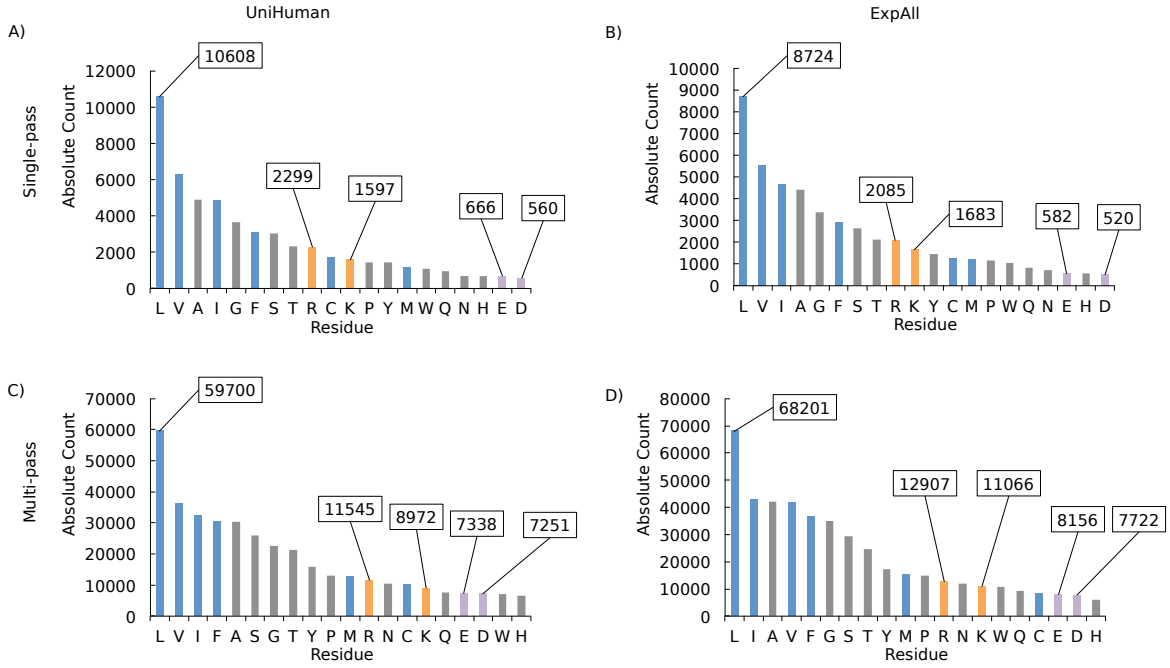

Figure 2.1: Negatively charged amino acids are amongst the rarest residues in TMHs and ±5 flanking residues. Bar charts of the abundance of each amino acid type in the TMHs with flank lengths of the accompanying ±5 residues from the (a) UniHuman single-pass proteins, (b) ExpAll single-pass proteins, (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 subcellular 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

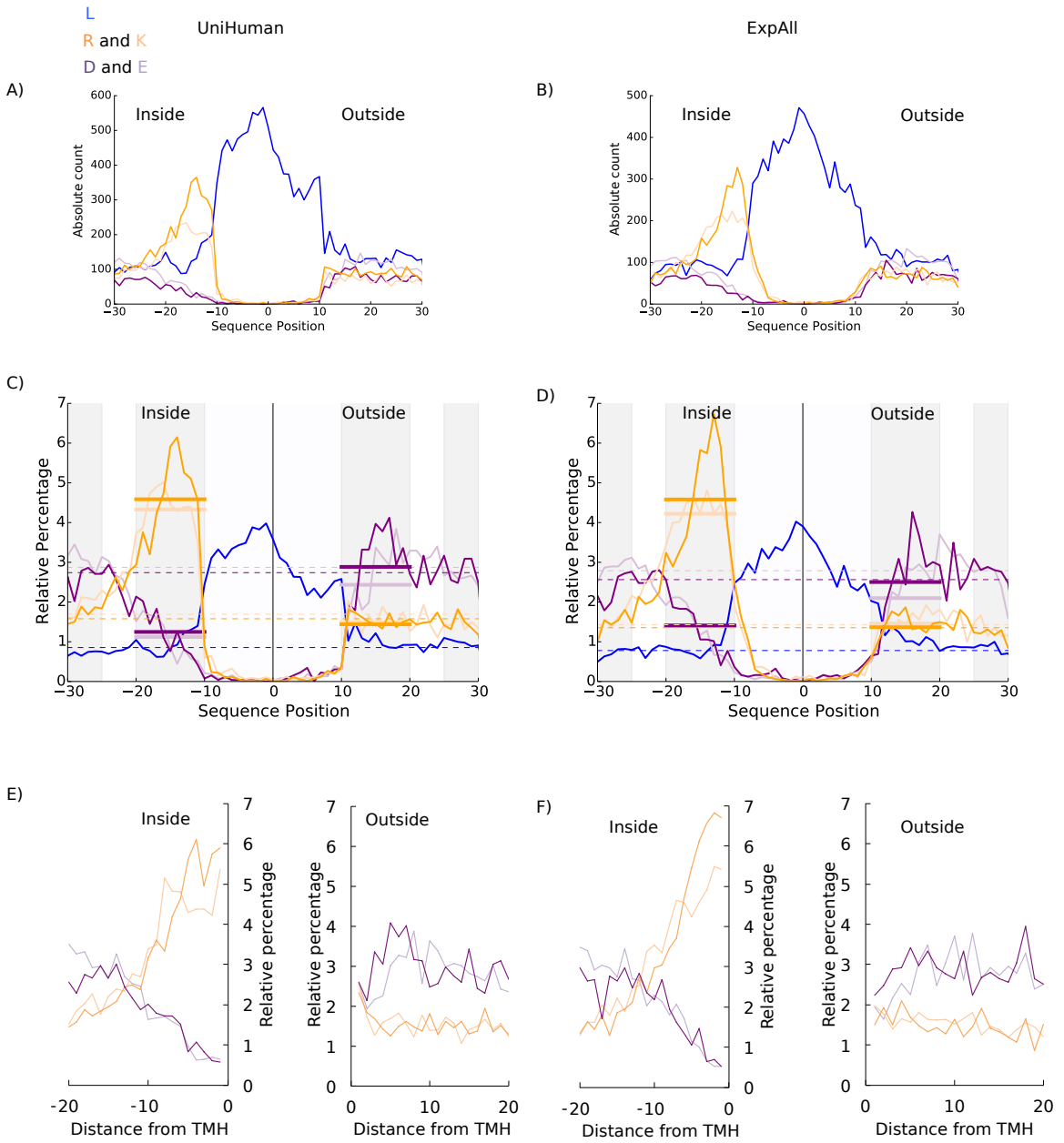

Figure 2.2: Relative percentage normalisation reveals a negative-outside bias in TMHs from single-pass protein datasets. All flank sizes were set at up to ±20 residues. We acknowledge that all values, besides the averaged values, are discrete, and connecting lines are illustrative only. On the horizontal 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**			
Dataset	**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,

