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 1, 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 [1]. Trans-membrane Protein (TMP)s include some of the most critical to life proteins as well as a large number of drug targets. However, the experimental inaccessibility of the Trans-membrane Helix (TMH) has hampered 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 [2], 27.1±5.4 residues based on 129 TMHs [3], 26.4 residues based on 45 TMHs [4], 25.3±6.0 residues based on 702 TMHs [5], 24.6±5.6 from 837 TMHs [6], and 28.6±1.6Å to 33.5 ±3.1Å from 191 proteins depending on membrane types [7]. 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 [8]. Secondly is that it is emerging that different membranes have different thicknesses [9], and that this is directly reflected in the hydrophobic lengths of the TMH [7, 10].

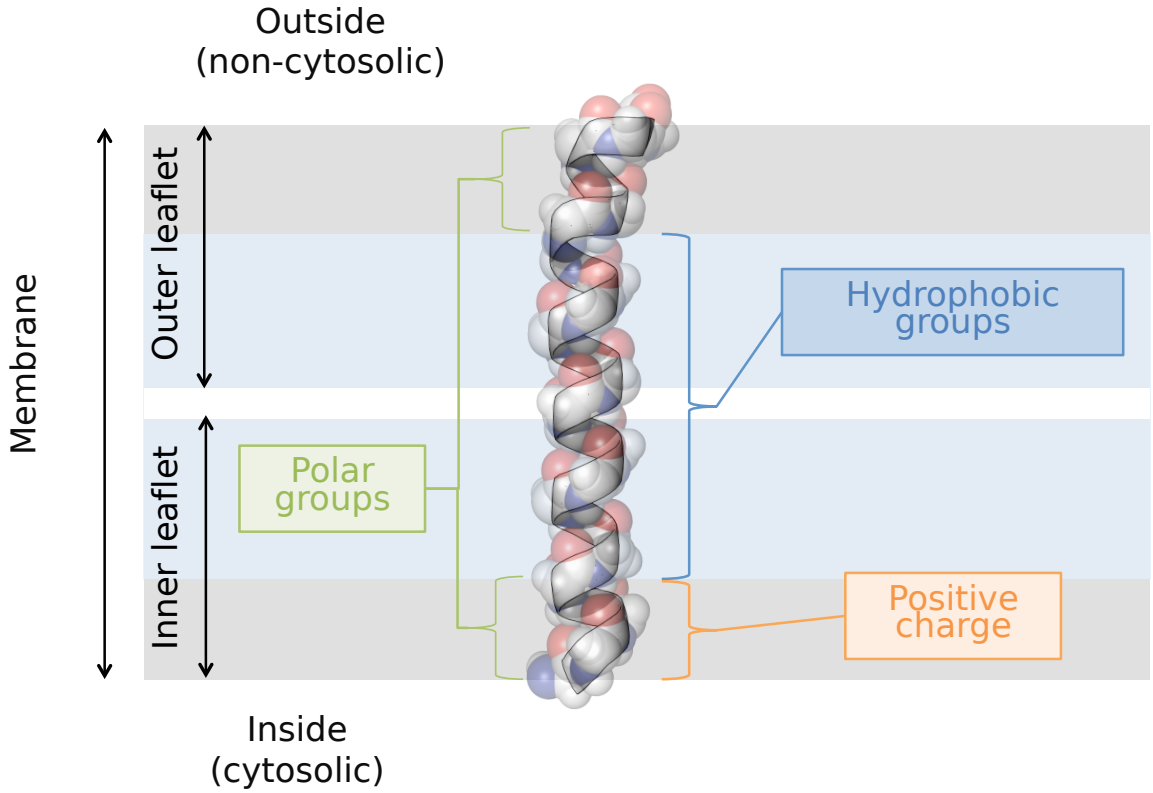

Figure 1.1: A cartoon showing the general components of the membrane and a typical TMH. The example used here for illustrative purposes is the trans-membrane region of therein (Protein Data Bank (PDB) 2LK9) [11]. Dark grey areas denote the area of lipid head groups. The residues found in these areas 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" [6, 7, 10].

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 [10].

A study by Baeza–Delgado *et al.* from 2013 [6] looked at TMHs in 170 integral membrane proteins from a manually maintained database of experimentally confirmed TMPs; MPTopo [12]. 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 [6]. 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 [6]. This is probably reflecting the different lipid composition of both leaflets of biological membranes [6].

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 [10]. 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 [10].

