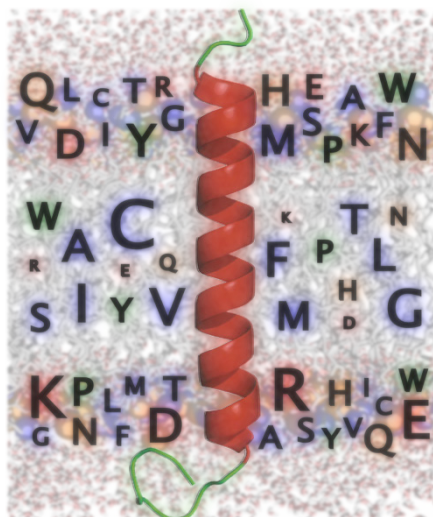


On the design of optimally inserted transmembrane helices



Carlos Baeza Delgado
Tesis Doctoral 2017

Directores:

Dr. Ismael Mingarro Muñoz
Dr. Marc A. Marti-Renom



VNIVERSITAT
DE VALÈNCIA

Doctorado en Biotecnología

Departament de Bioquímica i Biologia Molecular

ERI Biotecnologia i Biomedicina

On the design of optimally inserted transmembrane helices

Memoria presentada por Carlos Baeza Delgado

Para optar al Grado de Doctor por la Universitat de València

Directores: Dr. Ismael Mingarro Muñoz

Dr. Marc A. Martí-Renom

Ismael MINGARRO MUÑOZ, Catedràtic del Departament de Bioquímica i Biologia Molecular de la Universitat de València,
Marc A. Marti-Renom, Professor ICREA del Centre Nacional de Anàlisi Genòmic – Centre de Regulació Genòmica

MANIFESTEN que En Carlos Baeza Delgado, Llicenciat en Bioquímica per la Universitat de València, ha realitzat sota la nostra direcció al Laboratori de Proteïnes de Membrana de la ERI BioTecMed (Departament de Bioquímica i Biologia Molecular) la memòria titulada “*On the design of optimally inserted transmembrane helices*” que presenta per a optar al grau de Doctor. Alhora, aprofitem per a autoritzar la presentació i defensa d’aquesta memòria.

I perquè així conste i tinga els efectes oportuns, signem el present document en Burjassot el 24 de maig de 2017



Dr. Ismael Mingarro



MARC A. MARTI-RENO

Dr. Marc A. Marti-Renom

Para la realización de esta Tesis Doctoral, Carlos Baeza Delgado ha disfrutado de una Beca del Programa de Formación de Personal Investigador concedida por el Ministerio de Ciencia e Innovación (FPI, Convocatoria 2010), y de dos Ayudas para la Realización de Estancias Breves en el “Department of Biochemistry and Biophysics, Stockholm University” (Ayuda Estancias Breves 2012), concedidas por el Ministerio de Economía y Competitividad.

El trabajo se ha enmarcado dentro de los proyectos: “Membrane protein expression, folding and dynamics” (BFU2009-08401/BMC) financiado por el Ministerio de Ciencia e Innovación; “Membrane Protein Biology” (BFU2012-39482) financiado por el Ministerio de Economía y Competitividad; y “Modulación de interacciones proteína-proteína en apoptosis como diana terapéutica en procesos tumorales” de la Generalitat Valenciana (PROMETEOII/2014/061) de los que el Dr. Ismael Mingarro es Investigador Principal.

La Tesis ha sido realizada en el “Membrane Proteins Laboratory” (MemProt Lab) del Departament de Bioquímica i Biologia Molecular de la Universitat de València y parcialmente en el “Department of Biochemistry and Biophysics, Stockholm Center for Biomembrane Research” de la Universidad de Estocolmo, bajo la supervisión del Prof. Gunnar von Heijne, y en el Centro Nacional de Análisis Genómico (CNAG)/Centre de Regulació Genòmica (CRG) de Barcelona.



UNIVERSITAT
DE VALÈNCIA

cnag

centre nacional d'anàlisi genòmica
centro nacional de análisis genómico



Stockholm
University



MINISTERIO
DE ECONOMÍA
Y COMPETITIVIDAD



Agradecimientos

Hay muchas personas que, de una u otra manera, han estado implicadas en la realización de esta Tesis, y a las que me gustaría expresar mi más sincero agradecimiento.

En primer lugar, a Ismael y a Marc, directores de esta Tesis. Gracias por darme la maravillosa oportunidad de iniciarme en el fascinante mundo de la Ciencia, y por haberme permitido trabajar bajo vuestra dirección. Gracias por vuestra guía y vuestros consejos, por todo el conocimiento, esfuerzo y dedicación que habéis invertido en mí. Y, sobre todo, gracias por vuestra paciencia.

A todos los miembros del Grupo de Proteínas de Membrana con los que he tenido el placer de trabajar. A Luis y a Silvia, por guiarme en mis primeros pasos en el laboratorio. A Manolo, por ser mi compañero de Tesis por excelencia, gracias por todo el tiempo que hemos pasado juntos y por hacer el laboratorio más divertido. A María Jesús, gracias por tu ayuda y por todo el trabajo que haces por el laboratorio. Y a Brayan y a Natalia, es una alegría ver que en el laboratorio vuelve a ver nuevos futuros doctores después de tantos años.

A Davide, François, David y Fran, del grupo “Structural Genomics” del CNAG. Gracias por tratarme como uno más cuando he tenido que ir allí, por vuestra atención, por perder vuestro tiempo en explicarme tantas cosas y por resolver todas mis dudas, que fueron muchas. Gracias especialmente a François y Yasmina por acogerme en su casa una de las veces que me tocó ir a Barcelona.

Many thanks to Prof. Gunnar von Heijne and IngMarie Nilsson for welcoming me in their labs at the Stockholm University. It was an amazing scientific experience. Thanks also to all the people of the lab (Patricia, Aurora, Karin, Nina, Florian...) and the rest of the people at the DBB (Rickard, Carmen, Jake, Nurzian, Johannes, Bill...); it was a

pleasure to meet you all. Special thanks to the people of the “Stockholm family” and all the friends I met there: Pilar, Diogo and Cata, Marco, Rafa, Albert, Edurne, Ane, Damiano, Salome, Jim, PG... Thanks a lot for all the fun, for guiding me, for the beers, the dances, the dinners and the parties! Pau y Sara, muchas gracias por vuestra visita. Da igual el momento y el lugar, vosotros siempre estáis ahí. All of you made the wonderful memory I have of my time in the beautiful city of Stockholm. Tack så mycket!

Al denominado “Grupo Mocholí”: Ernest, Manolo, Carlos, Mercè y Natalia. No hay palabras para agradecer los innumerables buenos momentos que hemos compartido todos estos años: comidas, cafés, cervezas y mistelas, conversaciones, cenas y fiestas, arenajas, conatos de revolución en el Departamento, “mudanzas”, Oliba-ba (tu-tu-tururu!)... No sabéis lo mucho que ha significado durante este tiempo vuestra amistad, sin vosotros todo habría sido más difícil. Gracias por hacer el día a día más llevadero y por estar siempre disponibles, ya sea para desahogarse en un mal día o para compartir, en cualquier momento, conversaciones y risas con un café/cerveza. Os echaré mucho de menos.

Gracias a todas las demás personas del Departamento de Bioquímica con las que he coincidido y con las que también he compartido buenos momentos, cenas de departamento y conversaciones en los pasillos: Inma, Toni, Tian, María, Ana, Fany, Elena, Salva, Ester, Sara, y muchos más nombres que me dejo. Gracias también a todos los profesores del Departamento, por su colaboración y ayuda cuando la he necesitado.

A Pau y Sara, Ana y David, Gabi y Estela, Javi, Asahi, Alex, Alberto... Gracias por interesaros y preocuparos, por vuestra amistad y por formar parte de mi vida.

Y por último, pero no por ello menos importante, gracias a mi familia. A mis padres, Carlos y Manoli. Gracias por todo lo que habéis hecho por mí, por educarme y formarme como persona, por vuestro apoyo incondicional, por animarme en todo momento a seguir formándome y

por ser siempre un ejemplo a seguir, en todos los aspectos. A mis hermanos, María, Mateo y Marcos. Gracias por vuestro interés, por haber estado siempre ahí, por saber que siempre puedo contar con vosotros y por haber hecho mi vida tan alegre. Gracias especialmente a Eva, tú eres el pilar fundamental de mi vida desde mucho antes de empezar esta Tesis. Gracias por todo tu cariño, apoyo, comprensión y, sobre todo, gracias por darle sentido a mi vida. Y gracias a Lucía, eres la alegría de mi corazón. Gracias por haberme dado el más grande de los títulos que puedan existir.

A Eva y a Lucía

CONTENTS

ABBREVIATIONS.....	7
Amino acids, one and three letter code	8
SUMMARY	9
INTRODUCTION.....	11
I.1. BIOLOGICAL MEMBRANES	13
i.1.1. Membrane lipids	14
i.1.2 Membrane proteins	17
I.2. FOLDING AND STABILITY OF HELICAL MEMBRANE PROTEINS.....	20
i.2.1. The two-stage model.....	20
i.2.2. General structural features of membrane proteins	24
i.2.3. Driving forces in membrane proteins folding	26
<i>Van der Waals forces</i>	26
<i>Hydrogen bonds</i>	26
<i>Salt-bridge interactions</i>	27
<i>Aromatic-aromatic interactions</i>	27
<i>Motifs involving glycine residues</i>	28
I.3. BIOGENESIS OF HELICAL MEMBRANE PROTEINS	28
i.3.1 The Sec translocon.....	29
i.3.2. Co-translational targeting.....	31
i.3.3. Post-translational translocation.....	32

i.3.4. Insertion into the ER membrane.....	34
i.3.5. Translocon associated proteins.....	35
I.4. TOPOLOGY OF MEMBRANE PROTEINS	38
i.4.1. Classification of membrane proteins according to their topology.....	38
i.4.2. Topological determinants	39
<i>The positive-inside rule</i>	39
<i>Transmembrane length</i>	40
<i>Hydrophobicity</i>	41
<i>Long loops and globular domains</i>	41
<i>Lipid composition</i>	41
I.5. COMPUTATIONAL METHODS IN MEMBRANE PROTEINS	42
i.5.1. Statistical and machine learning methods	43
i.5.2. Hydrophobicity scales	45
i.5.3. Membrane protein predictors	47
<i>Transmembrane domain predictors</i>	47
<i>Topology predictors</i>	48
i.5.4. Membrane proteins databases	51
OBJECTIVES.....	53
METHODOLOGY.....	57
M.1. COMPUTATIONAL METHODS	59
m.1.1. Helix data sets	59
m.1.2. Amino acid propensity	60
m.1.3. Computational sequence design	61

m.1.4. Prediction of the ΔG values and probability of insertion	62
M.2. EXPERIMENTAL METHODS.....	64
m.2.1. Biological material.....	64
E. coli.....	64
<i>Growing conditions</i>	64
<i>Thermal shock transformation</i>	64
<i>Electric shock transformation</i>	65
m.2.2. DNA manipulation.....	65
<i>DNA isolation</i>	65
<i>Insert construction, vector preparation and ligation</i>	65
<i>Site-directed mutagenesis</i>	66
m.2.3. Glycosylation assay.....	66
m.2.4. <i>In vitro</i> transcription and translation	68
Enzymes and chemicals	69
RESULTS	71
R.1. STRUCTURE-BASED STATISTICAL ANALYSIS OF TRANSMEMBRANE HELICES.....	73
r.1.1. Helix length in membrane and water-soluble proteins	73
r.1.2. Amino acid composition of α -helices	74
r.1.3. Amino acid distribution in TM helices	77
<i>By taxonomic domains</i>	77
<i>By monotopic/polytopic proteins</i>	80
r.1.4. Position-dependent distribution of amino acid residues in TM helices	81

R.2. BIOLOGICAL INSERTION OF COMPUTATIONALLY DESIGNED SHORT TRANSMEMBRANE SEGMENTS	88
r.2.1. Predicted insertion capacity for designed sequences	88
r.2.2. Determination of experimental insertion by glycosylation assay	92
r.2.3. Correlation between predicted and experimentally determined insertion efficiencies	96
r.2.4. Analysis and mutants of sequences with low correlation between predicted and experimental insertion.....	101
DISCUSSION	107
D.1. STRUCTURE-BASED STATISTICAL ANALYSIS OF TRANSMEMBRANE HELICES.....	109
D.2. BIOLOGICAL INSERTION OF COMPUTATIONALLY DESIGNED SHORT TRANSMEMBRANE SEGMENTS	114
CONCLUSIONS.....	121
RESUMEN.....	125
R.1. INTRODUCCIÓN	127
<i>La membrana biológica</i>	<i>127</i>
<i>Proteínas de membrana</i>	<i>128</i>
<i>Plegamiento y estabilidad de proteínas de membrana helicoidales</i>	<i>129</i>
<i>Biogénesis de proteínas de membrana helicoidales.....</i>	<i>131</i>
<i>Topología de las proteínas de membrana</i>	<i>133</i>
R.2. OBJETIVOS	135
R.3. METODOLOGÍA	136
r.3.1. Métodos computacionales.....	136

<i>Bases de datos de α-hélices</i>	136
<i>Distribución de aminoácidos</i>	137
<i>Diseño computacional de secuencias</i>	138
<i>Predicción de los valores de ΔG_{app} y probabilidad de inserción...</i>	139
r.3.2. Métodos experimentales	141
<i>Material biológico</i>	141
<i>Manipulación del DNA</i>	142
<i>Ensayos de glicosilación</i>	143
<i>Transcripción y traducción in vitro</i>	144
<i>Productos químicos y enzimas</i>	145
R.4. CONCLUSIONES	147
BIBLIOGRAPHY	149
ANNEX I	165
ANNEX II	171
ANNEX III	181
ANNEX IV	193
ANNEX V	205
ANNEX VI	215

ABBREVIATIONS

ANN	Artificial neural network
BiP	Binding immunoglobulin Protein
DOPC	Dioleoylphosphatidylcholine
ER	Endoplasmic reticulum
HMM	Hidden Markov model
OPM	Orientations of Proteins in Membranes
OST	Oligosaccharyltransferase
PDB	RCSB Protein Data Bank
PDBTM	Protein Data Bank of Transmembrane Proteins
PPM	Positioning of Proteins in Membranes
RNC	Ribosome-nascent chain
SP	Signal sequence peptidase
SR	SRP receptor
SRP	Signal recognition particle
SS	Signal sequence
SVM	Support vector machine
TM	Transmembrane
TMH	Transmembrane helix

Amino acids, one and three letter code

Any amino acid	-	Xaa	-	X
Alanine	-	Ala	-	A
Arginine	-	Arg	-	R
Asparagine	-	Asn	-	N
Aspartic acid	-	Asp	-	D
Cysteine	-	Cys	-	C
Glutamic Acid	-	Glu	-	E
Glutamine	-	Gln	-	Q
Glycine	-	Gly	-	G
Histidine	-	His	-	H
Isoleucine	-	Ile	-	I
Leucine	-	Leu	-	L
Lysine	-	Lys	-	K
Methionine	-	Met	-	M
Phenylalanine	-	Phe	-	F
Proline	-	Pro	-	P
Serine	-	Ser	-	S
Threonine	-	Thr	-	T
Tryptophan	-	Trp	-	W
Tyrosine	-	Tyr	-	Y
Valine	-	Val	-	V

SUMMARY

The cell is the basic structural and functional unit of life and all cells are surrounded and delimited by biological membranes. Membranes play a crucial role in the existence of cells by isolating from and connecting with their environment. Biological membranes are mainly formed by lipids and proteins. Whereas lipids form a bilayer that prevents free diffusion of most molecules and ions, proteins control molecular trafficking and information flow across the membrane. These proteins embedded within biological membrane are called membrane proteins, characterized by the presence of sequence regions adapted to insert, fold and function in the complex environment of membranes. Proper insertion, folding and orientation into the membrane are essential for the correct function of membrane proteins.

This Thesis is focus on the transmembrane domains of α -helical membrane proteins, a type of molecules that comprise 20-30% of all genes in most sequenced genomes. However, due to the complexity of the environment in which they live, our knowledge about membrane proteins is still far away from that of soluble proteins.

INTRODUCTION

I.1. BIOLOGICAL MEMBRANES

Biological membranes are the boundaries that surround and maintain the cell integrity, acting as a barrier that separates the interior of the cell from the exterior environment. Nonpolar molecules are able to pass through the membrane following concentration gradients, but the hydrophobicity of biological membranes does not allow polar compounds and ions from crossing it by free diffusion, allowing the creation of electrical potential and concentration gradients across the membrane. Moreover, eukaryotic cell membranes contain organelles, producing the compartmentalization needed for the proper operation of metabolic pathways. Besides its role as a semi-permeable barrier, biological membranes are involved in cell division, communication with the exterior, molecular trafficking and biological reproduction.