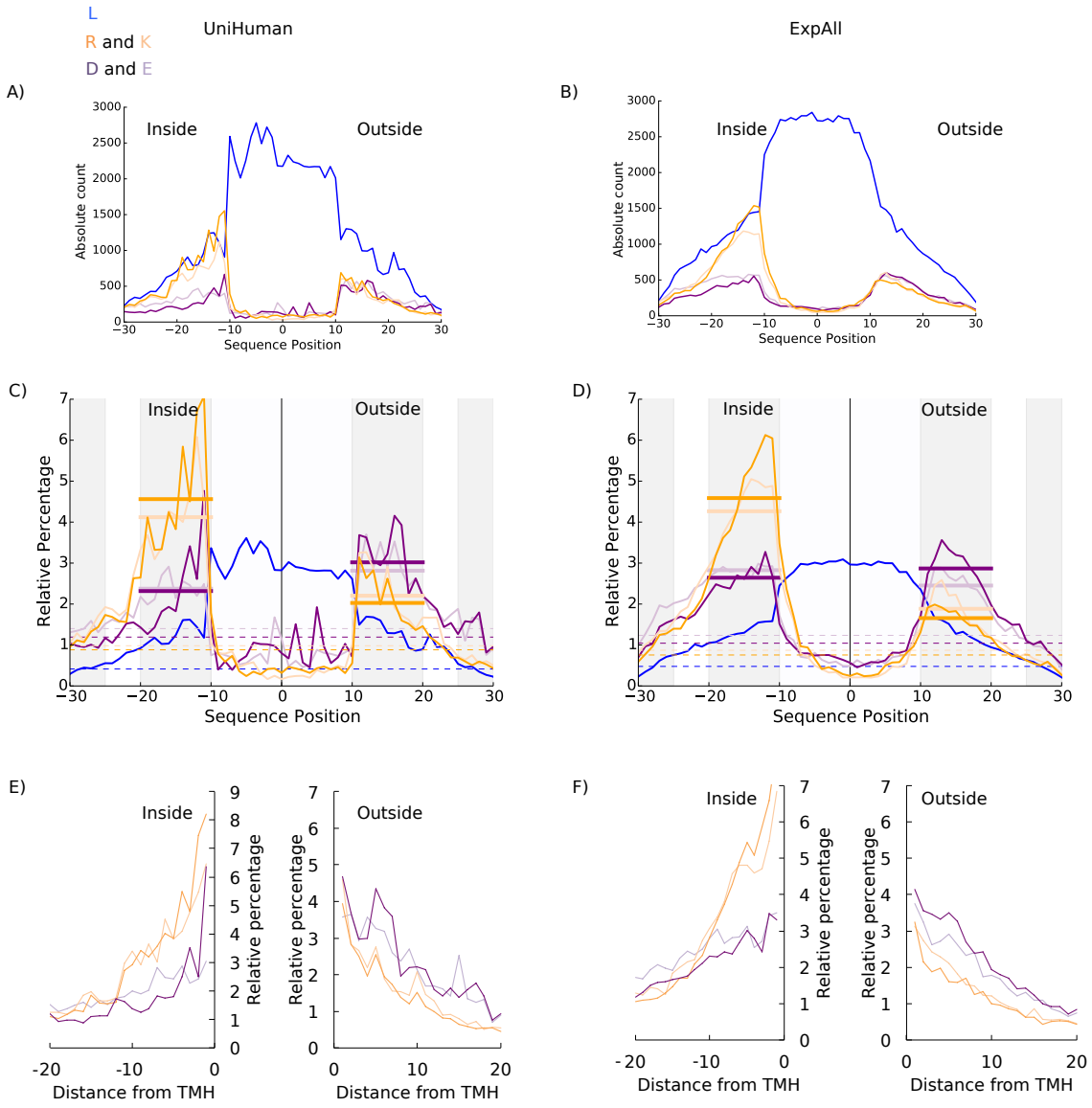

Figure 2.3: Negative-outside bias is very subtle in TMHs from multi-pass proteins. The meaning for the horizontal axis is the same as in Figure 2.2, with the negative sequence position numbers extending towards the cytoplasmic space. Leucine is in blue. Arginine and lysine are shown in dark and light 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).

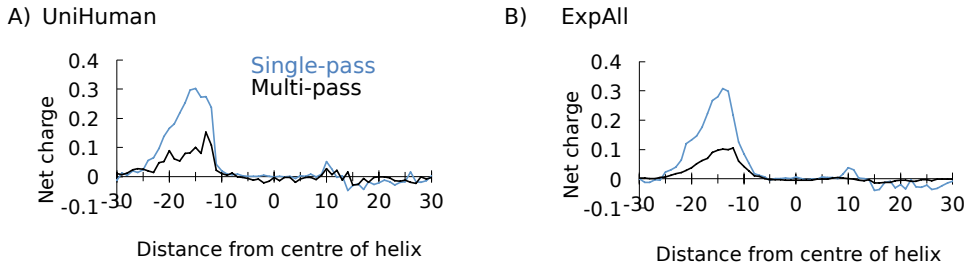

Figure 2.4: The net charge across multi-pass and single-pass TMHs shows a stronger positive inside charge in single-pass TMHs than multi-pass TMHs. The net charge per TMH plotted at each position; the positive-inside rule is stronger in TMHs from single-pass proteins than TMHs from multi-pass proteins. 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)

	Multi-pass					Database-defined flanks				Flanks after central alignment			
	Dataset	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
(A)	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