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 [13, 14]. 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 [6, 7, 10].

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" [6, 10, 15–17]. Not all aromatic residues are not found in the aromatic belt; phenylalanine has no particular preference for this region [16, 18]. However, it still remains unclear if this is to do with anchorage or translocon recognition [6].

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 [19]. 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 [20, 21]. 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 [22]. This has also been suggested to allow helices to adapt to

varying thicknesses of the membrane [23]. 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 [24]. 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 [25, 26].

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 [27, 28]. Simm *et al.* 2016 [27] 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 [29].

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 [28]. 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 [30].

Hessa's Biological Hydrophobicity Scale

This is arguably the most biologically relevant scale [28], 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 [15]. 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 [31].

White and Wimley Octanol – Interface Whole Residue Scale

This scale is calculated from two other scales; the octanol scale, and the interface scale [32]. 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 [33], Wolfenden [34], Chothia [35], Janin [36], Wolfenden [37], and the von Heijne scale [38]. The scales are normalised according to serine [39]. The automatic TRANSMEM annotation currently used in Uniprot is according to TMHMM [40], Memsat [41], Phobius [42] and the hydrophobic moment plot method of Eisenberg and coworkers [39].

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 [26]. 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 [26]. 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 [43]. Φ is the White and Wimley hydrophobicity [32] 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 [44–46], 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 [47].

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 [43, 48] 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 [8, 49]; 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 [1].

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 [29, 50, 51]. 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 [52], and crucially this function beyond anchorage can be revealed by sensitive, careful analysis of the sequence information alone [26].

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 [53].

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 [54].

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 [55–57]. 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} [29]; 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 [58]. 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 [59–66], as well as protein folding [67].

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 [68].

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 [10]. Furthermore, protein-lipid interactions have been shown to be determinants of membrane curvature [65], and undertake complex orientations and conformations to allow for hydrophobic mismatch [69]. 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 [70]. 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 [70]. 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 [71, 72]. Negativity was found in the mitochondrial matrix with a voltage across the mitochondrial membrane at 150mV [73]. No notable membrane potential has been identified in the Golgi [74, 75].

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 [15, 76].

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 [77]. Mutant studies ofSRP [77] 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 [78].

1.4.3 Post-Translational Translocation

Tail-Anchored Proteins Post-Translationally Insert

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 Soluble N-Ethylmaleimide-Sensitive Factor Attachment Receptor (SNARE) proteins. There is biomedical interest in SNARE drug delivery mechanisms.

SNAREs can fuse liposomes containing various drug payloads into the membrane. Notably, known SNARE TMHs are highly hydrophobic even compared to other tail anchored TMHs [79]. This hydrophobicity appears to be a determinate factor in the precise delivery mechanistic route that a Tail Anchor (TA) proteins use for insertion [80, 81], for which there is evidence demonstrating that are several mechanisms [81, 82].

Whilst most eukaryotic TA proteins are inserted into the ER.

1.4.4 Translocon Independent Membrane Insertion

Signal anchored proteins, proteins that contain a single hydrophobic segment that serves as both a mitochondrial targeting signal and a membrane anchor, as well as tail-anchored proteins have been shown to be able to spontaneously insert into the membrane independently from the translocon [50, 83, 84].

It is postulated that there are electrostatic factors in the flanking regions that contribute to this spontaneous membrane insertion. Our experimental collaborators in Stephen Highs group are interested in a small group of tail-anchored proteins that have very polar trans-membrane domains and are capable of liposome membrane insertion without insertion machinery, also known as spontaneous insertion. They have found that chimeric synaptobrevin, one of the first identified SNARE proteins, is capable of spontaneous insertion if the tail anchor domain is replaced by the TM domains belonging to a protein of known spontaneously inserting domains. Their studies have moved the focus of spontaneous insertion away from the loop regions and onto the physicochemical factors of the TMH itself. The idea that SNARE proteins are modular and capable of spontaneous insertion has significant implications for both biomedical application in liposome-based drug delivery and can aid future research for testing complex biological molecular networks [85, 86].

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 [87]. 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 [88]. 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 [6, 7, 10, 16, 17]. However, in *E. coli* negative residues do experience electrical pulling forces when traveling through the SecYEG translocon indicating that negative charges are biologically relevant [51]. 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 [8, 49]. The insertion and formation of the TMHs follow a complex thermodynamic equilibrium [29, 89, 90]. 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 [52]. 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 [26, 53, 91].