The two main components of membranes are lipids and proteins, molecules characterized by being amphiphilic, a property that constitutes the structure of a membrane in an aqueous solution. The lipids are organized in a double layer, which thermodynamically favour and stabilize the membrane by the interaction of hydrophobic acyl chains facing each other in the interior of the membrane excluding water molecules and the polar head groups oriented to the external aqueous environment.

One of the earliest models of cell membrane was the ‘fluid mosaic model’, proposed by Singer and Nicolson in 1972 (Singer and Nicolson, 1972). In this model, the membrane is seen as an ocean of lipids with few proteins floating around in it, and where lipids and proteins are in constant motion, freely moving laterally within the membrane matrix ([Figure 1](#)). However, new experimental approaches in the membrane field sustain a more crowded lipid bilayer with high number of proteins embedded into the membrane, limiting free lateral

INTRODUCTION

diffusion of molecules, suggesting that biological membranes are more mosaic than fluid (Engelman, 2005; Goñi, 2014).

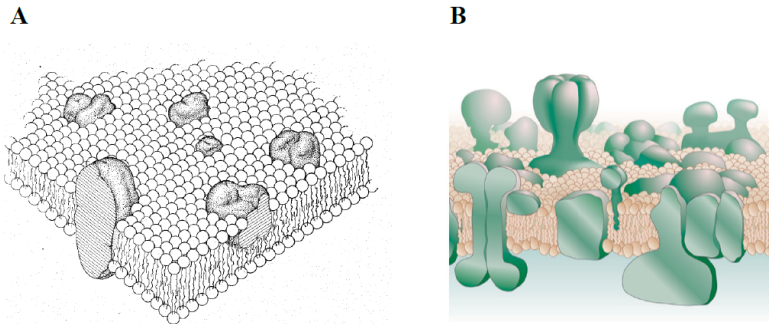


Figure 1. Membrane models. (A) Model structure of the cell membrane based on the Singer-Nicolson model (Singer and Nicolson, 1972), where lipids and proteins are freely diffusing laterally. (B) An update version of a more crowded membrane according to current knowledge (Engelman, 2005).

i.1.1. Membrane lipids

Lipids in biological membranes are considered as amphipathic molecules. They consist of a highly hydrophobic hydrocarbon tail and a hydrophilic phosphate-containing head group. By the hydrophobic effect, lipids spontaneously form a bilayer structure in aqueous solution. The hydrocarbon tails of the lipids associate between them in the hydrophobic core, excluding water molecules. The hydrophilic head groups shield the tails, forming an interface between the hydrophobic core of the membrane and the surrounding aqueous environment (Granseth *et al.*, 2005).

INTRODUCTION

There are three major types of lipids in biological membranes: phospholipids, glycolipids and cholesterol ([Figure 2 A](#)). Due to the diversity of chain length of the fatty acids and the different head groups that exists, there are a high number of different lipid species. One can find more than 100 different types present in a membrane of a prokaryotic cell such as *Escherichia coli* (Raetz and Dowhan, 1990), and more than 1000 in a eukaryotic cell (Sud *et al.*, 2007). In fact, cells use $\approx 5\%$ of their genes to synthesize all the lipids they need (van Meer *et al.*, 2008). Based on their structure, lipids can be divided into inverted conical with positive curvature, conical with negative curvature and cylindrical with no curvature ([Figure 2 B](#)).

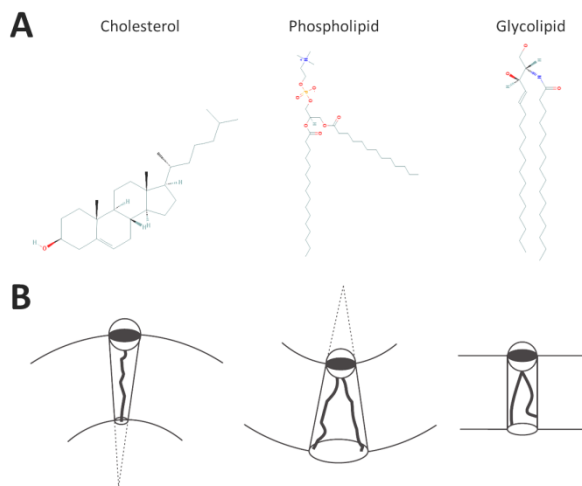


Figure 2. Different types of lipids in membrane. (A) According to their chemical composition: cholesterol, phospholipids (dipalmitoylphosphatidylcholine, DPPC) and glycolipids (C16 ceramide). (B) According to their structure: inverted conical (positive curvature), conical (negative curvature) and cylindrical (no curvature).

INTRODUCTION

Phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylserine (PS) and sphingomyelin (SM) are the most common lipids in the eukaryotic plasma membrane, whilst phosphatidic acid (PA) and phosphatidylinositol (PI) can be found at a lower level (Suetsugu *et al.*, 2014). The phospholipid composition varies between cell types and membranes. The plasma membrane is enriched in sphingolipids and cholesterol. This lipid composition of the plasma membrane contributes to a more rigid and stable membrane. In endoplasmic reticulum (ER) membrane, the scarcity of sphingolipids and cholesterol, and the abundance of unsaturated fatty acids results in looser lipid packing that is consistent with the role of this organelle in transporting and inserting newly synthesized lipids and proteins into the lipid bilayer (van Meer *et al.*, 2008).

In addition, membranes are asymmetric. That is, phospholipids are distributed into the two monolayers of the membrane asymmetrically (Daleke, 2003). Lipids can move between the two leaflets of the bilayer spontaneously in a very slow manner (Kleinfeld *et al.*, 1997), or this movement can be catalysed by enzymes like flippases (Kol *et al.*, 2004), which move PE and PS from outer to cytosolic leaflet, floppases (van Meer, 2011) that move phospholipids from cytosolic to outer leaflet, and scramblases (Williamson, 2015), which move phospholipids in either direction, toward equilibrium. Furthermore, anisotropy can also be found in the leaflet plane of the membranes. The lateral movement of membrane components allows different levels of organization. In some membranes, the lipid distribution is not random, but different types of lipids group into clusters enriched in cholesterol and sphingolipids. These microdomains, together with membrane proteins, form

highly organized structures known as lipid rafts (Sonnino and Prinetti, 2013).

According to the electronic density profile obtained by molecular dynamics simulation, a model bilayer formed solely by dioleoylphosphatidylcholine (DOPC) can be divided into four regions (MacCallum *et al.*, 2008) (Figure 3). Starting from the center of the bilayer, region I contains only hydrophobic tails of lipids, this region constitutes what is generally known as the hydrophobic core and occupies approximately 30 Å thick. Region II contains the beginning of hydrocarbon chains of the lipids and the initial part of the heads polar regions. Region II has the largest electronic density and is the most diverse, containing both hydrophobic as well as hydrophilic compounds. Region III begins at the maximum of total electronic density of the system and ends at the time when the most of the density comes from water, in this region we find the major part of the phosphate and choline groups. Finally, the IV region is the outer limits of the membrane, composed mainly of water and a small portion of the polar heads of the lipids. The sum of regions II and III is considered as the interfacial region, and represents about 15 Å of thickness on each side of the hydrophobic core.

i.1.2 Membrane proteins

Membrane proteins are those that reside and exert their function while inserted into biological membranes. They have crucial and specific roles, and are involved in many biological functions. In addition to maintaining the shape of the lipid bilayer, membrane proteins function as receptors, signal transducers, transporters, channels, motors or anchors and participate in signalling, transport, enzymatic processes and cell adhesion. Around 25% of the genes in the genome of fully sequenced prokaryotic and eukaryotic organisms encode membrane

INTRODUCTION

proteins (Krogh *et al.*, 2001), and they account above 50% of the drug targets (Overington *et al.*, 2006).

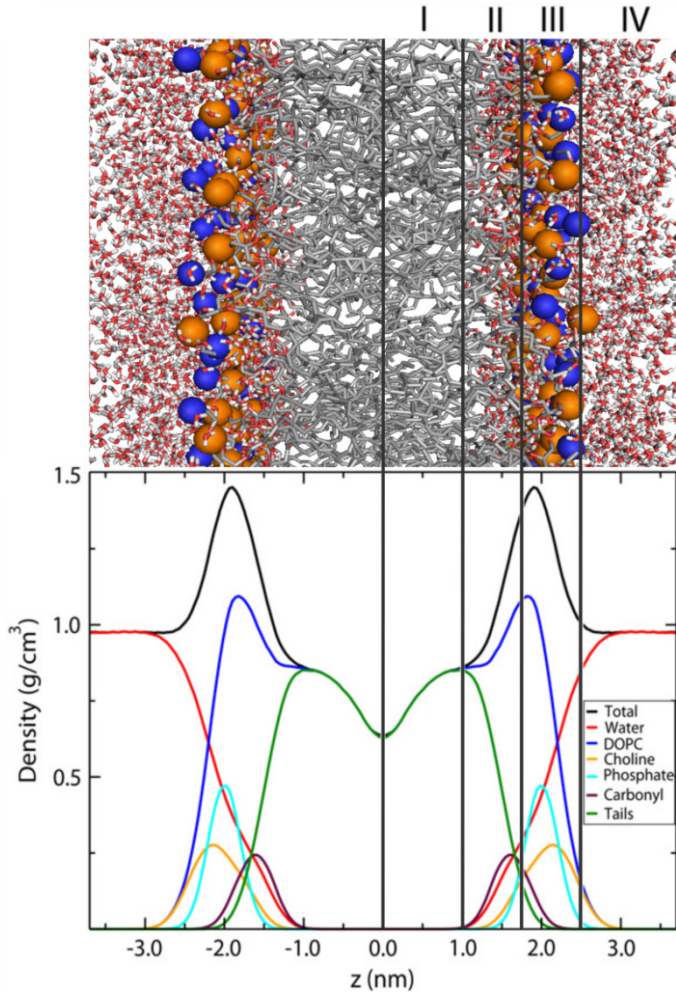


Figure 3. Density profile of a model membrane. **Upper panel**, molecular dynamic simulated structure of a membrane composed solely of dioleoylphosphatidylcholine (DOPC). Water is shown as red (oxygen) and white (hydrogen) cylinders. The lipid nitrogen and phosphate atoms are shown as blue and orange spheres respectively. The lipid tails are shown as thin grey lines. **Lower panel**, electronic density of the same membrane in which the four regions described can be differentiated. Total: total electronic density. Adapted from (MacCallum *et al.*, 2008).

INTRODUCTION

According to the interaction with the lipid bilayer, membrane proteins can be differentiated between integral and peripheral membrane proteins. Integral membrane proteins are buried inside the bilayer, surrounded by lipids. They are strongly attached to the membrane and can only be separated with hard treatments such as detergents, organic solvents or denaturants. Peripheral proteins do not span the hydrophobic core of the membrane, but they associate to the bilayer through other proteins or lipid groups covalently linked to the protein and reside at the membrane interface.

Determination of the three-dimensional structure of membrane proteins has revealed two main types of structural motifs in membrane proteins: α -helical bundles and β -barrels (Heijne, 1994; Vinothkumar and Henderson, 2010) (Figure 4). While most proteins in eukaryotic membranes are bundles of α -helices, β -barrels are found almost exclusively in outer membrane of bacteria and in mitochondria and chloroplast outer membranes. 2-3% of the genes of Gram-positive bacteria have been estimated to encode β -barrels (Wimley, 2003). In comparison, as stated above α -helical proteins represent about 25% of all open reading frames in fully sequenced genomes (Krogh *et al.*, 2001). These two types of secondary structures allow the insertion in the membrane of the polypeptide backbone by establishing intramolecular hydrogen bonds, thus reducing the polarity of the CO and NH groups of the peptide bond.

From here, this thesis focuses on α -helical proteins since they comprise the great majority of cell membrane proteins.

INTRODUCTION

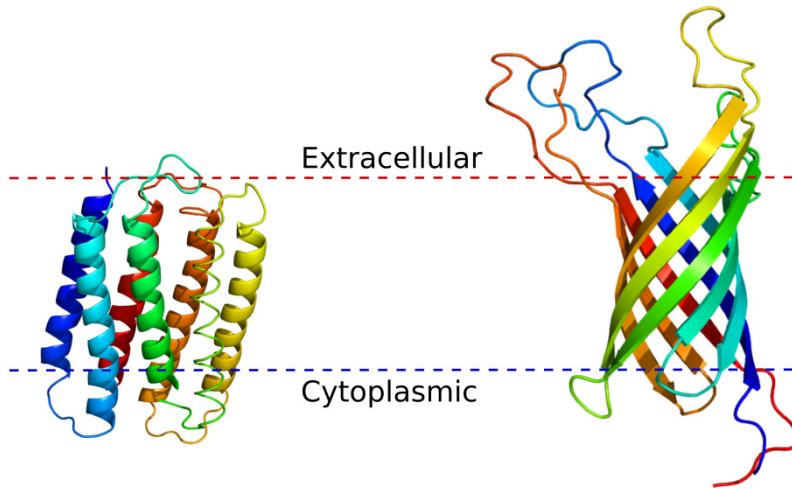


Figure 4. The two major types of structures in membrane proteins. Left: an α -helical bundle protein, bacteriorhodopsin from *Halobacterium salinarum* (PDB code: 2BRD). Right: a β -barrel protein, Outer Membrane Protein A (OmpA) from *Escherichia coli* (PDB code: 1G90). Structures are shown in cartoon representation, rainbow colored from blue (N-terminus) to red (C-terminus). Membrane boundaries obtained from the PPM server (Lomize *et al.*, 2012).

I.2. FOLDING AND STABILITY OF HELICAL MEMBRANE PROTEINS

i.2.1. The two-stage model

The current knowledge on the folding of α -helical membrane proteins is based on the model proposed in the early 1990s known as two-stage model (Popot and Engelman, 1990). According to this model, and giving rise to its name, the folding of membrane proteins occurs in two stages.

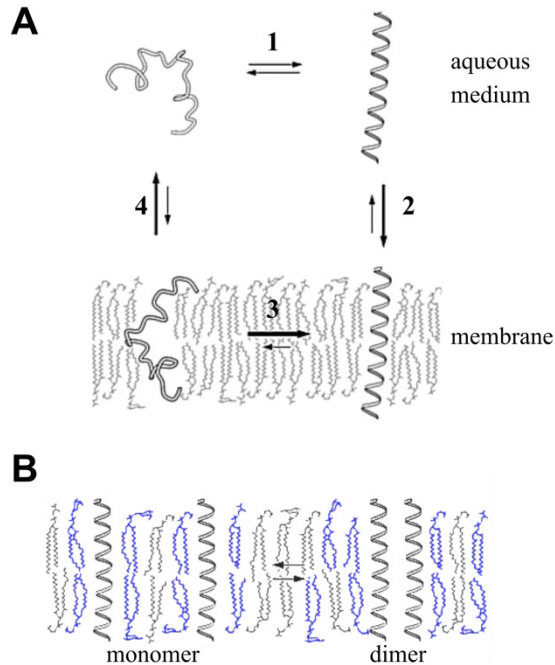


Figure 5. Two-stage model. (A) Thermodynamic equilibria for the formation of a hydrophobic helix in a lipid bilayer surrounded by water. (B) Lateral association (packing) of TMHs. Adapted from <http://medicine.yale.edu/lab/engelman/>.

In the first stage, hydrophobic sequences form α -helices when inserted into the membrane. In the presence of water and a lipid bilayer, a transmembrane α -helix (TMH) represents the most stable state for a nonpolar amino acid sequence. Since the number of hydrogen bonds of the polypeptide backbone is not critical in an aqueous milieu, it is considered that the equilibrium between an unstructured sequence and an α -helix will be governed by the side chains of the amino acids (transition 1 in Figure 5 A). A helix formed by hydrophobic residues has a high tendency to localize in the bilayer

INTRODUCTION

due to the reduction in the water entropy necessary for the same helix to remain in the aqueous milieu (transition 2 in [Figure 5 A](#)). The equilibrium of an unstructured sequence between water and lipid has been estimated to be in favor of water due to the loss of hydrogen bonds between the protein and water if this sequence enters in the bilayer (transition 4 in [Figure 5 A](#)) (White and von Heijne, 2008). Finally, disrupting a helix within the membrane is an unfavorable process, due to the energy penalty of breaking the hydrogen bonding pattern in an environment with a low dielectric constant such as that of the membrane core. Therefore, a hydrophobic sequence within the lipid bilayer is energetically prone to fold into an α -helical conformation (transition 3 in [Figure 5 A](#)).

In the second stage, probably occurring at the same time as insertion into the bilayer and acquisition of secondary structure, newly inserted transmembrane (TM) segments associate (pack) with those already in the membrane, leading to the formation of tertiary structures ([Figure 5 B](#)). The packing between two helices causes an increase in helix-helix and lipid-lipid interactions and a decrease in helix-lipid interactions.