	Multi-pass	Overlapping flanks				Database-defined viable* flanks				
	Dataset	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
(B)	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 (normalization 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 heatmap. 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

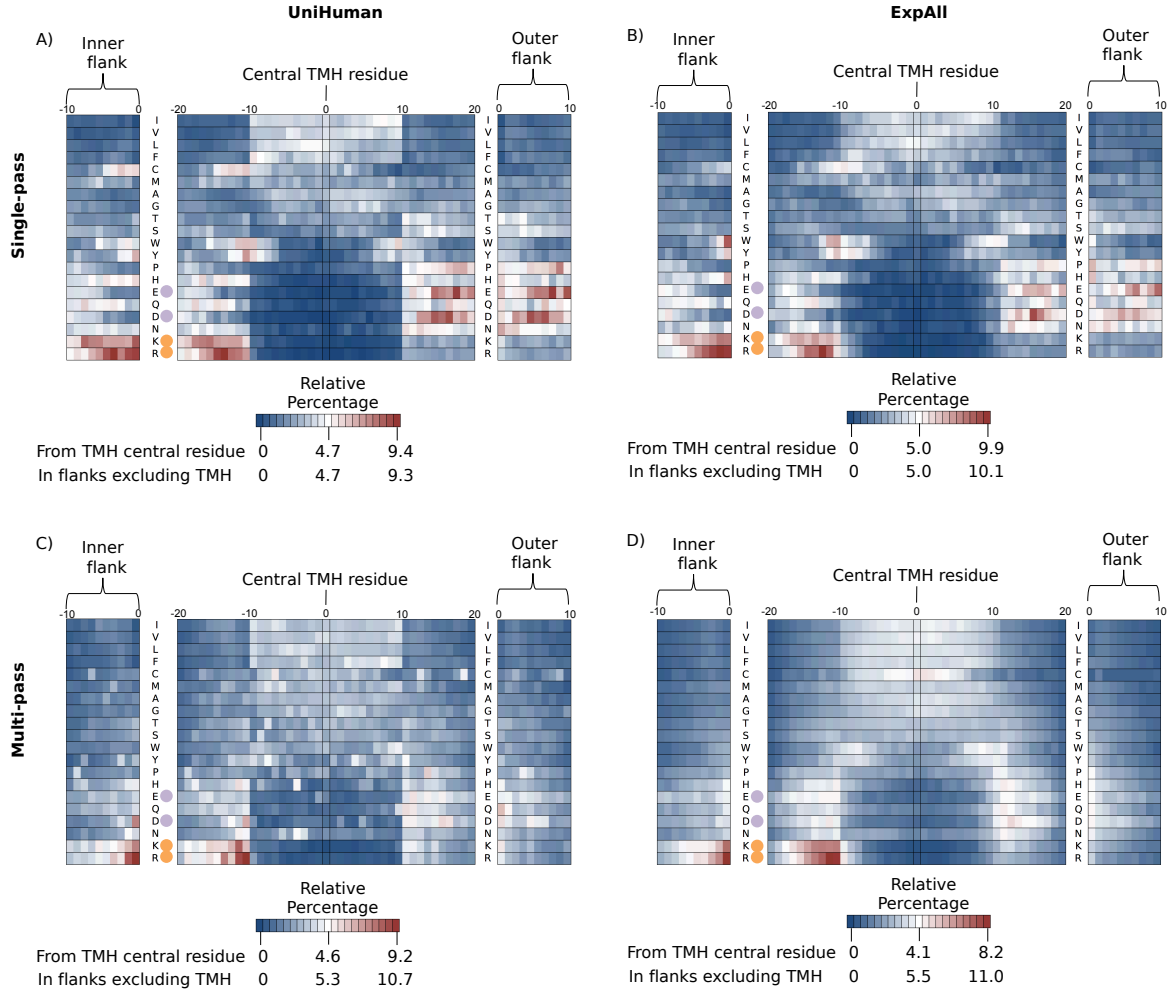

Figure 2.5: Relative percentage heatmaps from predictive and experimental datasets corroborate residue distribution differences between TMHs from single-pass and multi-pass proteins. The residue position aligned to the centre of the TMH is on the horizontal axis, and the residue type is on the vertical 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 heatmap 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).

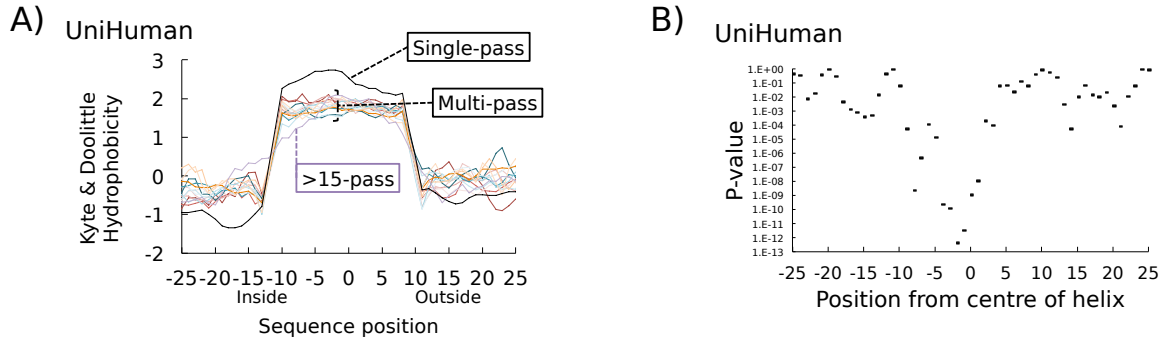

Figure 2.6: There is a difference in the hydrophobic profiles of TMHs from single-pass and multi-pass proteins. a The hydrophobicity of single-pass TMHs compared to multi-pass segments from the UniHuman dataset. The Kyte and Doolittle scale of hydrophobicity [**Kyte1982**] was used with a window length 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)