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 [1]. 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 [6, 10, 13, 16, 92, 93]. Such properties vary somewhat in length and intensity between various biological organelle membranes,

between prokaryotes and eukaryotes [94] and even among eukaryotic species studied due to slightly different membrane constraints [7, 10]. These biological dispositions are exploitable in terms of TM region prediction in query protein sequences [95, 96] and tools such as the quite reliable TMHMM [40, 97], Phobius [42, 98] or DAS-TMfilter represent todays prediction limit of TMHs hydrophobic cores within the protein sequence [99–101]. 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 [102–104] and thus, very valuable information.

The "positiveinside rule" reported by von Heijne [8, 13] 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 [13, 105]. As a trend, the "positive-inside rule" has since been confirmed with statistical observations for most membrane proteins and biological membrane types [6, 17, 106, 107]. 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 [7].

The positive-inside rule also received support from protein engineering experiments that revealed conclusive evidence for positive charges as a topological determinant [13, 88, 108, 109]. 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 [15].

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 [110].

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 [6, 10].

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 [88]; nevertheless, a sufficiently large negative charge can overturn the positive-inside rule [111, 112] and, thus indeed, negative residues are topologically active to a point. Negatively charged residues were observed in the flanks of TMHs [6], especially of marginally hydrophobic TM regions [113]. 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 [51, 114].

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 [105, 115]. 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 [6,

7, 10, 14, 16, 17].

There are many observations described in the literature that charged residues determine topology more predictably in single-pass proteins than in multi-pass TMH [112, 116]. 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 [117].

With sequence-based hydrophobicity and volume analysis and consensus sequence studies, Sharpe *et al.* [10] 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 [10].

A study by Baeza-Delgado *et al.* [6] analysed the distribution of amino acid residue types in TMHs in 170 integral membrane proteins from a manually maintained database of experimentally confirmed TMPs (MPTopo [12]) 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 [10].

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 [6, 7, 10] 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 [26, 53, 91]. 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 [118] 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 [119] 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/`.

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

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

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.

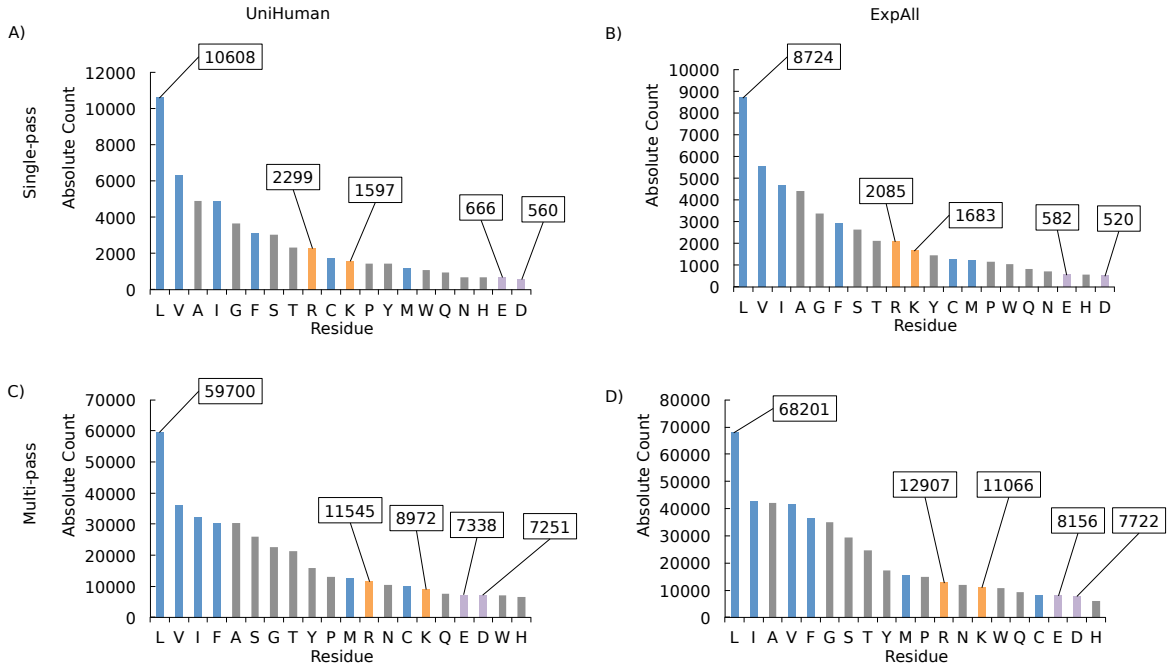

Figure 2.1: Negatively charged amino acids are amongst the rarest residues in TMHs and ±5 flanking residues. Bar charts of the abundance of each amino acid type in the TMHs with flank lengths of the accompanying ±5 residues from the (a) UniHuman single-pass proteins, (b) ExpAll single-pass proteins, (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 [15]. 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.* [120] 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% [120]).