In 1989, Jacobs and White proposed a three-step thermodynamic model for membrane protein folding based on structural and thermodynamic measurements of the partitioning of small hydrophobic peptides: interfacial partitioning, interfacial folding and insertion (Jacobs and White, 1989). The combining of their model with the two-stage model has led to the so-called four-step thermodynamic cycle or four-step model (White and Wimley, 1999). In this model, the four steps (partitioning, folding, insertion, and association) can proceed along an interfacial path, a water path, or a combination of the two ([Figure 6](#)). In addition, there is even a fifth

step in which different events can take place such as the binding of prosthetic groups, the folding of inter-helical loops, the entry of other regions of the protein into the hydrophobic core or oligomerization of different polypeptide chains into quaternary structure (Engelman *et al.*, 2003).

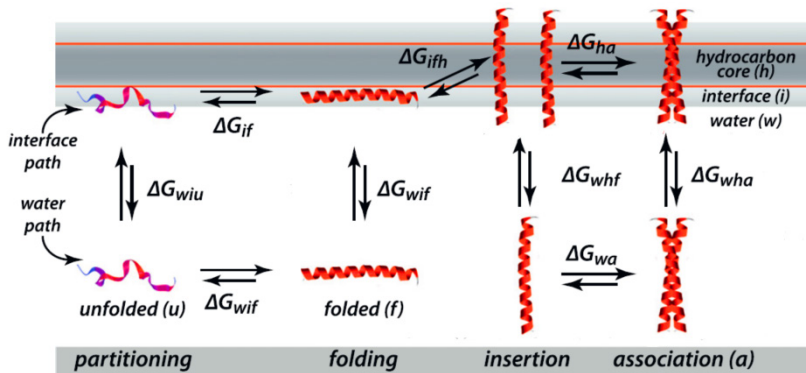


Figure 6. Four-step thermodynamic cycle for describing the energetics of the partitioning, folding, insertion, and association of an α -helix (red helices) in a lipid bilayer (gray). A polypeptide chain may partition from and to water (w), the interface (i) or the hydrophobic core of the membrane (h). In each of these environments the chain can be unfolded (u) or folded (f) adopting some type of secondary structure that neutralizes hydrogen bonds. Helices can associate (a) through interacting regions. The location of a peptide sequence will depend on the free energy variation (ΔG) associated with each of the transitions between the different. Adapted from (Cymer *et al.*, 2015).

Although these models do not necessarily mirror the actual biological assembly process of proteins into biological membranes, which will be discussed below, they can help to understand the thermodynamic constraints on membrane protein structure formation (Cymer *et al.*, 2015).

INTRODUCTION

i.2.2. General structural features of membrane proteins

The main determinant for a sequence to be inserted in the membrane is the global hydrophobicity value, which states that the tendency to insert into the membrane increases with the degree of the sequence hydrophobicity. This concept conflicts with the classical view that charged residues within the membrane are thought to be prohibited in the hydrophobic core. More recent studies have shown that this type of amino acid residues does not prevent the insertion into the bilayer of a sequence, but the insertion of a TM segment depends on its total hydrophobicity (MacCallum *et al.*, 2008; Martínez-Gil *et al.*, 2008; Ulmschneider *et al.*, 2017). A similar effect is observed with those amino acid residues that disfavour the formation of α -helical structure, such as Pro (Nilsson *et al.*, 1998) or Gly (Dong *et al.*, 2012).

Although the hydrophobicity is the main feature of TM sequences, a considerable fraction (> 30%) of the TMHs in multispanning membrane proteins are not hydrophobic enough to be efficiently inserted by themselves (Hessa *et al.*, 2007), they are only marginally hydrophobic. This finding suggests that membrane insertion of TMHs in multispanning membrane proteins in many cases may depend on sequence features extrinsic to the hydrophobic segment itself (Hedin *et al.*, 2010; Öjemalm *et al.*, 2012).

In addition to the requirements of hydrophobicity and helicity, a TMH must be long enough to be able to traverse the lipid bilayer. Considering a translation of 1.5 Å per residue in a canonical α -helix, about 20 residues are required to span the hydrocarbon core (~ 30 Å) of the membrane. However, the minimum hydrophobic length necessary to form a TMH has been investigated using model membrane-inserted hydrophobic peptides (Krishnakumar and London, 2007). These studies showed that for alternating Leu and Ala peptides (which have a hydrophobicity typical of natural TMHs), a length of 13

INTRODUCTION

consecutive residues is the minimum necessary to adopt a predominantly TM disposition in synthetic bilayers with a biologically relevant thickness. More recently, TM disposition of poly-Leu sequences were analyzed using synthetic peptides and oriented phospholipid bilayers, *in vitro* insertion into microsomal membranes and molecular dynamics simulations (Jaud *et al.*, 2009). Sequences with either blocks of or dispersed hydrophobic residues have also been analyzed using similar methods (Stone *et al.*, 2015). The picture that emerges from these studies is that lipid bilayers adapt to TMHs as short as 10–12 leucines long.

Hydrophobic mismatch occurs when the length of a TM segment does not match the thickness of the hydrophobic core of the membrane (Killian, 1998). Studies have indicated that both TMHs and lipids can adapt to minimize the effects of the mismatch (Figure 7). When the length of a TM sequence is longer than the thickness of the membrane (positive mismatch), it can tilt with respect to the normal axis of the bilayer, resulting in kinked or even interrupted helices (Bowie, 1997; Holt and Killian, 2010), or it can oligomerize to minimize the exposed hydrophobic parts. In contrast, when the TM segment is not long enough, it may result in aggregation or changes in the side chain orientation such as the ‘snorkeling’ effect, by which flexible aliphatic side chain of residues like Lys and Arg could face up into the head group region, so the positively charged end can interact with the phosphate groups of the lipid bilayer (Monné *et al.*, 1998; Killian and Von Heijne, 2000). In addition, lipids can also respond to hydrophobic mismatch by stretching or disordering their fatty acid chain, or by reorganizing membrane composition in microdomains or lipid rafts.

INTRODUCTION

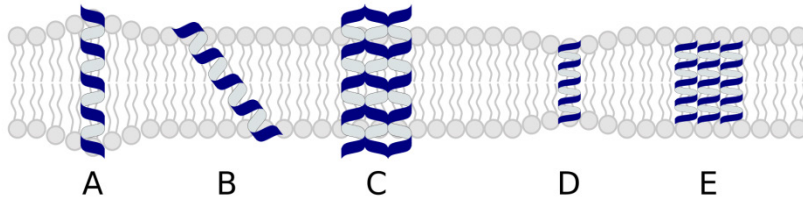


Figure 7. Possible adaptations to hydrophobic mismatch. In the case of too-long transmembrane peptides: (A) bilayer distortion, (B) peptide tilting and/or (C) peptide aggregation. For too-short transmembrane peptides: (D) bilayer distortion and/or (E) peptide aggregation.

i.2.3. Driving forces in membrane proteins folding

Van der Waals forces

Van der Waals interaction originates from the attraction between instantaneous atomic dipoles caused by fluctuation of electrons. Although this is a weak force when considered individually, the sum of all van der Waals interactions found in a protein can be a significant force in its folding. Moreover, residues in TM regions tend to be buried more often than residues in soluble proteins or extracellular regions, allowing a higher number of van de Waals interactions (Oberai *et al.*, 2009).

Hydrogen bonds

A hydrogen bond is the electrostatic attraction between a hydrogen atom covalently bound to a highly electronegative atom such as nitrogen (N) or oxygen (O) and another highly electronegative atom. The electronegative atom attracts the electron density from around the hydrogen nucleus and, by decentralizing it, leaves the hydrogen atom with a positive partial charge. Hydrogen atoms attached to carbon can also participate in hydrogen bonding when the carbon atom is bound

to electronegative atoms, resulting in a weaker hydrogen bond. In membrane proteins, TMHs backbones can be involved in such weak hydrogen bond (Jiang and Lai, 2002). Some studies have shown that strong hydrogen bonds involving Asn, Asp, Gln and Glu can drive helix oligomerization (Choma *et al.*, 2000; Zhou *et al.*, 2000; Meindl-Beinker *et al.*, 2006). The strength of the hydrogen bond depends on the distance, the chemistry and the arrangement of the atoms involved, and the nature of the surrounding milieu; small structural changes can easily break a hydrogen bond (Mottamal and Lazaridis, 2005).

Salt-bridge interactions

Salt-bridges or ion-pairs in proteins are formed when two oppositely charged groups are located within 4 Å distance. Buried charged residues have a stronger tendency to form salt-bridges than exposed residues (Donald *et al.*, 2011). Within the hydrophobic environment of the membrane, salt-bridge formation is a strong interaction due to the decrease of the di-electric constant. Even though charged residues (Asp, Glu, Lys, and Arg) are present at a low frequency level in TMHs (Bañó-Polo *et al.*, 2012), some studies have shown the importance of intramembrane salt-bridges in TM packing (Bañó-Polo *et al.*, 2013), as well as in the structure and function of helical membrane proteins (Sahin-Toth *et al.*, 1992; Donohue *et al.*, 1999; Hall *et al.*, 1999).

Aromatic-aromatic interactions

Strong attraction between aromatic rings has been long recognized as an important driving force in stabilizing nucleic acids, proteins, drug-protein complexes, and poly-aromatic macrocycles (Hunter and Sanders, 1990). A recent study has shown that the motifs (Q,W,Y)_{xx}(Q,W,Y) could drive self-oligomerization of TMHs in a model system, and that this oligomerization effect was stronger when

INTRODUCTION

the aromatic residues were close to the interface region of the membrane (Sal-Man *et al.*, 2007). Another study showed that Trp at various positions within a TMH promote homo-oligomerization in sequences containing a randomised heptad repeat pattern (Ridder *et al.*, 2005).

Motifs involving glycine residues

Different motifs involving residues with small side chain allow tight packing of TMHs. Side chains in helices form ‘knob-into-hole’ interactions, where ‘knobs’ are usually branched residues such as Val and Ile, which fits into ‘holes’ formed primarily by Gly residues. This kind of interaction was originally observed in the dimerization of the TMH of glycophorin A and described as a oligomerization motif (Lemmon *et al.*, 1992, 1994), which was more recently minimised (Orzáez *et al.*, 2005). The GxxxG motif together with the less common GxxxA have been found in many membrane proteins (Russ and Engelman, 2000). Other Gly containing motifs involved in helix-helix interactions are GxxGxxG, known as Gly zipper (Senes *et al.*, 2000) and GxxxxxxG or GG7 (Liu *et al.*, 2002). Statistic studies have shown that Gly zippers like GxxxG, (G,A,S)xxxGxxxG and GxxxGxxx(G,S,T) are present in more than 10% of all known membrane protein structures (Kim *et al.*, 2005).

I.3. BIOGENESIS OF HELICAL MEMBRANE PROTEINS

The vast majority of proteins are synthesized in the cytosol, where the ribosome translates mRNA codons into amino acids and add them to the growing polypeptide chain by catalyzing the peptide bond formation. Soluble cytosolic proteins simply fold as they emerge from

the ribosome. However, for secreted proteins and membrane proteins the process is more complicated as they have to cross the membrane, partially or totally. They use the machinery known as translocon for its insertion and transport. In eukaryotic cells the translocon is a multiprotein complex located on the ER membrane, consisting basically of a channel that crosses the lipid bilayer. The translocon allows soluble proteins to completely pass through the ER membrane, and TM fragments of the integral membrane proteins to be laterally inserted in the bilayer (Panzner *et al.*, 1995).

i.3.1 The Sec translocon

The universal protein-conducting channel is called Sec61 in eukaryotes and SecYEG in archaea and bacteria. The eukaryotic Sec61 complex consists of three subunits, α , β , and γ (Y, G and E in prokaryotes). Both α and γ subunits are highly conserved and are essential for cell viability. In 2004 the determination of the X-ray structure of the archaea Sec complex from *Methanococcus jannaschii* (Berg *et al.*, 2004) was a breakthrough in the study of translocation and membrane protein integration.

The Sec61 α (SecY) has 10 TM segments that form the aqueous pore of the translocon, that has an hourglass shape (Figure 8 A). It contains a constriction ring of six Ile residues in the cytoplasmic half of the membrane, maintaining the permeability barrier during protein translocation (Park and Rapoport, 2011). The α -subunit is divided into two halves (TM1-5 and TM6-10) that forms a ‘clam shell’ and has an interconnected hinge at the cytoplasmic loop between TM5 and TM6 (Figure 8 B). Furthermore, an extension of TM2 forms an α -helical ‘plug’ in the luminal side (Figure 8 A). This ‘plug’ blocks the channel in the closed state and moves away during protein translocation, conforming the open state of the translocon (Gogala *et al.*, 2014).

INTRODUCTION

Additionally, an opening between TM2 and TM7 creates a lateral gate that facilitates partitioning of TMHs into the membrane ([Figure 8 B](#)); it has been confirmed that two halves of the α -subunit move apart approximately 12 Å in the presence of a TM segment (Gogala *et al.*, 2014).

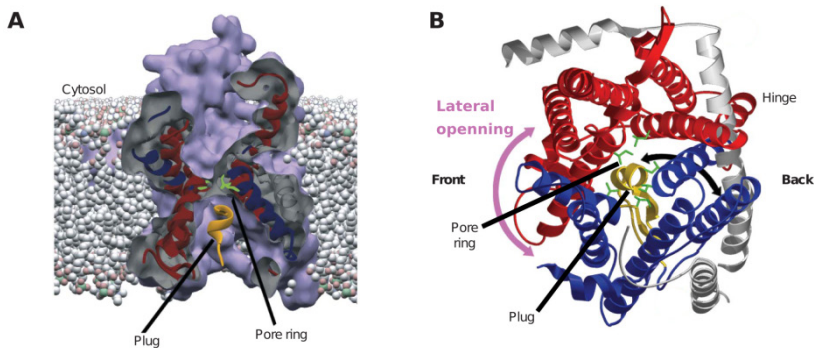


Figure 8. Structure of the translocon. (A) Cross-sectional view from the side of the crystal structure of the SecY complex from *Methanococcus jannaschii*. (B) View from the cytosol of the channel. The α -subunit consists of two halves, TM segments 1–5 and 6–10 (in blue and red, respectively), which can open the lateral gate at the front (purple double-headed arrow). The plug (TM2, in yellow) is in the centre of the α -subunit. Plug movement towards the back (black double-headed arrow) opens the channel across the membrane. The pore-ring Ile residues are indicated in green. Adapted from (Rapoport, 2007).

The role of Sec61 γ (SecE) is to clamp the two halves of Sec61 α , and is essential for cell viability and translocation (Lycklama *et al.*, 2013). However, less conserved Sec61 β subunit (SecG) is non-essential and its function is still unclear.

Therefore, the Sec translocase complex has the ability to open and close in two directions: perpendicular to the plane of the membrane to allow translocation of soluble proteins, and laterally to

allow the insertion of TM domains into the lipid bilayer. To reach the translocon, the targeting process can be either co- or post-translational.

i.3.2. Co-translational targeting

This mechanism, in which the translocation and insertion of the protein in the membrane is coupled to the translation of the nascent polypeptide chain, is the one used by secretion proteins and by the majority of the membrane proteins. The pathway starts when a signal sequence (SS) or the first TM of the growing nascent chain emerges from the ribosome and is recognized by the signal recognition particle (SRP) ([Figure 9](#)). The cleavable SS has a positively charged N-terminal region followed by a hydrophobic domain composed of 7-15 residues and a polar C-terminal region. When SRP binds to the SS it pauses the nascent chain elongation (Walter and Blobel, 1981; Mary *et al.*, 2010) and brings the whole ribosome-nascent chain (RNC)-SRP complex to the ER membrane by interacting with the SRP receptor (SR) (Akopian *et al.*, 2013). Interaction between the SRP and the SR requires GTP binding to both complexes. Subsequently, the RNC is transferred from the SRP to the translocon, and GTP hydrolysis triggers SRP-SR dissociation (Song *et al.*, 2000). Then, the ribosome can reinitiate the elongation of protein synthesis releasing the nascent chain from the exit tunnel to the translocon channel. The translocon will allow soluble domains to cross the membrane and hydrophobic TM segments to exit laterally into the lipid phase (Nyathi *et al.*, 2013) ([Figure 9](#)). Both SRP and SR have a GTPase domain. As mentioned above, they are bound to GTP when associated with the RNC, however, when GTP is hydrolyzed the whole complex is disassembled and SRP and SR are recycled for the next run (Akopian *et al.*, 2013). GTP hydrolysis is needed for the elongation of the polypeptide chain,

INTRODUCTION

but the movement through the translocon channel does not require energy.