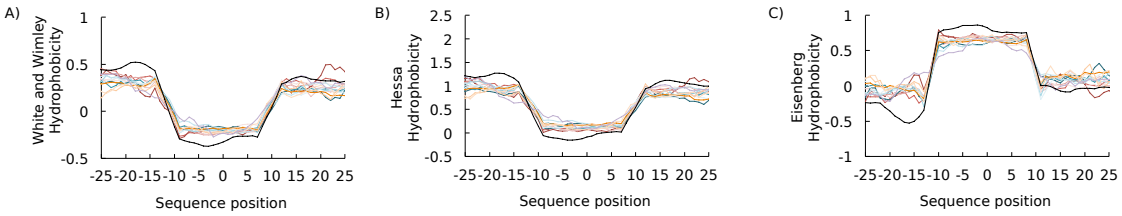

Figure 2.7: There is a difference in the hydrophobic profiles of TMHs from single-pass and multi-pass proteins. The difference in hydrophobicity between the single-pass and multi-pass datasets stratified by number of TMHs is not due to the choice of scale. As with Figure 2.6, UniHuman was stratified 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

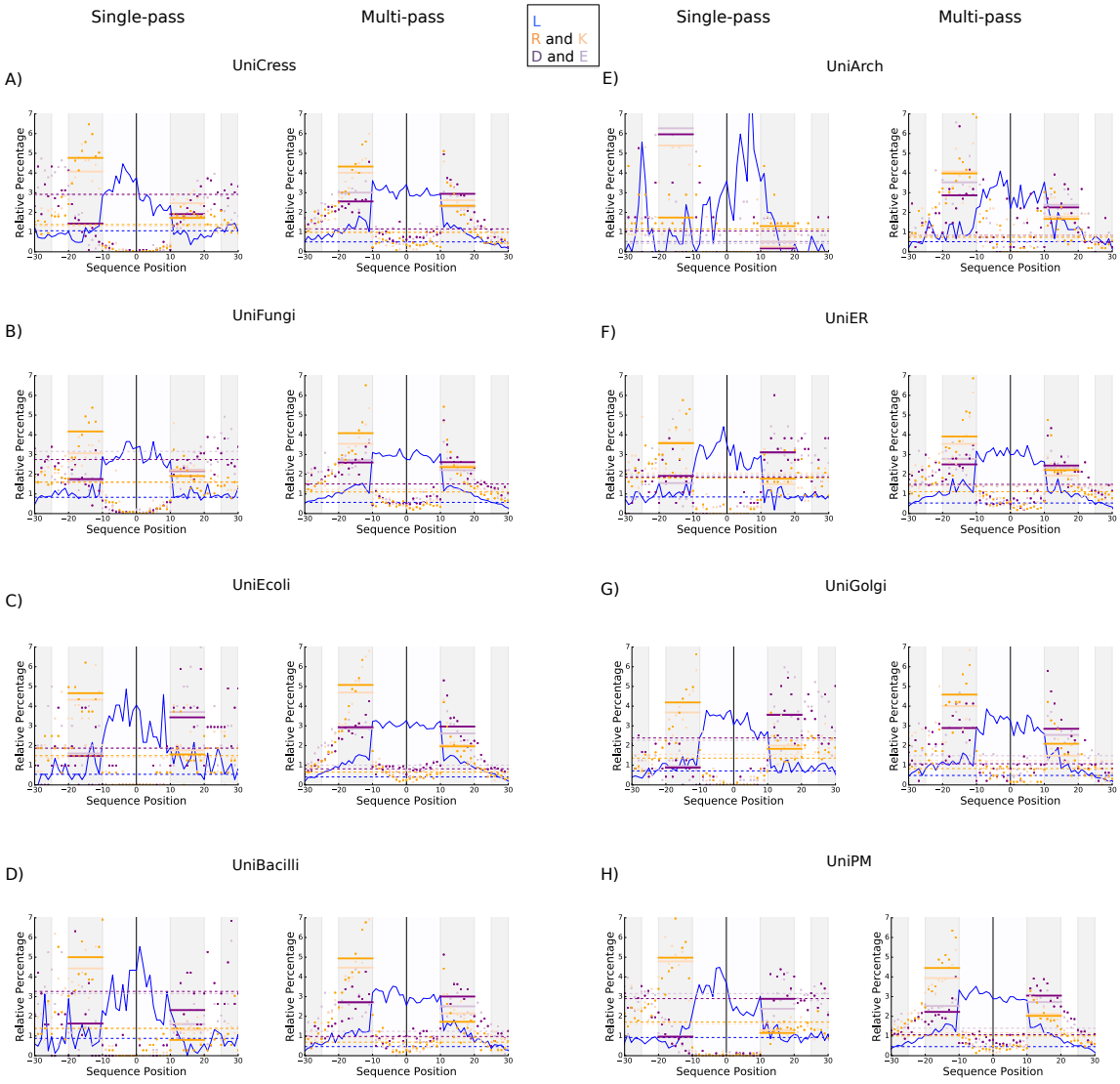

Figure 2.8: Comparing charged amino acid distributions in TMHs of multi-pass and single-pass proteins across different species and organelles. The relative percentage distribution of charged residues and leucine was calculated at each position in the TMH with flank lengths of ±20 in different datasets. 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,