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

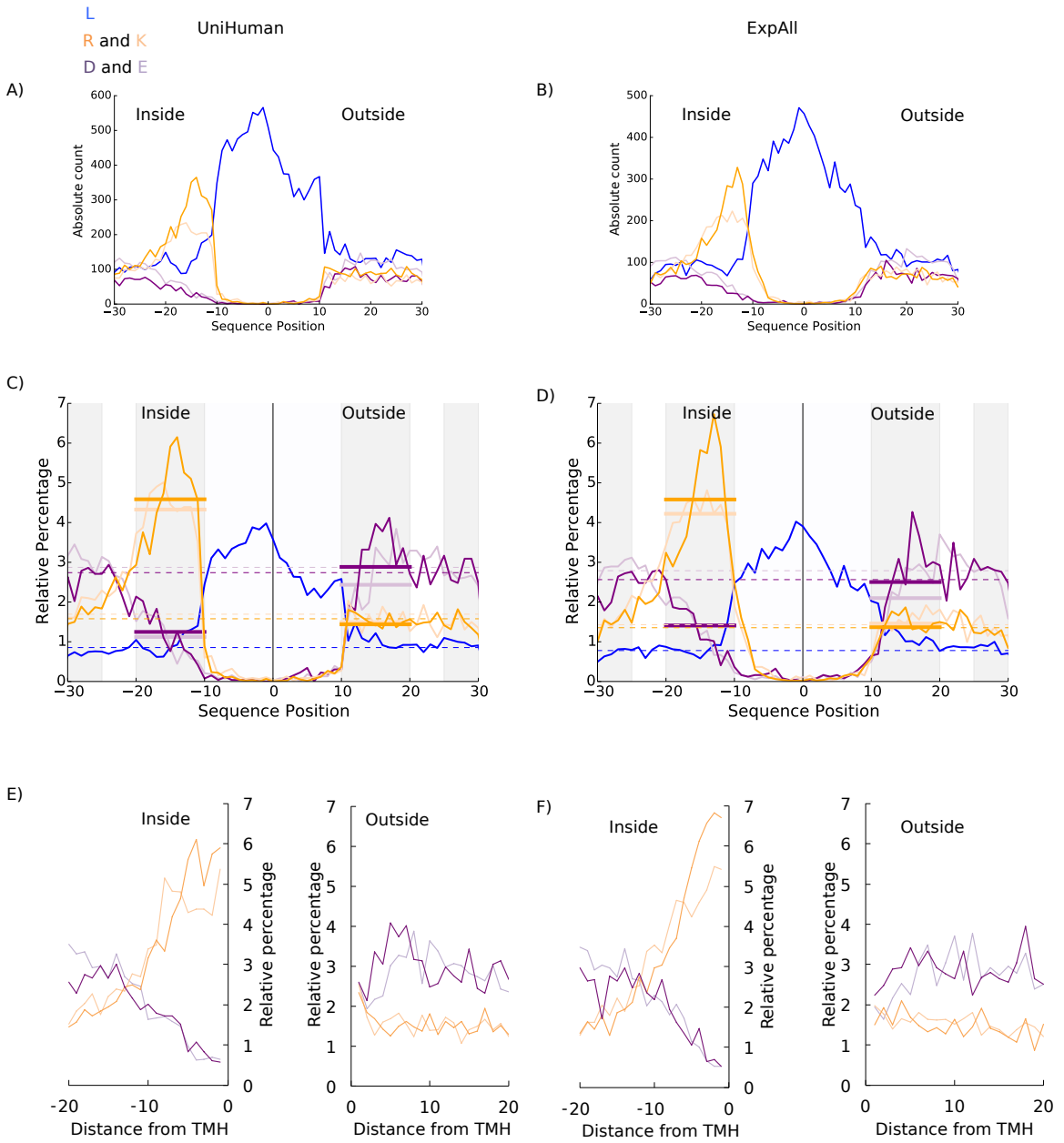

Figure 2.2: Relative percentage normalisation reveals a negative-outside bias in TMHs from single-pass protein datasets. All flank sizes were set at up to ±20 residues. We acknowledge that all values, besides the averaged values, are discrete, and connecting lines are illustrative only. On the horizontal 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 [119]; 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 Kruskal-Wallis 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 **??**). 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,