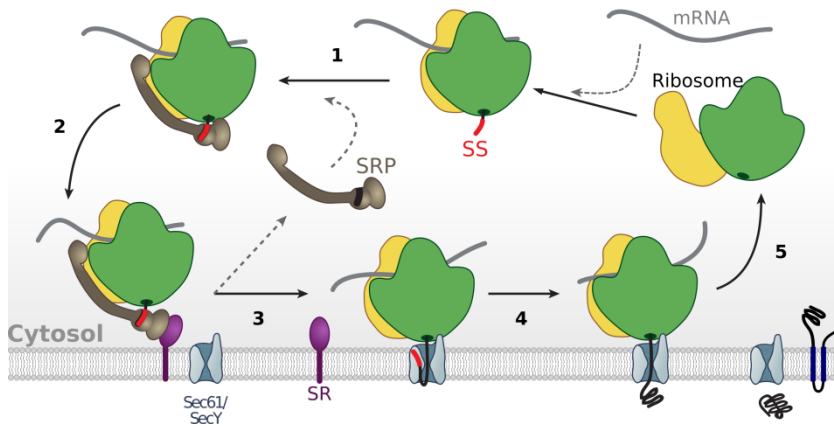


Figure 9. Model of co-translational translocation and insertion. 1: Binding of the SRP to a ribosome carrying a nascent chain with exposed signal sequence. 2: Binding of the RNC-SRP complex to the SR. 3: Release of SRP, binding of the ribosome to the Sec61 (SecY) channel, and transfer of the nascent chain into the channel. 4: Translocation of the polypeptide chain, signal sequence cleavage, and folding of the polypeptide on the other side of the membrane or lateral insertion for helical membrane proteins. Step 5: Termination of translocation or insertion and dissociation of the ribosome into its two subunits. Adapted from (Park and Rapoport, 2012).

i.3.3. Post-translational translocation

In the post-translational pathway, the protein is completely synthesized by soluble ribosomes in the cytosol and thereafter targeted and inserted into the ER membrane. This pathway is used mostly by soluble proteins, such as secretory proteins, which possess only moderately hydrophobic signal sequences that cause them to scape recognition by the SRP during their synthesis (Rapoport, 2007). These

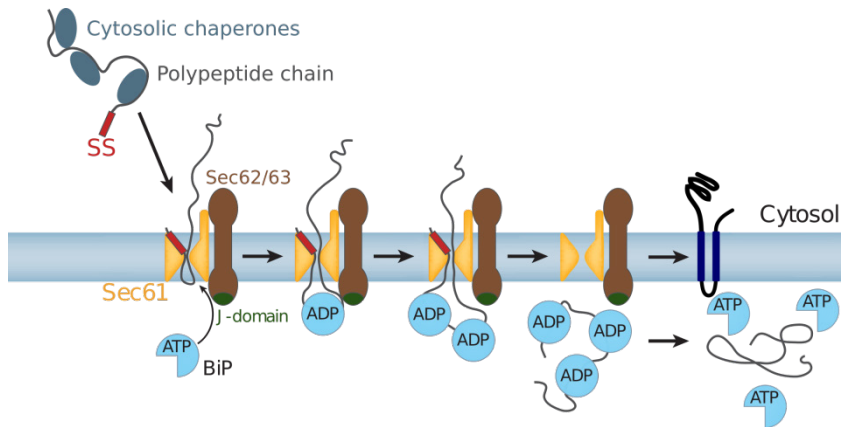


Figure 10. Model of post-translational insertion/translocation in eukaryotes. Before targeting the translocon, the polypeptide is maintained in its unfolded state in the cytosol by the binding of chaperones (dark blue). Once the polypeptide is located in the translocon, it will be inserted into the membrane (membrane proteins) or translocated through the channel to ER lumen (soluble/secretory proteins). BiP (light blue) binding to the translocating polypeptide chain on the ER lumen prevents it from returning to the cytosol. The Sec61 complex is indicated in yellow, Sec62/63 in brown with the J-domain in green. Adapted from (Rapoport, 2007).

proteins need to remain unfolded or loosely folded after their release from the ribosome to cross the membrane. After protein release, this pathway begins with the binding of cytosolic chaperones while the nascent chain emerges from the ribosome to prevent premature folding before the polypeptide is transported through the channel. The substrate is later targeted to the Sec61 complex, which interacts with the Sec62/Sec63 complex in eukaryotes ([Figure 10](#)). The chaperones release the nascent chain while it is translocated through the channel (Plath and Rapoport, 2000; Park and Rapoport, 2012). On the luminal side of the ER the chaperone BiP (Binding immunoglobulin Protein) binds to the polypeptide chain. BiP is an ATPase that provides energy for translocation (Panzner *et al.*, 1995). The mechanism is the following: BiP interacts with Sec63 in its ATP-bound state, when

INTRODUCTION

ATP is hydrolyzed BiP binds to the polypeptide chain in its ADP state, preventing it from sliding back to the cytosol. When the nascent chain has sufficiently moved into the ER lumen, the next BiP chaperone binds (Park and Rapoport, 2012). This process is repeated until the polypeptide has completely traversed the channel. Once the translocation is terminated, exchange of ADP for ATP opens the peptide-binding pocket and releases BiP molecules (Figure 10).

i.3.4. Insertion into the ER membrane

As membrane proteins traverse the translocon channel, TM segments are recognised by the translocon and inserted into the lipid bilayer by the lateral opening of the channel. Once the channel plug is opened, the lateral gate of the translocon is opening and closing continually, exposing the regions of the proteins that are currently in the translocon (aqueous) channel to the hydrophobic milieu of the membrane. Hydrophobic segments will be directed to the membrane through the lateral opening of the translocon. The size of this aperture suggests that TM segments leave the channel one at a time (Heinrich and Rapoport, 2003) or in pairs (Saurí *et al.*, 2005, 2007), although there is also evidence of insertion of several helices as a bundle (Sadlish *et al.*, 2005) (Figure 11).

In this sense, the translocon complex would allow interactions between different TM domains of the same protein before the integration process has completely finished, which facilitates the insertion in the bilayer of segments that would not be able to integrate by themselves, like marginally hydrophobic TM segments (White and von Heijne, 2008; Hedin *et al.*, 2010; Tamborero *et al.*, 2011; Öjemalm *et al.*, 2012).

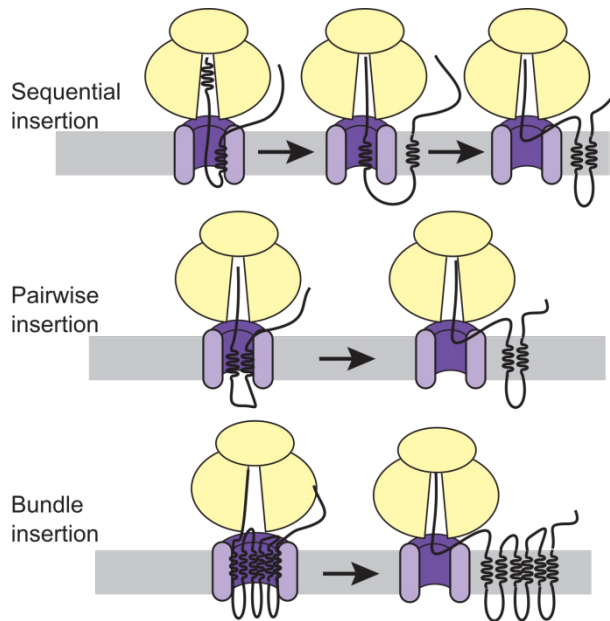


Figure 11. TM integration. During membrane protein biogenesis, the TMHs are transferred laterally from the proteinaceous environment of the translocon into the lipid bilayer. This may occur sequentially with each TMS integrating independently (up), in a pairwise fashion (middle), or in groups (down). Adapted from (Skach, 2009).

i.3.5. Translocon associated proteins

In addition to the Sec61 complex, the translocon machinery has in eukaryotes many other components involved in the translocation/insertion process.

The *signal peptidase* (SP) complex is a protease that removes the N-terminal signal sequence from secreted proteins and some membrane proteins on the luminal side of the membrane (Paetzel *et al.*, 2002). SP has a substrate specificity for small and uncharged residues, such as Ala (Dalbey and von Heijne, 1992). Following the

INTRODUCTION

removal of the signal sequence, the polypeptide is folded, modified, and salt bridges are formed. Moreover, the signal sequence left in the membrane is further cleaved by a signal peptide peptidase (Weihofen *et al.*, 2002).

N-linked glycosylation is one of the most important covalent protein modifications in mammals that occurs on about 1/3 of the proteome, and it is essential for the folding of the protein and even for oligomerization, quality control and transport (Igura *et al.*, 2008). *N*-linked glycosylation occurs co-translationally by the *oligosaccharyltransferase* (OST) multi-subunit protein complex that is closely associated with the translocon (Chavan *et al.*, 2005; Pfeffer *et al.*, 2014). It transfers an oligosaccharide to the Asn residue of the glycosylation site (Asn-X-Ser/Thr, being X any amino acid except Pro). The catalytic site of the OST is located on the luminal side of ER membrane (Igura *et al.*, 2008; Pfeffer *et al.*, 2015), so only glycosylation sites located at the ER lumen can be modified (Figure 12).

TRAM (translocating chain-associating membrane protein) is a glycoprotein with eight TM segments and an N-/C-terminal cytosolic orientation (Tamborero *et al.*, 2011) that is involved in the translocation and insertion of secreted and membrane proteins. It has also been suggested to function as a membrane protein chaperone for short, weak signal sequences and poorly hydrophobic TM segments like those harbouring abundant charged residues (Goerlich *et al.*, 1992; Heinrich *et al.*, 2000).

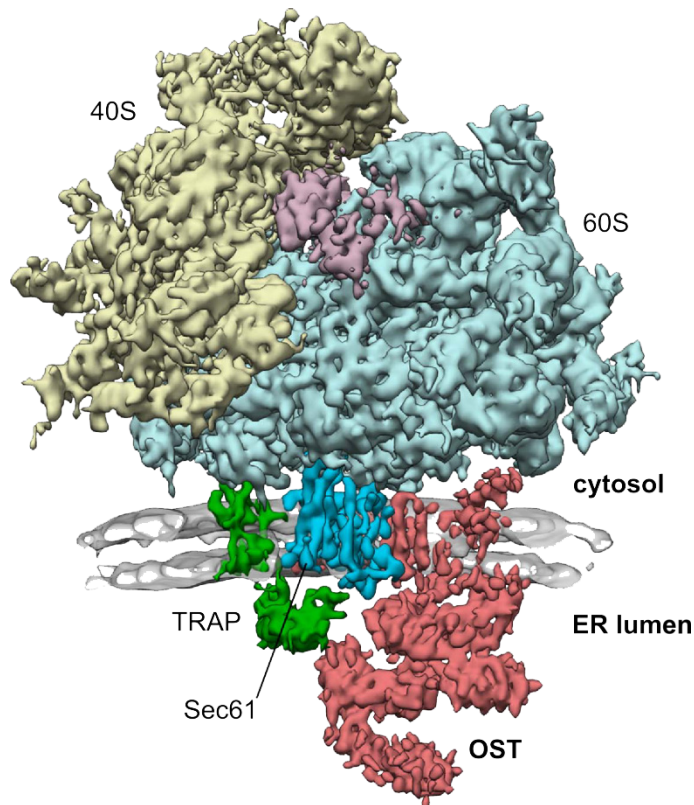


Figure 12. Overall structure of the ER membrane-associated mammalian ribosome. Segmented densities for the 40S (yellow) and 60S (light blue) ribosomal subunits, translation elongation factors (magenta), Sec61 (blue), TRAP (green) and OST (red) complexes. Density for the ER membrane is shown in grey. Adapted from (Pfeffer *et al.*, 2015).

Other component of the translocon machinery is *TRAP* (translocon-associated protein). Although its main function is still unclear, it has been proposed to be involved in the co-translational insertion or translocation of newly synthesized eukaryotic proteins. *TRAP* is a tetrameric protein complex (α , β , γ and δ) of integral membrane proteins (Hartmann *et al.*, 1993) and is associated with ribosome–Sec61 complexes with a 1:1 stoichiometry (Ménétret *et al.*,

INTRODUCTION

2008). It has been proposed that TRAP facilitates the initiation of protein translocation (Fons *et al.*, 2003), although the details of the mechanism remain unknown. The α , β , and δ subunits are single spanning membrane proteins, whereas the γ subunit crosses the membrane four times (Baño-Polo *et al.*, 2017). More recently, the molecular organization of the TRAP complex have been revealed by cryo-electron tomography (Pfeffer *et al.*, 2017).

I.4. TOPOLOGY OF MEMBRANE PROTEINS

i.4.1. Classification of membrane proteins according to their topology

The topology of a membrane protein (number of TM segments and the relative location of its extramembranous domains on either side of the membrane) is fundamental to being able to carry out its biological function. In general, a protein can adopt a single topology in the membrane, although cases have been identified in which the same sequence is able to be inserted in the bilayer with two opposite topologies (Rapp *et al.*, 2006, 2007; Seppälä *et al.*, 2010). According to their topology, four large groups of membrane proteins can be identified (Goder and Spiess, 2001).

Type I proteins are those with cleavable SS, therefore they have their N-terminal end in the extracytoplasmic region. There are some cases of proteins with a short loop between the SS and the first TM domain, in which an inversion of the topology has been observed once the SS has been cut (Stewart *et al.*, 2001). Type II and Type III are SS-deficient membrane proteins in which the first TM segment is oriented with its N-terminal towards the cytosol or the lumen,

respectively (Figure 13). In multispanning proteins is the SS or in its absence the first hydrophobic segment the main responsible for the targeting as well as for establishing their global topology. In addition to these three groups it could be considered a fourth (type IV) formed by proteins anchored to the membrane by its C-terminal end. The insertion mechanism of these proteins will necessarily locate the large N-terminus towards the cytosol.

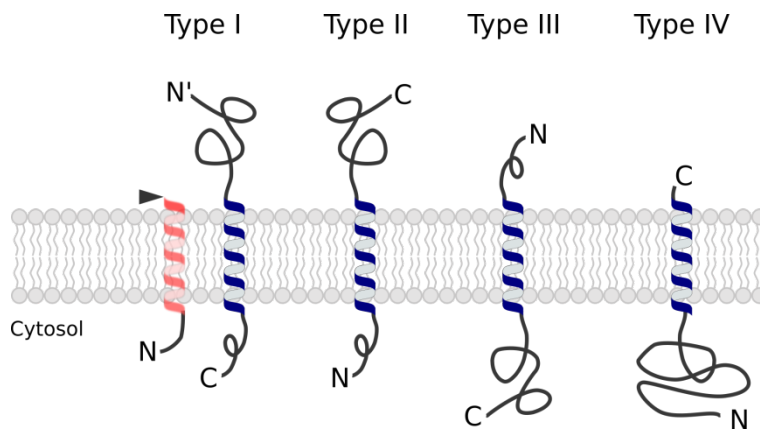


Figure 13. Types of membrane proteins according to their topology. Type I are N_{out} with cleavable SS. Type II and III do not have SS, their first TM segment has a $N_{in}-C_{out}$ and $N_{out}-C_{in}$ topology respectively. Type IV commonly also lacks SS and is anchored to the membrane by the C-terminus, with the N-terminus oriented to the cytosol. Classification based on (Goder and Spiess, 2001).

i.4.2. Topological determinants

The positive-inside rule

The major determinant for the topology of a membrane protein is the so called positive-inside rule, which states that non-translocated loops contain two to four times more positively charged residues (Arg and

INTRODUCTION

Lys) than the translocated ones (von Heijne, 1986a). Charges at the N-terminal side of the SS may influence the orientation of the N-terminal hydrophobic segment of a membrane protein. If it has positive charges, the N-terminal part will remain in the cytosol and the C-terminus will be exposed to the outside or the ER lumen (von Heijne, 1986b). One explanation could be that the positive charges are arrested in the cytosol (Johansson *et al.*, 1993; Fujita *et al.*, 2011; Yamagishi *et al.*, 2014), probably because they interact with the negatively charged lipid head groups (Van Klompenburg *et al.*, 1997). In addition, the translocon may be involved in the initial orientation of the TM by following the positive inside rule; conserved negative charges at the Sec61p (homologous to mammalian Sec61 α) of yeast seem to interact with the nascent chain providing the driving force for signal orientation (Goder *et al.*, 2004). More recently the *Charge Balance Rule* (Bogdanov *et al.*, 2014) was proposed as an extension of the positive-inside rule, and it suggests that the net zero charge of neutral lipids reduces the translocation potential of negatively charged residues in favour of the cytoplasmic retention potential of positively charged residues. This explains why positively charged residues are more potent topological signals than negatively charged residues.

Transmembrane length

The length of the hydrophobic sequence can also determine the orientation of TM segments. Longer sequences have a preference to localize the N-terminus in the lumen (Sakaguchi *et al.*, 1992; Wahlberg and Spiess, 1997; Eusebio *et al.*, 1998). Cleavable signal sequences have shorter hydrophobic segments than TM segments orienting the N-terminal part to the cytosol (Nilsson *et al.*, 1994), confirming the TM length as a determinant in defining the orientation of the hydrophobic helix.

Hydrophobicity

Besides the length of the sequence, its degree of hydrophobicity and the hydrophobicity gradient can influence the orientation of a TM domain in the bilayer. The more hydrophobic the TM segment is the more it tends to insert with an N_{out}-C_{in} orientation independently of the charges of the flanking region (Wahlberg and Spiess, 1997; Goder and Spiess, 2003). An explanation for this observation is that very hydrophobic helices may rapidly exit the lateral gate by interacting with the hydrophobic core of the lipid bilayer (Heinrich *et al.*, 2000). In contrast, natural signal sequences (N_{in}-C_{out}) are less hydrophobic and remain longer in the translocon allowing them to reorient during nascent chain translation (Whitley and Mingarro, 2014).