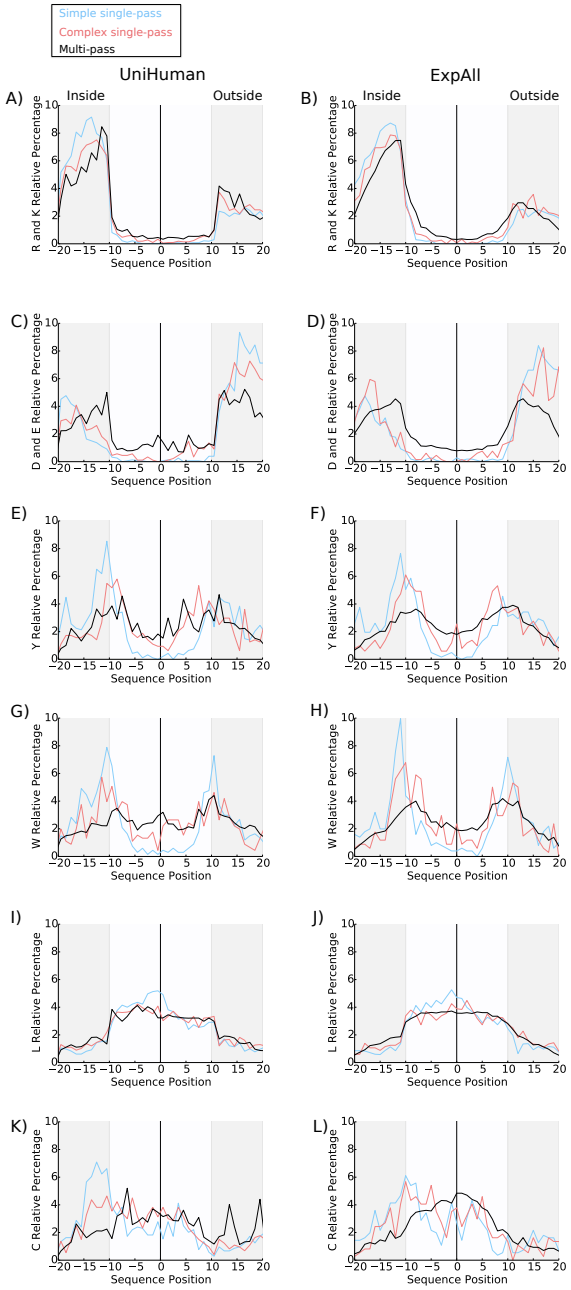

Figure 2.9: Comparing the amino acid relative percentage distributions of simple and complex TMHs from single-pass proteins and TMHs from multi-pass proteins. Comparing the amino acid relative percentage distributions of simple and complex TMHs from single-pass proteins and TMHs from multi-pass proteins. 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 localizations, 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 standardized 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 benefitted from the central alignment approach and the first two studies from another normalization 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.

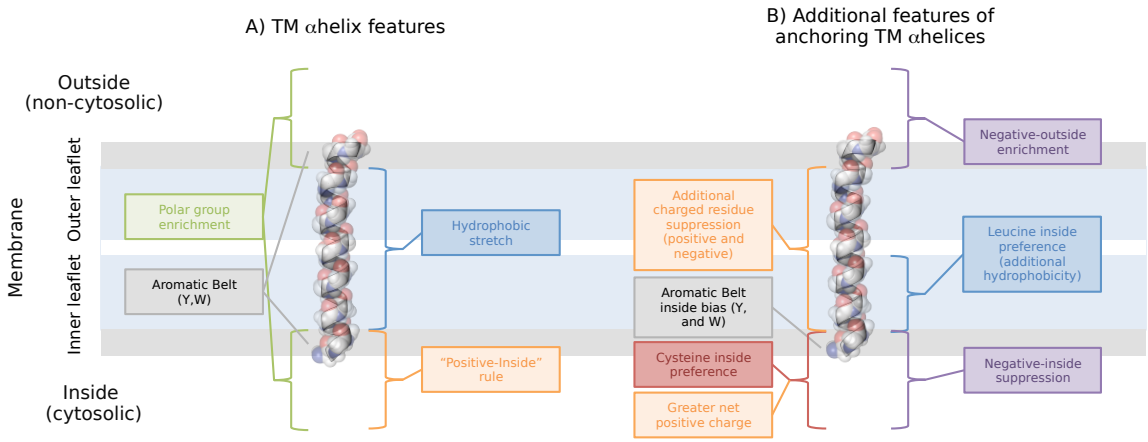

Figure 2.10: Residue distributions of transmembrane anchors. A view showing additional residue distribution features that TMHs with an anchorage function display.a The more classic model of a TMH showing the "positive-inside" rule [**VonHeijne1989**], the hydrophobic core [**Kyte1982**], the polar enrichment 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 localization 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	313

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 etal. [**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).

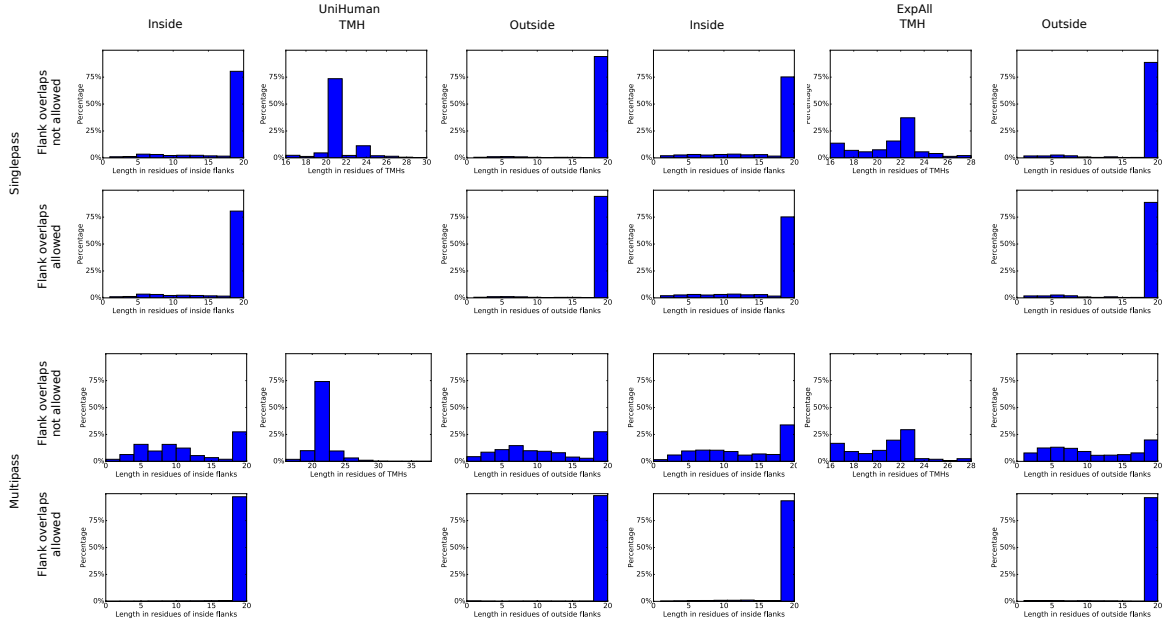

Figure 2.11: The lengths of flanks and TMHs in multi-pass and single-pass proteins in the UniHuman and ExpAll dataset. On the horizontal axis are the lengths of the TM segment regions in residues. On the vertical axis are the percentages of the population. There are three regions: the inside flank, the 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 emphasizing 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 normalization reveals the most preferred residue types at given sequence positions.