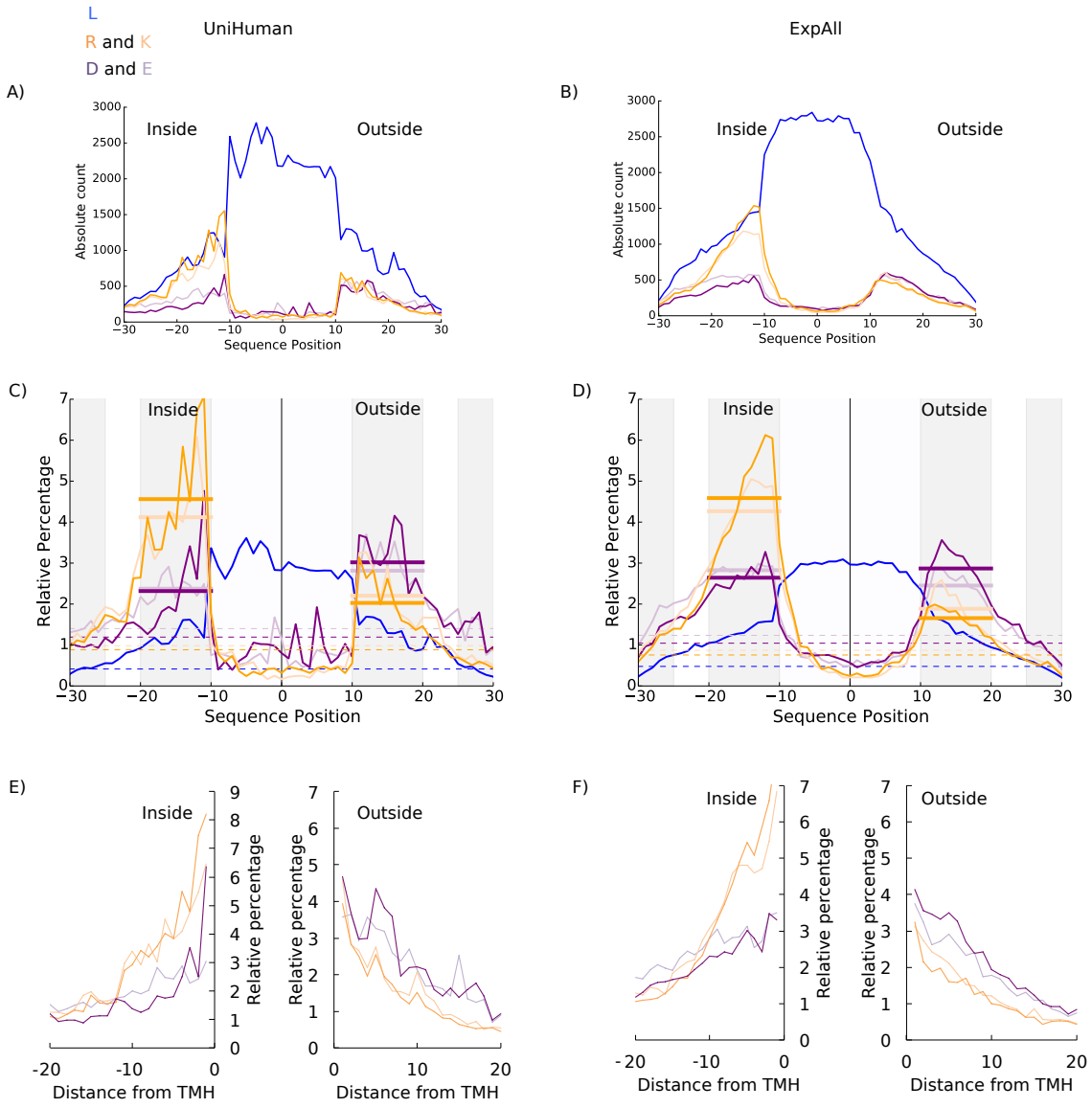

Figure 2.3: Negative-outside bias is very subtle in TMHs from multi-pass proteins. The meaning for the horizontal axis is the same as in Figure 2.2, with the negative sequence position numbers extending towards the cytoplasmic space. Leucine is in blue. Arginine and lysine are shown in dark and light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