Long loops and globular domains

Long loops reach further from the membrane surface to localize in the aqueous environment and behave as other globular (soluble) domains. If the N-terminal tail is long or rich in charged residues it will stay in the cytoplasm and will not be translocated (Andersson and von Heijne, 1993; Nouwen *et al.*, 2009), leading to the N_{in}-C_{out} orientation of the segment. Similarly, it has been observed that the folding state of an extramembrane domain preceding a TM segment precludes its translocation, and consequently forces the TM segment towards an N-terminal cytoplasmic orientation (Denzer *et al.*, 1995). However, if the N-terminal tail is translocated, the following TM domain will adopt a N_{out}-C_{in} orientation.

Lipid composition

In addition to affecting the assembly and structure of membrane proteins, lipid composition can also affect their topology. Some studies have shown that membrane proteins from *E. coli* can adopt

INTRODUCTION

different topologies depending on the lipid composition (Bogdanov *et al.*, 1996; Zhang *et al.*, 2003, 2005). The topology of lactose permease (LacY) was largely altered when inserted in a membrane lacking phosphatidylethanolamine (PE), signifying the importance of lipid influence on topology. Interestingly, by inducing PE synthesis post-assembly of LacY, the TM domains could reorient to their native orientation, indicating that the process is completely reversible (Bogdanov *et al.*, 2002).

I.5. COMPUTATIONAL METHODS IN MEMBRANE PROTEINS

Despite the enormous advances in the field, it is still a challenge to work with membrane proteins, as they are unstable outside the membrane environment, and then difficult to handle biochemically. To stabilize them outside the membrane, they need to be reinserted into a kind of lipid, like detergent milieu, or by adding ligands or inhibitors, or by introducing point mutations to generate a more stable protein in solution (Vinothkumar and Henderson, 2010). Moreover, experimental methods used to determine the three-dimensional structure of proteins are expensive, time-consuming and require crystallization, which is especially difficult for membrane proteins. Although the progress in the determination of membrane protein structures follows an exponential growth (White, 2004) ([Figure 14](#)), there is still a huge gap between the number of structures of membrane proteins and globular ones. All these difficulties underline the importance and the need for automated computational tools to identify and to predict the potential structure of membrane proteins.

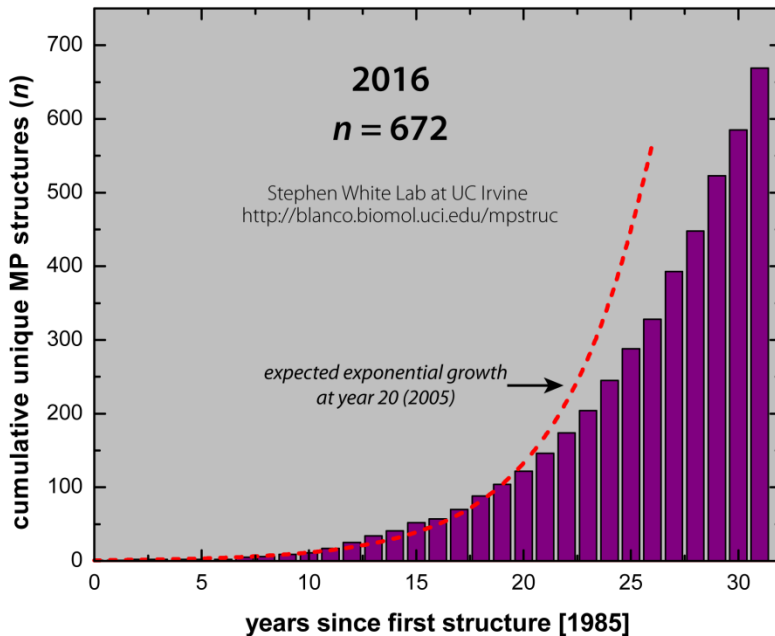


Figure 14. Growth in TM protein structure determination since the release of the first structure (1985), showing an exponential trend. This exponential trend has been reduced in the last decade relative to the expectations in 2005 (red dashed line). Figure created at Stephen White's lab (<http://blanco.biomol.uci.edu/mpstruc/>) (White, 2004).

i.5.1. Statistical and machine learning methods

Early predictions of TM segments generally used a scale-based analysis, which follows a four-step procedure: **1.** Each amino acid is given a value derived from its physico-chemical properties, generating a 'propensity scale'. **2.** Using these values, the amino acid sequence is transformed into a sequence of values generating a plot of propensity values among the sequence. **3.** Smooth the plot by taking the average propensity value in a window of n residues and plot the average at the center of the window (sliding window average). **4.** Identify TM

INTRODUCTION

regions by scanning the smoothed plot for regions with values higher than a predetermined threshold.

Modern approaches in membrane protein topology prediction are machine learning methods. The goal of machine learning is to develop algorithms which can learn by themselves to solve problems. Many algorithms used in topology prediction are based on hidden Markov models (HMMs) and artificial neural networks (ANNs).

HMMs are probabilistic models that are suitable for modelling a wide range of sequence based problems (Rabiner and Juang, 1986). It consists of a set of interconnected states, each of which emits an observable output symbol. To each state there are two types of parameters, emission probabilities, which are the probabilities of emitting each symbol, and transition probabilities, the probabilities of moving from one state to some other state. When predicting the topology of TM proteins using HMMs (Sonnhammer *et al.*, 1998; Tusnady and Simon, 1998; Krogh *et al.*, 2001; Bernsel *et al.*, 2008; Viklund and Elofsson, 2008), the basic idea is to build a model, where each state is labelled with either “Membrane”, “Inside” or “Out-side” and define the possible transitions between the states so that they agree with the grammar of topology. Then, using sequences with experimentally known topologies, the transition and emission probabilities are adjusted. One advantage of using HMMs is that the final prediction is guaranteed to be globally optimal, since the predicted location for a particular residue depends on the most likely state path through the model for the whole sequence.

ANNs are a set of algorithms for information processing, which are designed to mimic the functioning of brain synapses (McCulloch and Pitts, 1943; Hebb, 1949; Rosenblatt, 1958). As with HMMs, ANNs in membrane topology predictions (Rost *et al.*, 1995; Rost *et al.*, 1996; McGuffin *et al.*, 2000; Jones, 2007) are generally used to

classify each residue in a sequence as either “Inside” (i), “Outside” (o) or “Membrane” (M). For each position in the sequence, the input is commonly a window of residues centered on this position, and the output is a value between 0 and 1 for each of the structural categories (i, o and M). Using ANNs for topology prediction usually involves two steps, a parameter estimation step (training) and a prediction step. In the parameter estimation step the weights of the neural network are optimized using sequences with known topology. These estimated parameters are then used for predicting sequences with unknown topologies. Neural networks are most commonly used for residue level predictions.

i.5.2. Hydrophobicity scales

Hydrophobicity scales are tables that assign a value to each amino acid residue according to their tendency to interact with surrounding water. Based on this, an estimation of which amino acids are preferred in membrane regions can be made. For membrane proteins prediction, these tables are used to define the free energy needed to insert an amino acid into the hydrophobic core of the membrane. Hydrophobicity scales are based on the partitioning of amino acids between two immiscible liquid phases, chromatographic techniques or accessible surface area calculations. Some of the numerous available scales are chronologically outlined below.

One of the most frequently cited hydrophobicity scale, the Kyte-Doolittle scale (Kyte and Doolittle, 1982), combines accessible surface area measurements in globular proteins with water-vapor partitioning preferences. By implementing this scale in a sliding-window approach, it was possible both to distinguish exterior from interior in globular proteins, and to identify TM regions in membrane proteins.

INTRODUCTION

Another hydrophobicity scale specifically adapted to TMHs is the Goldman-Engelman-Steitz scale (Engelman *et al.*, 1986), in which a semi-theoretical approach is taken, accounting for the attachment of side chains to an α -helical backbone structure.

The Wimley-White scale (White and Wimley, 1999) in addition to compute the hydrophobicity of isolated amino acids, takes into account the contribution of the backbone peptide bonds to partition into the bilayer. Partitioning of pentapeptides into palmitoyloleoylphosphatidylcholine (POPC) bilayer interfaces (Wimley and White, 1996) and n-octanol (Wimley *et al.*, 1996) were used to build this scale.

The Hessa scale (Hessa *et al.*, 2005, 2007) is a biological hydrophobicity scale based on experimental results. The authors challenged the Sec61 translocon with a large set of systematically designed TM sequences and measured the efficiency of membrane integration for each one. Then, they used the quantitative data generated in this way to measure contributions from individual residues and to calculate the energy needed to insert each amino acid.

The Zhao-London scale (Zhao and London, 2006) is based on propensities of amino acids in protein structures. They applied hydrophobicity analyses to databases of soluble and transmembrane proteins of known structure and this information was used to define a refined hydrophobicity-type TM sequence prediction scale. The refinement procedure involved adjusting scale values to eliminate differences between the average amino acid composition of populations TM and soluble sequences of equal hydrophobicity.

i.5.3. Membrane protein predictors

Transmembrane domain predictors

The *DAS* server, presented in 1997 (Cserzö *et al.*, 1997), uses the so-called Dense Alignment Surface (DAS) method, which is based on low-stringency dot-plots of the query sequence against a collection of non-homologous membrane proteins using a previously derived special scoring matrix. In 2002, the method was updated to *DAS-TMfilter* (Cserzo *et al.*, 2002), which in a second prediction cycle predicts TM segments in the sequences of the TM library.

SOSUI (a Japanese word that means “hydrophobic”) (Hirokawa *et al.*, 1998) classifies and predicts secondary structure of membrane proteins taking into account the known helical potentials of the given amino acid sequence. *SOSUI* uses 4 characteristics of the amino acids in its prediction: ‘hydrophobicity index’ (Kyte and Doolittle, 1982), ‘amphiphilicity index’ (weighted presence of amphiphilic amino acids and their localization), the charge of the amino acids and the length of the sequence.

The *ΔG Prediction Server* (<http://dgpred.cbr.su.se/>) is based on experimental results from systematically designed 19-residue long amino acid sequences that have been expressed and tested for TM insertion using an *in vitro* assay (Hessa *et al.*, 2005, 2007). Given the amino acid sequence of a putative TM helix, the server gives a prediction of the corresponding apparent free energy difference, ΔG_{app} , for insertion of this sequence into the ER membrane. The server runs in two different modes, for two different types of queries: prediction of ΔG_{app} for membrane insertion of a potential TMH or scan a full protein sequence for putative TMHs.

INTRODUCTION

Topology predictors

One of the first topology predictors was *TopPred* (von Heijne, 1992). It uses the Goldman-Engelman-Steitz scale (Engelman *et al.*, 1986) to create a hydrophathy plot of the sequence, with a sliding window approach. All hydrophobicity peaks above a first cut-off value are identified as 'certain' TM segments, whereas all peaks below this cut-off but above a second lower cut-off value are marked 'putative' TM helices. Additionally, it solved the orientation problem by considering the positive inside rule (von Heijne, 1986a). Few years later, an improved version, named TopPred II, was compiled to include a n option in which the unfavourable free energy of membrane insertion of charged residues in the TM segments can be reduced by means of the "Charge-pair Energy" parameter if they can form $i, i+3$ or $i, i+4$ charge-pairs (Claros and Heijne, 1994).

MEMSAT (MEMbrane protein Structure And Topology) is based on the combination of a scoring matrix developed with a dynamic programming algorithm (Jones *et al.*, 1994). The scoring matrix defines five different states: inside loop, outside loop, inside helix end, helix middle, outside helix end. Latter versions of the program have introduced main changes, as the use of ANNs and sequence profiles (McGuffin *et al.*, 2000; Jones, 2007).

PHDhtm was published in 1995 as the first ANN-based method incorporating evolutionary information (Rost *et al.*, 1995; Rost *et al.*, 1996). The method takes a sequence profile as input and predicts a preference score for each profile column based on a sliding window of neighbouring columns. A topology is then generated using a dynamic programming algorithm, and the overall orientation is determined by the positive-inside rule.

INTRODUCTION

The *TMHMM* (TransMembrane Hidden Markov Model) (Sonnhammer *et al.*, 1998) and *HMMTOP* (Tusnády and Simon, 1998) methods were introduced in the late nineties and use HMMs for topology prediction. An advantage of HMMs is that biological knowledge can be easily translated into flexible models. The models presented by both groups show many similarities: both separate the helix into a core and two different end states and they differentiate between inside/outside loop. Later on, the *PRO-TMHMM* (Viklund and Elofsson, 2004), which is a development of *TMHMM*, was introduced. *PRO-TMHMM* uses sequence profiles rather than single sequences that can be scored against the model to improve prediction performance.

The first version of *Phobius* (Käll *et al.*, 2004) was published in 2004. The main aim was to create a topology predictor that could identify signal peptides, in turn reducing prediction errors by assigning TM regions as signal peptides and *vice versa*. Knowledge about the appearance of signal peptides also helps determining the orientation of a protein inside a membrane. Shortly after the first version was released, an extension called *PolyPhobius* (Käll *et al.*, 2005) was released. Its improvement is a hidden Markov decoding algorithm that uses evolutionary information by including for each position in a multi-sequence alignment the average of the previous label probability. This helps to improve the prediction performance for both TM regions and signal peptides. *Philius* (Reynolds *et al.*, 2008) is based on the work done to create *Phobius*. Similarly, it can predict signal peptides in addition to the TMHs. *Philius* also provides confidence scores for its predictions and predicts the signal peptide cleavage site.

TMDET (Tusnády *et al.*, 2004) is an approach to distinguish between transmembrane and globular proteins using structural

INTRODUCTION

information only and to locate the most likely position of the lipid bilayer. An automated algorithm determines the membrane planes relative to the position of atomic coordinates, together with a discrimination function which is able to separate transmembrane and globular proteins even in cases of low resolution or incomplete structures such as fragments or parts of large multi-chain complexes.

SCAMPI (Bernsel *et al.*, 2008) is based on the idea to take biological knowledge and translate it into a computational model. It uses a simple HMM with only two optimized parameters: the positive-inside rule and the Hessa hydrophobicity scale (Hessa *et al.*, 2005). Although it is a simple method, it achieves results similar to those of more complex methods (Peters *et al.*, 2016).

OCTOPUS (Viklund and Elofsson, 2008) modelled for the first time a variety of regions that do not completely span the membrane. It is the first method to implement an HMM with an alphabet consisting of residue preferences predicted by an ANN. Shortly after the release of *OCTOPUS*, *SPOCTOPUS* (Viklund *et al.*, 2008) was released to address a well-known weakness of many prediction algorithms that often confuse TM regions with signal sequences.

Finally, *TOPCONS* was introduced in 2009 (Bernsel *et al.*, 2009). It is a meta-predictor that runs five other predictors (*OCTOPUS*, *Philius*, *PolyPhobius*, *SCAMPI* and *SPOCTOPUS*) to combine their output into a topology profile, and uses this profile as input for a simple HMM. Basically, it filters predictions provided by the other methods to assess the likely global topology. In 2015 an updated version was published (Tsirigos *et al.*, 2015), in which the HMM was extended such that it is now able to separate between signal peptides and the rest of the protein.

i.5.4. Membrane proteins databases

Several databases have been built as repositories of membrane protein sequences and/or structures, and for the purpose of helping computational biologists to develop and test their prediction methods. Some of the most frequently used membrane proteins databases are described below.

Stephen White's manually curated database (<http://blanco.biomol.uci.edu/mpstruc/>) contain a current list of membrane protein structures determined by X-ray, NMR and electron diffraction with links to the *Protein Data Bank* (PDB) (<http://www.rcsb.org/>) (Berman *et al.*, 2000) and *PubMed* (<https://www.ncbi.nlm.nih.gov/pubmed>) entries. From the same group, *MPtopo* (Jayasinghe *et al.*, 2001) is a curated database of membrane proteins with experimentally validated TM segments.

PDBTM (Protein Data Bank of Transmembrane Proteins) database (<http://pdbtm.enzim.hu/>) (Tusnády *et al.*, 2004, 2005a; Kozma *et al.*, 2013) is a comprehensive and up-to-date transmembrane protein structures selection of the PDB. The *PDBTM* database was created by scanning all PDB entries with the *TMDET* algorithm (Tusnády *et al.*, 2005b) which also predicts the membrane orientation of protein structures.

The *OPM* (Orientation of Proteins in Membranes) database (<http://opm.phar.umich.edu/>) (M. A. Lomize *et al.*, 2006) provides as a significant novelty spatial arrangements of membrane proteins with respect to the hydrocarbon core of the lipid bilayer. *OPM* includes all unique experimental structures of transmembrane proteins and some peripheral proteins and membrane-active peptides. Each protein is positioned in a lipid bilayer of adjustable thickness with the PPM (Positioning of Proteins in Membranes) algorithm (A. L. Lomize *et*

INTRODUCTION

al., 2006; Lomize *et al.*, 2011). PPM program can be applied to newly determined experimental protein structures or theoretical models.