Our second normalization 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 is the total abundance of residues of just amino acid type 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 as function of r reveal the prefered positions of residues of type i. The difference in and 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 residuescale [**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:

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

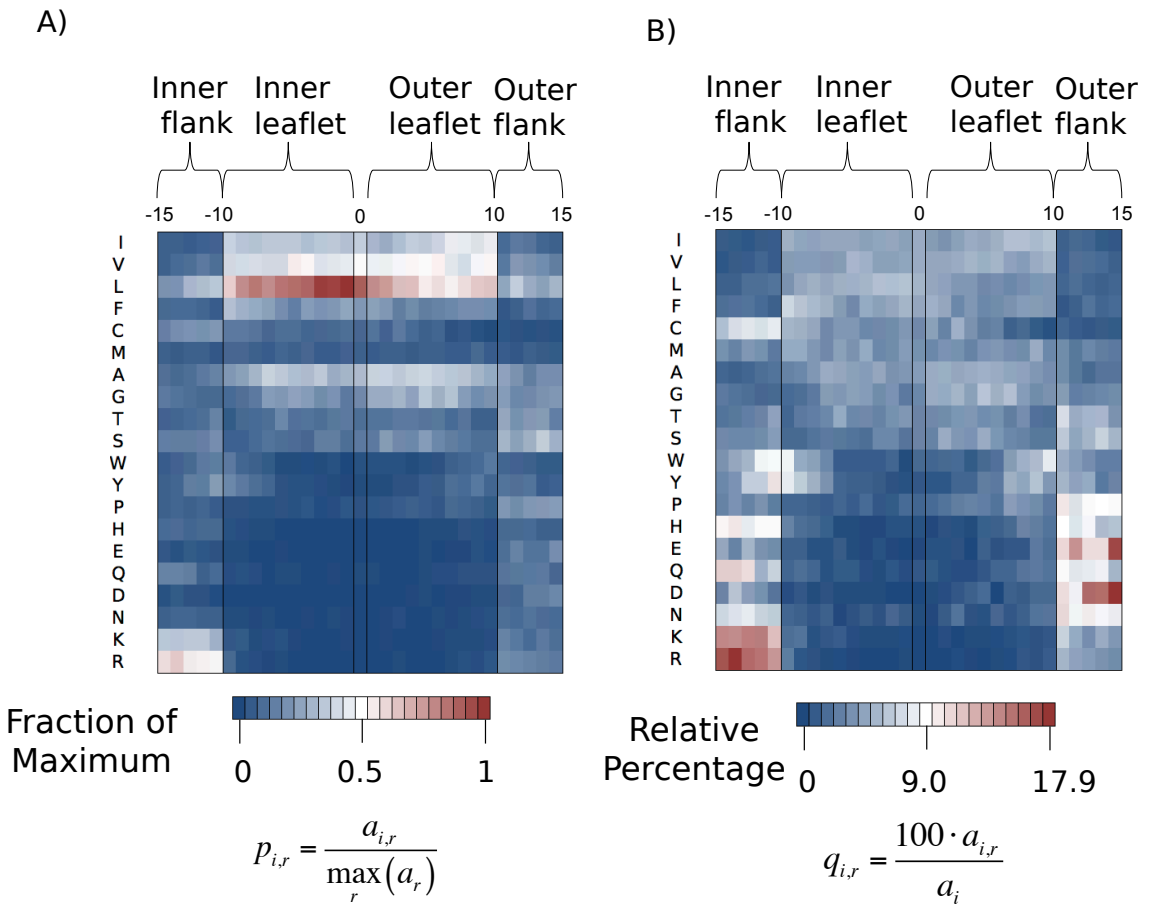

Figure 2.12: Relative percentage heatmaps from the predictive datasets calculated by fractions of the absolute maximum and by the relative percentage of a given amino acid type. The residue position aligned to the centre of the TMH is on the horizontal axis, and the residue type is on the vertical axis. 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.

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. ope

Chapter 3

Tail-Anchored Protein Datasets

3.1 Abstract

3.2 Introduction

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

The pipeline generates a list of singlepass proteins with a transmembrane domain close to the C terminal, that are not splice isoforms. A previous study by Kalbfleisch *et al.* published in Traffic 2007 (8: 1687-1694) predicted 411 tail anchor proteins [**Kalbfleisch2007**]. The tools developed herein are openly available for reapplication to other datasets. Notably, known SNARE transmembrane helices are highly hydrophobic even compared to other TA transmembrane helices. We compare Kyte and Doolittle hydrophobicity profiles of our filtered human protein list against the profiles of previously known SNARE and TA proteins. This provided a list of potential SNARE proteins in addition to potential spontaneously inserting TA proteins similar to cytochrome b5 which have the least hydrophobic transmembrane helices.

Tail-anchored proteins are a topologically distinct class of intracellular proteins

defined by their single carboxy-terminal transmembrane domain with a cytosolic-facing amino-terminus.

Tail-anchored proteins are involved in a range of key cellular functions including protein translocation and apoptosis. Additionally, within the tail-anchored class of proteins are a set of vesicle fusion proteins called SNARE proteins. There is biomedical interest in SNARE drug delivery mechanisms.

SNAREs can fuse liposomes containing various drug payloads into the membrane. Notably, known SNARE TMHs are highly hydrophobic even compared to other tail anchored TMHs [**Kalbfleisch2007**]. This hydrophobicity appears to be a determinate factor in the precise delivery mechanistic route that a TA proteins use for insertion [**Rabu2008**, **Rabu2009**], for which there is evidence demonstrating that are several mechanisms [**Rabu2009**, **Johnson2013**].

Whilst most eukaryotic TA proteins are inserted into the ER.

3.3 Results

3.3.1 Tail-Anchored Protein Datasets Are A Moving Target

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 21119 TMH residues (28791 including uptt to ±5 flanking residues) from 987 UniProt protein records.

We compared redundant versions of these two datasets to the S1 dataset from a previous method [**Kalbfleisch2007**], that aimed to gather TA proteins in the human genome from the NCBI, to see how many records are shared, how many are now obselete, and how many are unique.

Figure 3.1 shows that already a study from 2007 [**Kalbfleisch2007**] has 175 record ids of 222 records (78.8%) that do not share overlap the up-to-date manually curated UniProt dataset [**TheUniProtConsortium2014**]. Of the 166 unique records of that 2007 dataset, 92 records do have location annotation in UniProt that the scripts herein

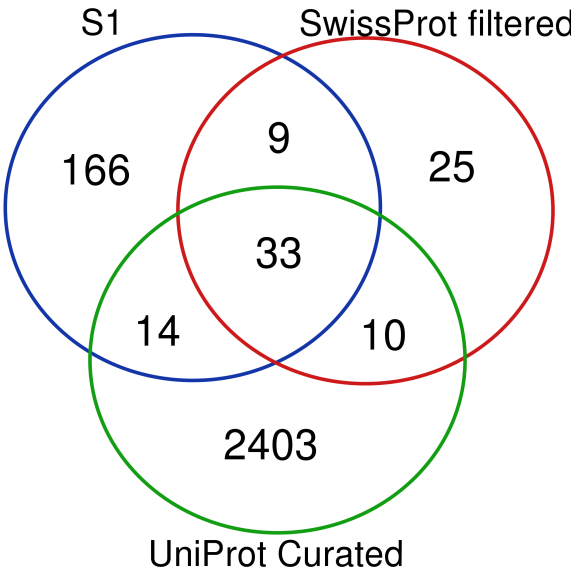

Figure 3.1: A venn diagram showing tail anchored protein UniProt ids present in each of the datasets as well as those present in multiple datasets. The number of ids present in redundant versions of i) the supplementary materials table of a previous study predicting the complete set of human tail anchored proteins denote by S1 [**Kalbfleisch2007**], ii) the Swissprot dataset filtered according to typical TA features limited to the human proteome [**TheUniProtConsortium2014**], and iii) 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 later statistical analysis.

use for topological determination, leaving 74 records without location annotation. This leaves 92 of 222 (41.4%) records that originally fitted criteria that no longer fit those same criteria. If we exclude those lacking suitable annotation i.e ids from S1 that are found in either SwissProt with the filters (9), the curated UniProt list (14), or both (33), compared to the 92 that have annotation contradicting the original prodictions, 37.8% of the ids overlap.

Equivalent criteria were applied to the entire SwissProt database and then restricted to the human proteome dataset. 43 of these 77 records (55.8%) are in the curated UniProt TA dataset leaving 34 records that meet the criteria out of the manually curated set (44.2%of the filtered Swissprot dataset). 42 of the 77 (54.5%) records from SwissProt filtered human dataset can be found in the original S1 list. A further consideration is that, after removing redundant proteins, this method picked up 46 Archaeal and 66 bacterial records.

Datasets are a moving target as they are constantly updated with more accurate and reliable tools. Perhaps unsurprisingly, as a trend, this shows that up-to-date datasets improve the reliability of this automated predicted method and that there is

a large degree of what we now believe to be mistakes that occured in older prediction tools. These automated criteria still do not fully align with the manually curated list, which is bound to change too, especially considering only 973 of a non-redundant (90% CD-HIT threshold [**Huang2010**, **Wu2011**]) set of those 2460 proteins of the UniProt manually curated set contained annotation for the transmembrane boundary residues.

Note that these numbers are not absolutely certain. 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 list that excluded proteins that were either hypothetical or splice isoforms. That being said, this "conversion" step prevents us from directly comparing the entire original S1 dataset.

3.3.2 Species Variation

When comparing the average Kyte & Doolittle [**Kyte1982**] hydrophobicity values for the TMHs from humans and mice, *A. thaliana*, and *S. cerevisiae*, we see little difference between the mean values all at $\tilde{2}$.3-2.6 with the TMH, and at $\tilde{1}$.3-1.6 when considering residues in close proximity to the TMH ($\pm$5 residues)(Figure 3.1). Indeed, we see no strong observable statistical differences in hydrophobicity ($P - value > 8.30E - 02$ in the UniProt curated list, $P > 3.35E - 01$ in the SwissProt automatically filtered list)(Table 3.2).

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 that can be observed in terms of hydrophobicity [**Sharpe2010**], especially in TMH anchors [**Baker2017**]. However, in these TA protein datasets we see no such patterns at this sample size.

Here, we are dealing with datasets at least an order of magnitude smaller than those broad studies which could explain the absence of the effect, however this only goes to show that if there is an effect in TA proteins, it is indeed weak.