MemProtMD (<http://sbc.bioch.ox.ac.uk/memprotmd/beta/>) (Stansfeld *et al.*, 2015) is a database of all known membrane proteins identified in the PDB inserted into simulated lipid bilayers using Coarse-Grained Molecular Dynamics simulations. The simulations are analyzed for protein-lipid interactions, identifying lipid binding sites, and revealing local bilayer deformations plus molecular access pathways within the membrane. The coarse-grained models of membrane protein/bilayer complexes are then transformed to atomistic resolution for further analysis and simulation.

OBJECTIVES

OBJECTIVES

The general objective of this thesis is the understanding of the insertion and assembly of TM α -helices into biological membranes, and how this process is affected by hydrophobic sequence length and amino acid composition. During the study of these transmembrane α -helices, this thesis has addressed the following specific objectives:

- To describe the differences between TM and water-soluble helices in terms of length and amino acid composition.
- To study the amino acid distribution patterns statistically along TM helices.
- To analyze the predicted insertion of computationally designed sequences with different length and amino acid composition.
- To validate experimentally our predictions by analysing its membrane integration capacity using an *in vitro* translation/glycosylation system.

METHODOLOGY

M.1. COMPUTATIONAL METHODS

m.1.1. Helix data sets

Two data sets of water-soluble and TM helices were obtained from the PDB (Berman *et al.*, 2000) and the MPTOPO databases (Jayasinghe *et al.*, 2001), respectively.

First, the water-soluble dataset was built by selecting a total of 4,405 structural chains deposited in the PDB (as of November 17th, 2011) that passed the following criteria were selected: (i) their total secondary structure had more than 60% of α -helices and no β -strands; (ii) their crystallographic resolution was 2.0 Å or higher; and (iii) the word MEMBRANE did not appear in the “TITLE” nor the “DESCRIPTION” fields of the PDB file. Furthermore, to remove redundancy, the 4,405 chain sequences were compared to each other with the CD-HIT program (Huang *et al.*, 2010) and pairs resulting in sequence alignments with 80% or higher identity were discarded. The final set of 930 non-redundant PDB chains was parsed to identify a total of 7,348 helices from “HELIX” fields of each PDB chain entry. Thus, the data set of water-soluble helices contained 930 non-redundant and high-resolution protein structures, 7,348 α -helices and 108,277 amino acids.

Second, all α -helical membrane proteins deposited in the MPTOPO database (last updated on January 19th, 2010) (Jayasinghe *et al.*, 2001), and thus with known membrane insertion topology, were selected. The initial set was further filtered by: (i) removing any entry of unknown structure as based on the MPTOPO entry classification (*i.e.*, keeping only entries described as “3D_helix” and “1D_helix”); (ii) removing redundant pairs at 80% sequence identity by applying the CD-HIT program (Huang *et al.*, 2010). The final data set of TM helices contained 170 non-redundant structures, 837 TM helices, and

METHODOLOGY

20,079 amino acids. Furthermore, to properly analyze the amino acid propensities in single membrane spanning TM helices, we discarded any helix shorter than 17 amino acids or larger than 38 amino acids. The resulting TM data subset contained 792 TM helices, and 19,356 amino acids.

m.1.2. Amino acid propensity

We calculated three different amino acid measures: (i) probability and percent, (ii) Odds, and (iii) LogOdds. The probability (p_i) of an amino acid i is defined as:

$$p_i = \frac{n_i}{N} \quad [1]$$

where i is the amino acid type (one of the 20 amino acids), n_i is the observation count of the amino acid i , and N is all amino acids in the data set. Similarly, the percent of a given amino acid i is defined as its probability multiplied by 100. The Odds (O_i) of an amino acid i is defined as:

$$O_i = \frac{p_{i,c}}{(1-p_{i,c})} / \frac{p_{i,r}}{(1-p_{i,r})} \quad [2]$$

where $p_{i,c}$ is the probability of the amino acid i in the class c (for example, TM helix) and $p_{i,r}$ is the probability of the amino acid i in the class r (for example, water-soluble helix). Similarly, the LogOdds of a given amino acid i is defined as the logarithm in base 10 of its Odds. Briefly, Odds higher than 1 (or positive LogOdds) indicate over-occurrence of the amino acid type in the class. Odds smaller than 1 (or negative LogOdds) indicate under-representation of the amino acid type in the class.

m.1.3. Computational sequence design

Using the created TM helices dataset described above (792 TM helices and 19,356 amino acids), we generated a series of designed sequences of a given length by populating them with an increased number of amino acid types.

First, we generated sequences with only Leu amino acid type (the most common in TMs) of lengths ranging from 9 to 25 residues. Next, we included the second most common amino acid type in TMs (that is, Ala) with its relative probability compared to Leu. Again, we generated 1,000 sequences of each length where the sequences were obtained by shuffling a string of L and A letters with the proportions between them as found in our data set (58% L and 42% A). The same procedure was repeated each time including a new amino acid type following the order found in our dataset, in terms of abundance, until we had all 20. (See [Figure 15](#)).

Next, we generated a set of sequences with both, amino acid propensity as well as positional effect by taking TMs residues and annotating their position with respected the center of the TM helix. In such case, we computed the probability of an amino acid type in each of the positions of a TM starting from the central one (position 0) and increasing the positive number as we approached the cytoplasmic side of the TM or a negative number as we approached the extracellular side of the TM. The designed sequences were built taking into account the position by selecting amino acids for each pool across the TM helices. Similarly to the non-position dataset, 1,000 sequences of each different length were generated for each amino acid type compositions. (See [Figure 15](#)).

METHODOLOGY

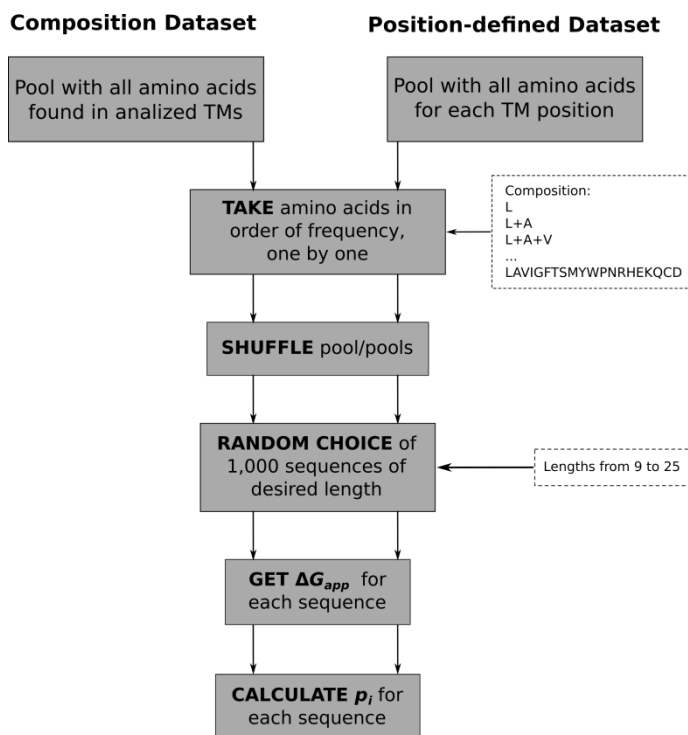


Figure 15. Computational flowchart for designing TM sequences as well as predicting their insertion probability using the experimentally based ΔG Prediction Server (<http://dgpred.cbr.su.se/>).

m.1.4. Prediction of the ΔG values and probability of insertion

All sets of 1,000 sequences of desired length, amino acid composition and non-positional or positional effect were used to generate a series of predicted insertion capacity scores, which were predicted using the experimentally based ΔG Prediction Server (<http://dgpred.cbr.su.se/>) (Hessa *et al.*, 2005, 2007). In this program, the apparent ΔG is described as:

$$\Delta G_{app} = -RT \ln K_{app} \quad [3]$$

where $T = 298$ K, $R = 0.0019872$ kcal/(K·mol) and the K_{app} is the equilibrium constant between the fraction of molecules inserted (f_i) and the fraction of molecules translocated (f_t):

$$K_{app} = \frac{f_i}{f_t} \quad [4]$$

Since

$$f_i + f_t = 1 \quad [5]$$

and the fraction of molecules inserted (f_i) is the same than the probability of insertion (p_i), we can rewrite the equation 3 as follows:

$$\Delta G_{app} = -RT \ln \left(\frac{p_i}{1-p_i} \right) \quad [6]$$

Thus, by reorganization of equation 6 we can obtain the probability of insertion (p_i) from the ΔG value given by the predictor:

$$p_i = \frac{e^{\left(\frac{\Delta G_{app}}{-RT}\right)}}{1 + e^{\left(\frac{\Delta G_{app}}{-RT}\right)}} \quad [7]$$

All the analysis mentioned above were performed with Python (Python Software Foundation, version 2.7, <https://www.python.org/>),

METHODOLOGY

R (The R Foundation, <https://www.r-project.org/>) and RStudio (<https://www.rstudio.com/>).

M.2. EXPERIMENTAL METHODS

m.2.1. Biological material

E. coli

The strain used for the routine extraction of plasmid DNA was *E. coli* DH5 α (genotype: *dlacZ* Δ M15 Δ (*lacZYA-argF*) U169 *recA1 endA1 hsdR17(rK-mK+)* *supE44 thi-1 gyrA96 relA1*) (Taylor *et al.*, 1993). The competent cells were prepared in the laboratory following the protocol from (Sambrook and Russell, 2001).

Growing conditions

For the growth of the bacteria we used LB media (*Luria Bertani*, composed of yeast extract at 0.5% (w/v), triptone at 1% (w/v) and NaCl at 1% (w/v)) liquid or solid (adding bacteriological agar at 2% (w/v)). The media was autoclaved for 20 minutes at 1 atmosphere pressure and 121°C and supplemented with the appropriate antibiotic. Cells were grown at 37°C and 250 rpm (liquid LB) overnight (at least 12 hours).

Thermal shock transformation

To introduce the plasmids into the bacteria, 10 ng of DNA was incubated on ice along with 50 μ L of competent cells for 30 minutes. After this period the cells were subjected to a thermal shock of 42°C for 45 seconds and subsequently incubated for 5 minutes on ice. 500 μ L of LB was then added and the cells were recovered at 37°C for 45

minutes under stirring. Cells were collected by sedimentation at 3,000 rpm for 5 min, 450 μ L of the supernatant was removed and the rest was resuspended with the pellet and seeded on an LB plate with the appropriate antibiotic. The plate was grown at 37°C overnight.

Electric shock transformation

In those occasions that high transformation efficiency was required, the transformation of DNA was made by electroporation. In this case 50 μ L of electrocompetent cells were incubated with 5 ng of plasmid DNA. The cells were then subjected to an electric shock of 1.8 kV/cm² for 5 milliseconds. The distance between the electrodes the electroporation cuvettes was 0.1 cm. After the electric shock the cells were recovered in LB medium by incubating them for 45 minutes at 37°C under agitation, after which the cells were seeded on an LB plate plus the required antibiotic. The plate was grown at 37°C overnight.

m.2.2. DNA manipulation

DNA isolation

Plasmids isolation from *E. coli* and agarose gel purification were made using commercial kits (*GeneJET Plasmid Miniprep Kit* and *GeneJET Gel Extraction Kit* respectively) purchased from Thermo (Ulm, Germany), following the manufacturer's instructions.

Insert construction, vector preparation and ligation

Tested sequences were constructed using two double-stranded oligonucleotides with 5' overlapping overhangs at the ends. Pairs of complementary oligonucleotides were first annealed at 85°C for 10 min followed by slow cooling to 30°C, after which the two annealed double-stranded oligonucleotides were mixed, incubated at 65°C for 5

METHODOLOGY

min and cooled slowly to room temperature. The resulting oligonucleotides were then purified in a 2% agarose gel, phosphorylated, ligated into the vector and transformed into *E. coli*. The vector was previously digested with the required restriction enzymes, dephosphorylated and purified in a 1% agarose gel.

Site-directed mutagenesis

Mutations at the designed TM segments were obtained by site-directed mutagenesis using the commercial kit *Quickchange* from Agilent Technologies (Santa Clara, CA, USA). The reaction mix (2.5 μ L of simple buffer 10x, 1 μ L of dNTPs 25 mM, 250 ng of both oligonucleotides and 2.5 U of Pfu Turbo polymerase) was subjected to 25 amplification cycles (denaturing 50 seconds at 95°C, annealing 1 minute at 58-60°C and elongation 12 minutes at 68°C) in a *Eppendorf Mastercycler Personal* (Hamburg, Germany). The reaction was then digested with *DpnI* enzyme to degrade the parental DNA and transformed into *E. coli*.

All oligonucleotides used in the production of constructs by the annealing method and in site-directed mutagenesis are listed in Annex I.

m.2.3. Glycosylation assay

To analyze the experimental insertion of computed sequences into ER membranes, we used the pGEM-Lep plasmid. In this plasmid, the *E. coli* leader peptidase protein (Lep) is under a SP6 promotor. Lep consists of two TM segments (H1 and H2) connected by a cytoplasmic loop (P1) and a large C-terminal domain (P2), and inserts into ER-derived rough microsomes (RMs) with both termini located in the lumen (Figure 16). The designed sequence (“TM-tested”) was engineered into the luminal P2 domain and flanked by two acceptor

sites (G1 and G2) for *N*-linked glycosylation. The engineered glycosylation sites can be used as membrane insertion reporters because G1 will always be glycosylated by the OST complex (see [Figure 12](#)) due to its native luminal localization, but G2 will be glycosylated only upon translocation of the analyzed region through the microsomal membrane. A singly glycosylated construct in which TM-tested is inserted into the membrane has a molecular mass ~2.5 kDa higher than the molecular mass of Lep expressed in the absence of microsomes; the molecular mass shifts by ~5 kDa upon double glycosylation (*i.e.*, membrane translocation of the TM-tested sequence).

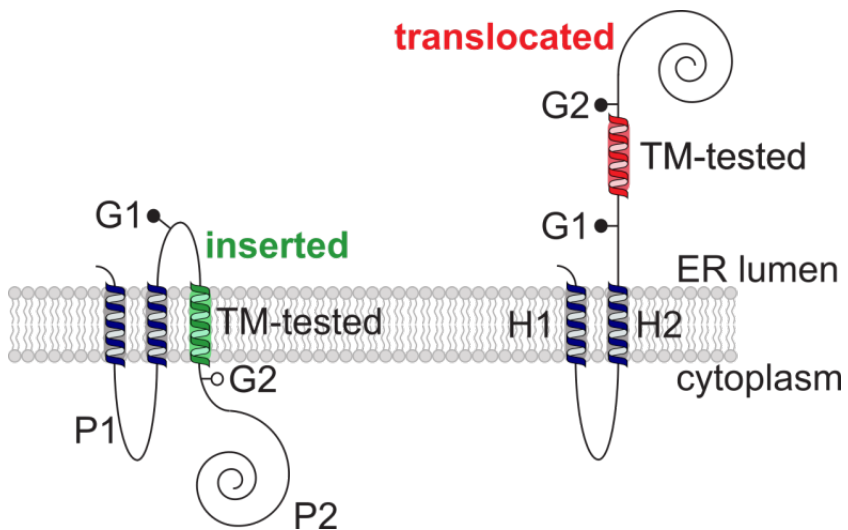


Figure 16. Experimental insertion of segments into the ER membrane. Schematic representation of the engineered Lep model protein. Lep, consisting of 2 TM segments (H1 and H2, blue) and a large luminal domain (P2), inserts into rough microsomes in an N_{lum} - C_{lum} orientation. Computationally designed TM sequences were engineered into the P2 domain with flanking glycosylation sites (G1 and G2). For sequences that integrate into the membrane (green), only the G1 site is glycosylated (left), whereas both G1 and G2 are modified for sequences (red) that do not integrate into the membrane (right).

METHODOLOGY

m.2.4. *In vitro* transcription and translation

Constructs in pGEM-Lep were transcribed and translated in the TNT SP6 Quick Coupled System (Promega). A total of 75 ng DNA template, 0.5 μL ^{35}S -Met (5 μCi) and 0.25 μL microsomes (tRNA Probes) were added at the start of the reaction, and samples were incubated for 90 min at 30 °C. Translation products were diluted in 50 μL of loading buffer (Tris-HCl 625 mM pH 6.8, glycerol 10% (v/v), SDS 2% (w/v), β -mercaptoethanol 4% (v/v) and bromophenol blue 0.025% (w/v)) and analyzed by SDS-PAGE. The gels were quantified using a Fuji FLA-3000 phosphoimager and Image Reader 8.1j software. The membrane-insertion probability of a given TM-tested sequence was calculated as the quotient between the intensity of the singly glycosylated (f_{1g} , inserted) band divided by the summed intensities of the singly glycosylated and doubly glycosylated bands (f_{2g} , translocated):

$$p_i = \frac{f_{1g}}{f_{1g} + f_{2g}} \quad [8]$$

Likewise, the experimental ΔG was calculated with the next formula:

$$\Delta G_{app}^{exp} = -RT \ln K_{app}^{exp} \quad [9]$$

where, in this case, the experimental equilibrium constant (K_{app}) is defined as the quotient between the intensity of the singly glycosylated (f_{1g} , inserted) band divided by the intensity of the doubly glycosylated band (f_{2g} , translocated):

$$K_{app}^{exp} = \frac{f_{1g}}{f_{2g}} \quad [10]$$

Enzymes and chemicals

All enzymes as well as plasmid pGEM1, the TNT SP6 Quick Coupled System and rabbit reticulocyte lysate were from Promega (Madison, WI, USA). ER rough microsomes from dog pancreas were from tRNA Probes (College Station, TX, USA). [³⁵S]Met were from Perkin Elmer (Waltham, MA, USA). The restriction enzymes were purchased from Roche Molecular Biochemicals. The DNA purification kits and site-directed mutagenesis kit were from Thermo (Ulm, Germany). All the oligonucleotides were purchased from Sigma-Aldrich (Switzerland). All TM segment inserts and mutants were confirmed by sequencing of plasmid DNA with the Sequencing Service from MacroGen Europe (Amsterdam, the Netherlands).

RESULTS

R.1. STRUCTURE-BASED STATISTICAL ANALYSIS OF TRANSMEMBRANE HELICES

r.1.1. Helix length in membrane and water-soluble proteins

Length distributions for 837 TM and 7,348 water-soluble helices found in high-resolution structures were analyzed, which turned to be very different ([Figure 17](#)).

Helices in TM proteins are in average $24.0 (\pm 5.6)$ amino acid residues long, this result slightly differs from previous data obtained using databases with 45 (Bowie, 1997) and 129 (Ulmschneider *et al.*, 2005) TM helices, where average helix length was 26.4 and 27.1 amino acid residues, respectively. Because the translation per residue in a canonical helix is 1.5\AA , a stretch of about 20 consecutive hydrophobic residues can span the 30\AA of the hydrocarbon core of biological membranes. Indeed, the more prevalent ($\sim 12\%$) length for TM helices in our data set was 21 residues ([Figure 17](#)).

TM helices shorter than 17 residues as well as larger than 38 residues were excluded in later analysis since they may not cross entirely the membrane ([Figure 17](#), inset a) or may contain segments parallel to the membrane ([Figure 17](#), inset b). In the case of water-soluble helices all lengths were included in our analysis because no restrictions in terms of length can be assumed for water-soluble proteins in an aqueous milieu.

RESULTS

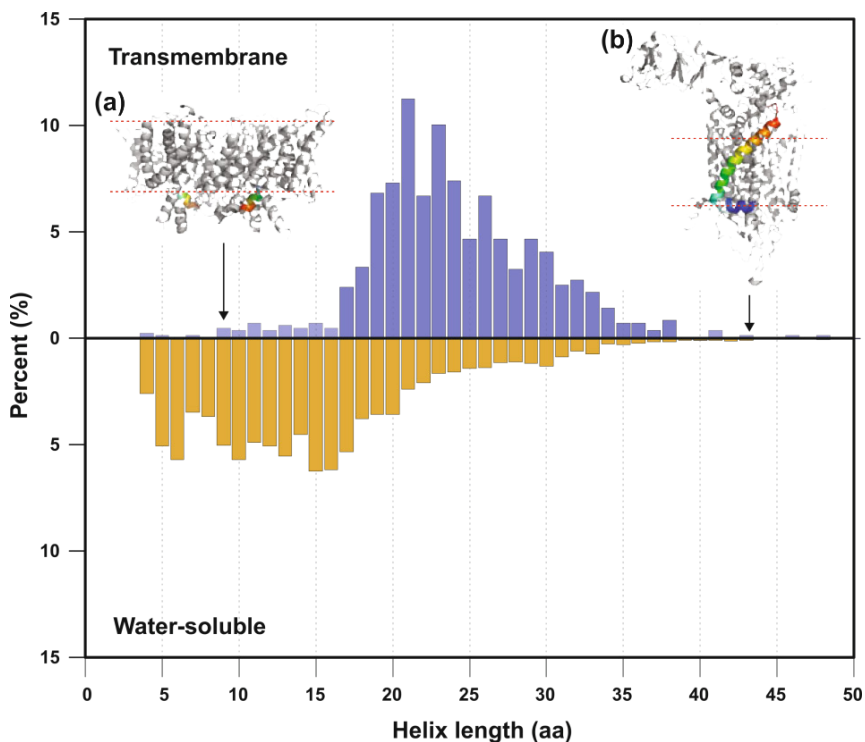


Figure 17. Length distributions for TM and water-soluble helices. Transmembrane helices are shown in blue (pale blue corresponds to discarded lengths) and water-soluble helices are shown in orange. **(a)** Example of a short nine-amino-acid-length helix in the CIC chloride channel from *E. coli* (1KPK entry in PDB). Membrane boundaries were obtained from the PPM server (Lomize *et al.*, 2012). The selected membrane is shown in rainbow coloring from the N-terminal (blue) end to the C-terminal (red) end. **(b)** Example of a large 43-amino-acid-length helix in the chicken cytochrome BC1 complex (1BCC entry in the PDB); the N-terminus of the helix (blue) lies at the membrane/water interface. Representation as in inset **(a)**.

r.1.2. Amino acid composition of α -helices

The amino acid composition for both TM and water-soluble helices was examined (Figure 18). TM helices of lengths between 17 and 38 residues were selected from the MPTOPO database (Jayasinghe *et al.*,

2001), which included helical segments that do completely span the hydrophobic core of the membrane.

As expected, hydrophobic residues Leu, Ala, Val and Ile constitute the bulk of the amino acids in the TM region accounting for almost half (47.0%) of all residues. Similarly, these residues are also frequently found in helices of water-soluble proteins (34.1%). However, there are differences in composition of the two types of helices. Despite sharing the same structural features, the differences between the two types of helices are reflected by their preferential occurrences measured by the logarithm of the Odds of finding a given amino acid in a TM helix with respect to its frequency in a water-soluble helix ([Figure 18](#), bottom panel). While charged and polar residues are much more frequently found in helices from water-soluble proteins, Trp, Gly and Phe have higher propensities in TM helices. Interestingly, in contrast to their conformational preferences in water, the helical propensities of residues such as Val, Ile, Phe and Met are notably increased in the membrane environment. Significantly, Gly and Pro are more frequent in TM helices relative to water-soluble helices.

A comparison of the amino acid frequency between TM and water-soluble helices confirmed that strongly polar residues (Glu, Lys, Asp, Arg, and Gln) are more prevalent in water-soluble helices ([Figure 19](#)). These residues constitute only 8.2 % of the residues within TM helices compared to 30.9 % in water-soluble helices. Conversely, hydrophobic amino acids (Leu, Val, Ile, Gly, and Phe) are over-represented in TM helices ([Figure 19](#)). Interestingly, although being the second more abundant residue in TM helices ([Figure 18](#)), Ala is not over-represented in this type of helices likely due to its limited hydrophobicity (Nilsson *et al.*, 2003).

RESULTS

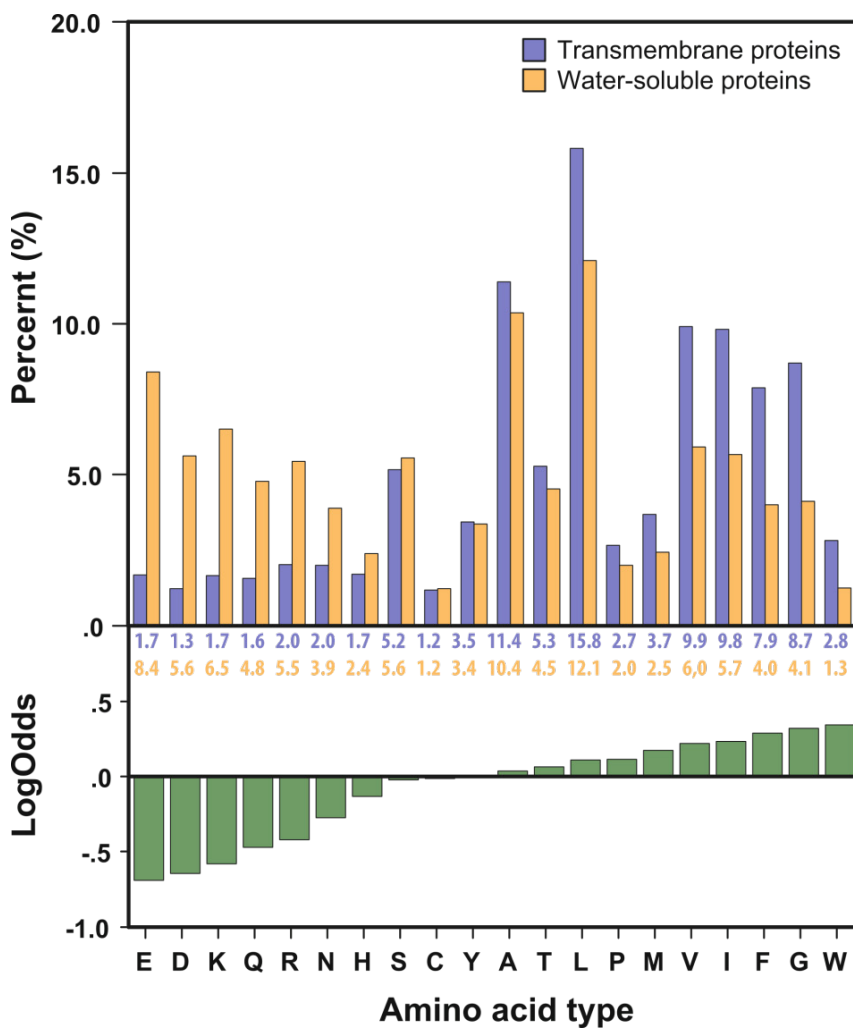


Figure 18. Amino acid type distribution from TM and water-soluble helices. (Upper plot) Amino acid type distribution for TM helices in blue and for water-soluble helices in orange. (Lower plot) LogOdds values for comparison of the relative abundance of each amino acid type in TM and water-soluble helices. Amino acid types are ordered by LogOdds values.

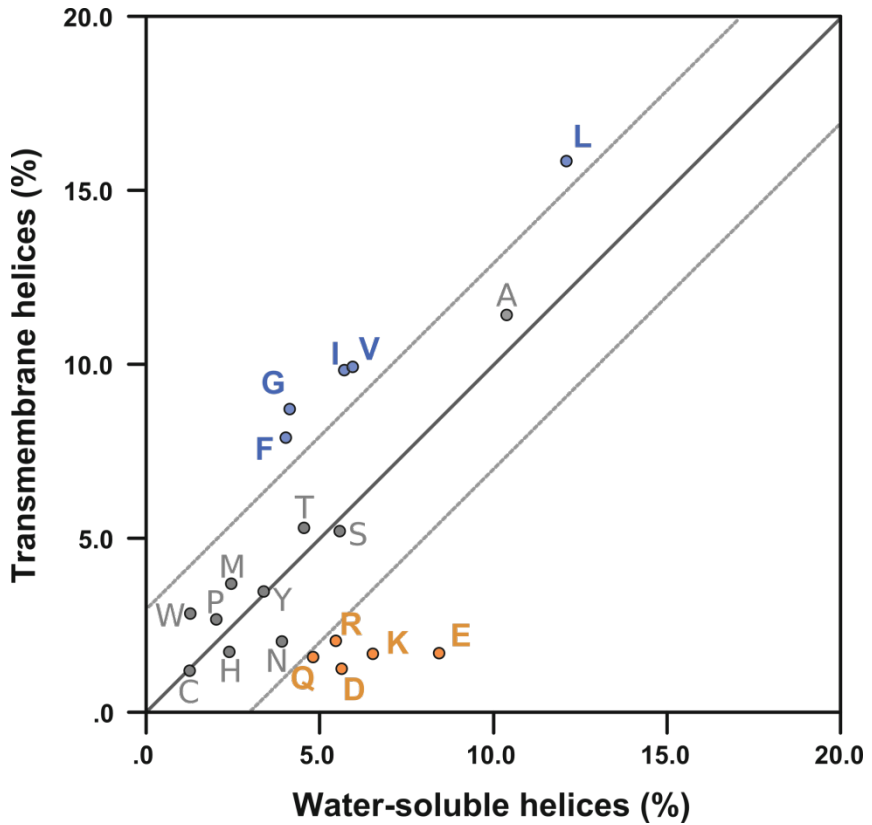


Figure 19. Amino acid type percentage comparison between TM and water-soluble helices. Blue colored amino acids are over-represented (difference >3 % points) in TM helices compared with water-soluble helices. Orange colored amino acids are over-represented (difference >3 % points) in water-soluble helices compared with TM helices. Dashed grey lines indicate a cut-off of 3 % difference points.

r.1.3. Amino acid distribution in TM helices

By taxonomic domains

We decided to analyze the amino acid distribution in TM helices according to taxonomic domains.

RESULTS

We found a total of 46 different species in our database: 1 from *Virus*, 9 from *Archaea*, 29 from *Bacteria* and 11 from *Eukarya* domain. *Bacteria* is the domain that most contributed to our database (70.6 % of the total of 19,356 amino acids analyzed), followed by *Eukarya* (23.7 %) and *Archaea* (5.6%) ([Table 1](#)). Due to the low presence of amino acids from viruses (0.1 %), we decided not to take into account this domain for further separate analysis.

	SPECIES	PROTEINS	TMs	AMINO ACIDS
Virus	1	1	1	22
Archaea	5	9	43	1,083
Bacteria	29	118	566	13,656
Eukarya	11	42	182	4,595
Total	46	170	792	19,356

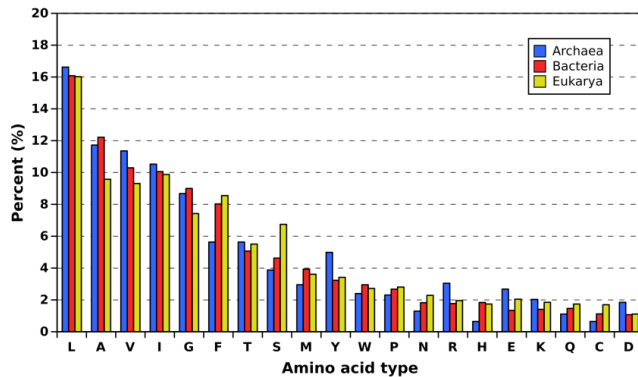
Table 1. Number of species, proteins, TMs and amino acids found in our database according to its taxonomic classification.

The amino acid distribution showed small differences among *Archaea*, *Bacteria* and *Eukarya* domains ([Figure 20 A](#)). Compared to *Bacteria* and *Eukarya* domains, in *Archaea* (blue) Tyr, Arg, Asp and Glu residues are over-represented, while Phe and His are less present, as well as Met, Cys and Asn with minor differences. In eukaryotic proteins (yellow), Ser and Cys have an increased percentage, while Ala, Gly and Val are down-represented. In *Bacteria* (red), Glu and Lys have a reduced presence. Despite these minor differences, the amino distribution is similar in the three domains, with a high correlation coefficient between them (0.97 *Archaea-Bacteria*, 0.95 *Archaea-Eukarya* and 0.98 *Bacteria-Eukarya*). The length distribution of all TM segments was also analyzed according to taxonomic classification ([Figure 20 B](#)). The results showed a slightly displacement of the *Bacteria* distribution peak towards shorter

RESULTS

lengths, probably due to the reduced thickness of bacterial outer membrane. On the other hand, *Archaea* have utterly different membrane phospholipids, with branched isoprene chains instead of fatty acids, L-glycerol instead of D-glycerol, and ether linkages instead of ester linkages, which can affect hydrophobic thickness.

A



B

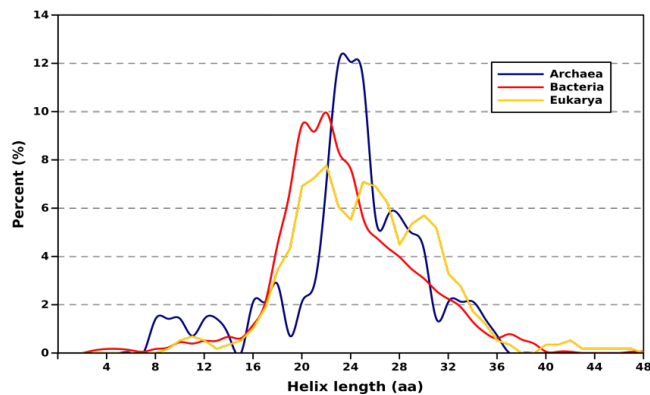


Figure 20. Amino acid and length distribution by taxonomic domains. *Archaea* is showed in blue, *Bacteria* in red and *Eukarya* in yellow. (A) Amino acids are ordered according to the total frequency (see [Figure 18](#)). (B) The lines represents the moving average ($n=3$) of length distributions.