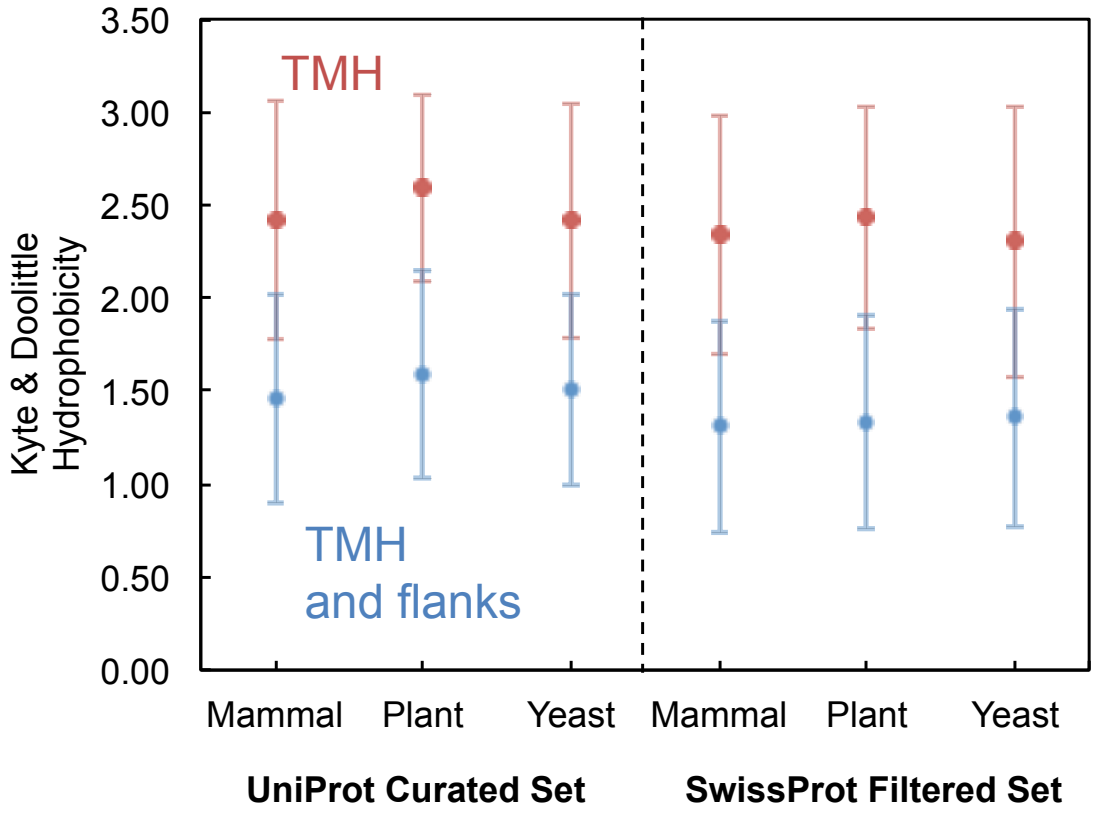

Table 3.1: Average values of species datasets from UniProt manually curated set and SwissProt automatically filtered dataset.
The average hydrophoicity values from the Kyte & Doolittle scale [**Kyte1982**] and the GlobProt scale [**Linding2003**] for both the TMH and the TMH±5 residues. Values are shown 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=38 and Mouse N=37) to *A. thaliana* (N=60) representing plants and *S. cerevisiae* (N=31) 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.

3.3.3 Organelle Membrane Variation

3.3.4 Spontaneous insertion may be achieved by polar patches in the TMH

In addition, predicted insertion machinery dependent TAs with TMHs that are more polar on average than spontaneously inserting TA protein cytochrome b5 have been highlighted.

Table 3.2: 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=38 and Mouse N=37) to *A. thaliana* (N=60) representing plants and *S. cerevisiae* (N=31) 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 predicting the boundary of TMHs is difficult, according to the Kyte & Doolittle hydrophobicity scale [**Kyte1982**]. Disorder was calculated in the same way using the GlobProt scale [**Linding2003**]. The score column refers to the statistical score obtaineed 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. The Bahadur column refers to the Bahadur slope, an interpretation of the P-value that accounts for the sample size powering the test [**Bahadur1967**, **Bahadur1971**].

3.4 Discussion

Given the large biochemical distinction between TA proteins with different terminal destinations, it is tempting to conclude that TMHs contain the necessary biological factors to determine their targetting. It is indeed possible that our observations are adaptations to the membrane environment, and this would not be unreasonable, except that TA proteins would be expected to experience similar adaptations at a species level, for which at this sample size such an effect is unobservable. Here, we postulate that there is indeed biological information held within the TMH itself that allows the protein to be specifically targetted to a destiantion. This is almost certainly aided by other factors and is part of a system with several redundant mechanisms.

Signal anchored proteins, proteins that contain a single hydrophobic segment that serves as both a mitochondrial targeting signal and a membrane anchor, as well as tail-anchored proteins have been shown to be able to spontaneously insert into the membrane independently from the translocon [**Elisa2012**, **Lan2000**, **Colombo2009**]. 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**].

3.5 Methods

3.5.1 Building a List of Tail-Anchors

Steps carried out by Kalbfleisch *et al.* published in Traffic 2007 (8: 1687-1694) [**Kalbfleisch2007**], were recreated using up to date tools. Whilst their study

focussed on the human proteome, here we take into account the entire TrEMBL and Swiss-Prot database and then stratify the datasets by the organism at the end of the pipeline.

Swiss-Prot Tail Anchored Dataset According to Filters

There were 557012 protein records downloaded from Swiss-Prot 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 robust meta-analysis of TMH prediction using TMHMM [**Krogh2001**], Memsat [**Jones2007**], 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 glstmhs 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 and their coworkers [**Kalbfleisch2007**]. Running CD-Hit 4.5.3 on the WebMGA webserver [**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]")`.

TrEMBL Tail Anchored Dataset According to Filters

111425234 records were storeed in the TrEMBL database at 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. 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 15 residues of the C terminal residue. In those record's annotations, no more than 1 appeared in any given species, so they were omited from the Swiss-Prot list sequence redundancy protocol to avoid representing a well annotated record with a poorly annotated record.

UniProt Curated List

A query for `locations:(location:"Single-passtypeIVmembraneprotein[SL-9908]")` was used in UniProt which returned 2460 UniProtKB IDs; 463 Swiss-Prot results and 1997 TrEMBL results. Running these records through CD-HIT at 90% redundancy yielded 309 Swiss-Prot records and 808 TrEMBL records [**Huang2010**, **Wu2011**]. Of those, 987 proteins from 973 records (308 from Swiss-Prot, and 665 from TrEMBL) had the `TRANSMEM` annotation indicating a bone fide TMH. No further filters were applied to this list. Proteomes represented by more than 20 records include *A. thaliana* (60 records), Humans (38), Mouse (37), and *S. cereveisiae* (31).

401 were annotated under the Mitochondrion location (query `locations:(location:"Mitochondrion[SL-0173]")`) 39 from Swissprot and 362 automatically assigned in TrEMBL. 401 in the ER (query `locations:(location:"Endoplasmicreticulum[SL-0095]"`), 98 from SwissProt and 303 automatically annotated in TrEMBL. 1 TrEMBL record (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 annoted SwissProt records; 37 in the PM (query `locations:(location:"Cellmembrane[SL-0039]")`, and 82 in the Golgi (query `locations:(location:"Golgiapparatus[SL-0132]")`).

Remapping Previous Dataset

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

3.5.2 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. 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.

3.5.3 Calculating Sequence Entropy

Sequence 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 anchorage of a TMH [**Wong2011**, **Wong2012**, **Baker2017**]. here we 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 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 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**].

3.5.4 Statistics

The null hypothesis of homogeneity of two distributions was examined with the KS, the KW and the 2-sampled T-test statistical tests. These tests were all ran through the Python scipy stat package v0.17 python 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 a permutated set that exponentially increases as N increases, the P-value is a strong function of N. We rely on the (absolute) Bahadur slope (B) as a measure of distance between two distributions [**Bahadur1967**, **Bahadur1971**, **Sunyaev1998**, **Baker2017**]:

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

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

Chapter 4

Identifying Intramembrane Folds Using Sequence Complexity

4.1 Abstract

4.2 Introduction

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

GPCRs have long be known to be overrepresented among genomes [**Remm2000**].

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