RESULTS

By monotopic/polytopic proteins

We also examined the differences in the amino acid distribution in monotopic (single-pass) and polytopic (multi-pass, with two or more TM segments) integral membrane proteins.

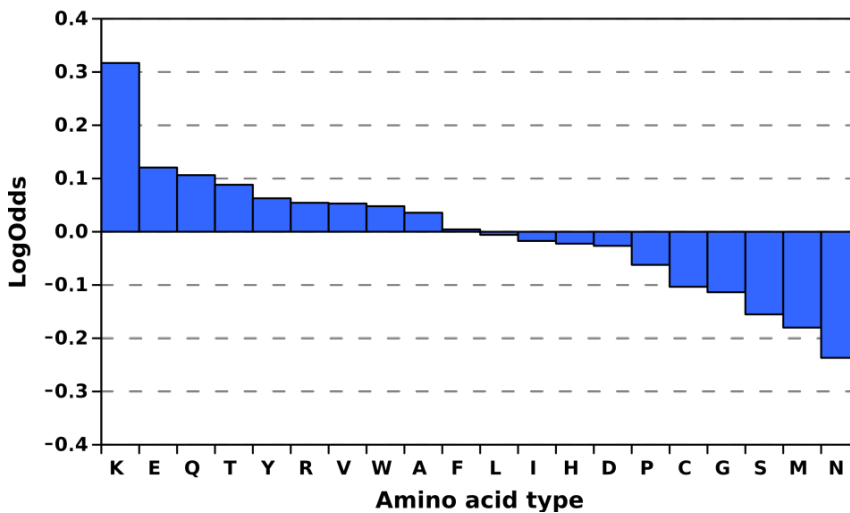


Figure 21. LogOdds for comparison of amino acid composition between TM helices from monotopic and from polytopic proteins. Positive LogOdd values are indicative of prevalence in monotopic proteins.

Only 51 of the 792 TM segments from our database are from monotopic proteins, which comprise 1,322 residues (6.8 % of the total of 19,356). To be able to compare the amino acid distribution among both groups we calculated and plotted the LogOdds ([Figure 21](#)). Gly and Pro are under-represented in monotopic proteins. Both residues are considered as helix breakers, therefore the observed under-representation in monotopic proteins could be explained because TM interactions in polytopic proteins can compensate the helix destabilization caused by Gly or Pro. Similarly, one can expect

charged and polar residues to be under-represented in monotopic proteins, but only Asn and Ser clearly are. Asp, His and Arg are quite similar in both groups, whereas surprisingly Gln, Glu and Lys (specially this latter one), are over-represented in monotopic proteins. Nevertheless, the correlation coefficient between amino acid distribution of TM helices from monotopic and from polytopic proteins is high (0.98).

r.1.4. Position-dependent distribution of amino acid residues in TM helices

We then analyzed the position-dependent distribution of amino acids in the 792 TM helices from our complete database.

A comparison of the amino acid frequency at different positions in a TM segment, taking as reference the TM center, confirmed that about half of the natural amino acid residues have similar distributions at positive positions (towards inside the cell) than at negative positions (towards outside the cell) ([Figure 22](#)). It was found that not only the strongly hydrophobic residues but also Gly and the hydroxylated residues Ser and Thr are equally distributed along the hydrophobic core of the membrane.

Sulphur containing Met or Cys are also frequent at different locations within the hydrophobic core, but a relative prevalence can be observed in a region that would correspond with the initial portion of the polar headgroups of the phospholipids, consistent with the slightly amphipathic nature of these amino acids and in agreement with its distribution in the lipid bilayer recently obtained from molecular dynamics simulation (MacCallum *et al.*, 2008).

While Phe has a flat distribution in TM helices, behaving as a hydrophobic residue, Trp, Tyr and Pro residues are distributed in a

RESULTS

biased manner: they are found preferentially at the ends of the bilayer (i.e. at the interface between the hydrophobic core of the bilayer and the bulk water). At this location, aromatic residues may serve as anchors for the TM helices into the membrane. In fact, Trp and Tyr positioned 7 to 9 residues away from the center of a TM segments result in a reduction in free energy (Hessa *et al.*, 2007), which nicely correlates with the present statistical distribution from

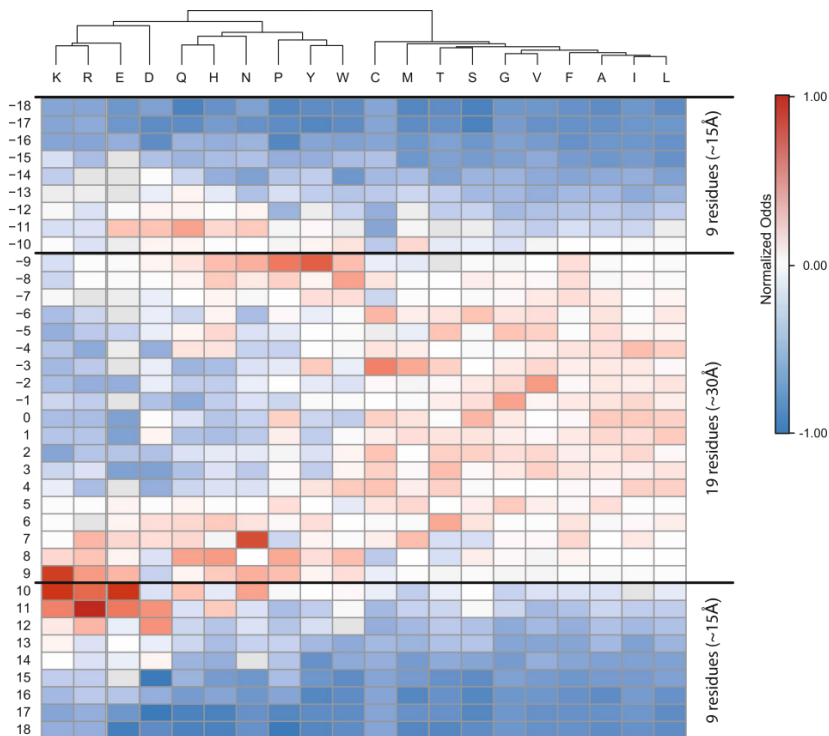


Figure 22. Amino acid type and position distribution in TM helices. Each amino acid type and its positioning in the TM helix is represented by its position-normalized Odds (that is, for each column the Odds are normalized to an average of zero and a standard deviation of unity). The amino acids are clustered (dendrogram at the top) on the basis of their positional normalized Odds within the helices. Positively labeled positions indicate the cytoplasmic side of the membrane and its flanking region whereas negatively labeled positions are indicative of extra-cytoplasmic regions.

three-dimensional structures ([Figure 22](#)). A similar distribution is observed for Pro residues, although an increased presence is detectable towards the center of the bilayer. Indeed, thirteen TM helices with known structure from our database have a Pro residue at the 0 (central) position, which in all cases results in a kink in the helix. Nevertheless, it should be noted that the interfacial preference of these three residues is somehow more pronounced at the non-cytoplasmic interface. This was also observed in the case of aromatic residues (Trp and Tyr) in a membrane protein prediction analysis using sequence information from 107 genomes (Nilsson *et al.*, 2005).

The distribution pattern for Asn, His and Gln, corresponds to an interfacial preference close to the end of the TM regions, which is consistent with the amphipathic nature of these molecules. In good agreement with our data, this pattern was previously reported for His residues (Ulmschneider and Sansom, 2001).

Among the 792 TM helices included in our database, 586 helices (74.0%) contained at least one ionizable residue (Asp, Glu, His, Lys or Arg) among the entire sequence and 461 (58.2%) helices contained at least one ionizable residue within the hydrophobic region (that is, the central 19 amino acid residues). A summary of the statistics is presented in [Figure 23](#). Since the energetic cost of inserting an ionizable group in the hydrophobic environment of the membrane is very high (White and Wimley, 1999), charged amino acids should generally be excluded from the hydrophobic core of the TM helices. However, charged amino acids consistently clustered at the TM flanking regions ([Figure 22](#)). For example, acidic (Asp and Glu) residues result in an increased distribution at both cytoplasmic and extra-cytoplasmic side of the membrane, although with some prevalence for the cytoplasmic region. Positively charged (Arg and Lys) residues distribution is even more asymmetric between opposite

RESULTS

sides of the membrane, in agreement with the positive-inside rule (von Heijne, 1992).

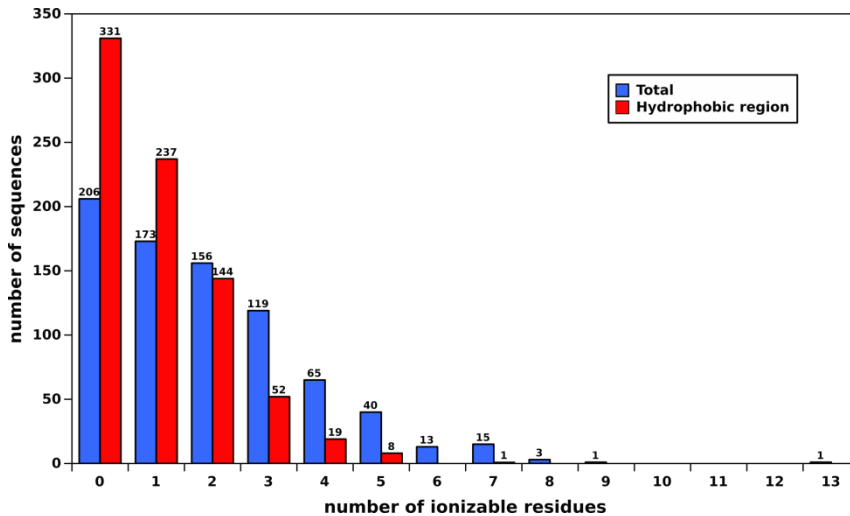


Figure 23. Ionizable residues in TM helices. Blue bars represents the number of helices with each number of ionizable (D, E, H, K or R) residues among the entire sequence, whereas red bars represents the number of helices with ionizable residues in the hydrophobic region (the central 19 amino acid residues).

When considered globally, charged residues (Lys, Arg, Glu and Asp) cluster preferentially near the cytoplasmic end of the TM segments (Figure 24, red line). This effect was already noted in a previous structure-based analysis that included a lower number of structures available at the time (Ulmschneider *et al.*, 2005). On the contrary, although polar residues (Gln, His, and Asn) mimic the distribution pattern of charged residues avoiding the more hydrophobic region of the bilayer, they show a preference for the extra-cytoplasmic region (Figure 24, orange line). Aromatic residues (Trp, Tyr and Pro) are more abundant around 8 to 9 residue positions away from the center of the membrane, that is, within the interface

RESULTS

region, but with some bias toward the extra-cytoplasmic interface (Figure 24, green line). The rest of natural amino acids (Cys, Met, Thr, Ser, Gly, Val, Phe, Ala, Ile and Leu) are more abundant at the center of the bilayer, within 7 amino acid positions on both sides of the membrane normal, but they are also very frequently found beyond this boundary as noted by their overall proximity to the Odds value of 1 for positions >10 on both sides of the center of the membrane (Figure 24, blue line). Interestingly, the amino acid distribution patterns at both interface regions are slightly different. There is a sharper transition from mainly hydrophobic to charged, polar and aromatic residues at the cytoplasmic side of the membrane (positions 6 to 8) compared to that at the extra-cytoplasmic side (positions -5 to -9).

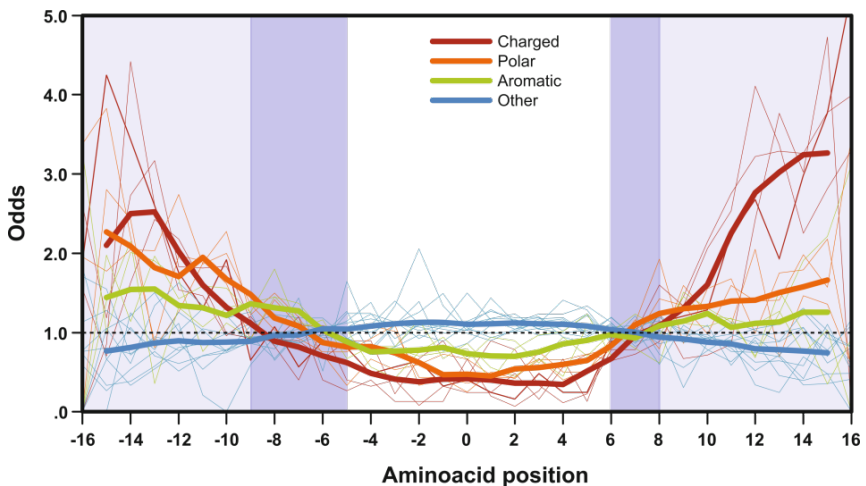


Figure 24. Most likely positions of amino acid groups in a membrane. Thin lines represent the positional Odds for each amino acid individually, whereas thick lines represent the average positional Odds for each group of amino acids obtained from Figure 6. Amino acid types are grouped as in the dendrogram in Figure 22, i.e. charged amino acids (red KRED), polar amino acids (orange QHN), aromatic amino acids plus Pro (green PYW), and the other amino acids (blue CMTSGVFAIL).

RESULTS

Finally, we analyzed and plotted the odd ratio for each amino acid in three regions in a membrane, that is, taking the hydrophobic TM region as the central 19 positions ($\sim 30\text{\AA}$) and 9 residue positions ($\sim 15\text{\AA}$) on both sides as the extra-cytoplasmic (from -10 to -18 residues) and cytoplasmic (from 10 to 18) flanking regions ([Figure 25](#)). Hydrophobic amino acids (blue colored) populated preferentially the hydrophobic center. However, this trend is not observed for the more prevalent residues in TM segments (for example Leu, [Figure 18](#)), which are also frequently found at the flanking regions. Trp, Tyr, and Pro (green) have a minor increase for the extra-cytoplasmic flanking region. The absence of higher differences for the distribution of these residues is probably due to their precise location at the interface between the hydrophobic core and the flanking hydrophilic environment. Polar (orange) residues (Gln, His, and Asn) have a preference for both flanking regions since they are energetically unfavorable within the membrane core. These residues do not ionize at the physiological pH and are able to donate and accept hydrogen bonds simultaneously. Such an effect translates into a higher preference of Gln, His and Asn for the rich hydrogen bond network environment of the interface. Charged residues (red) are underrepresented at the hydrophobic core and resulted in preferences for the cytoplasmic flanking region being acidic residues more prevalent at the extra-cytoplasmic flanking region. Furthermore, basic residues are strong topological determinants that heavily populate the cytoplasmic flanking region.

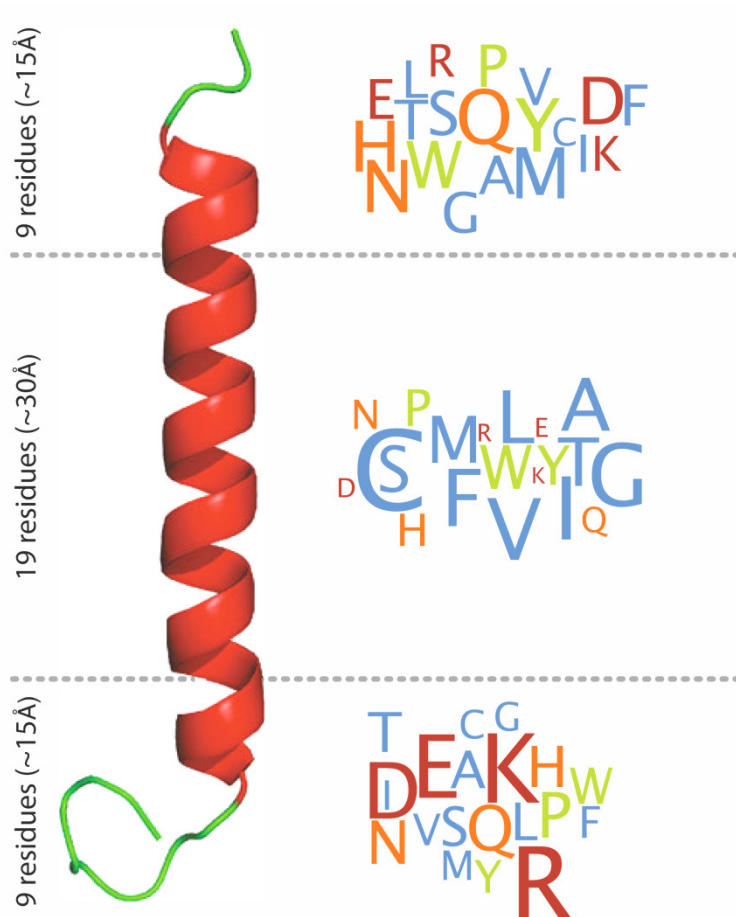


Figure 25. Amino acid location prevalence in a membrane. Letter size is proportional to the odds (relative prevalence) of finding a given amino acid in the three regions in a membrane (i.e., from top to bottom outer, membrane, and inner regions). Amino acids are colored as in [Figure 24](#).

RESULTS

R.2. BIOLOGICAL INSERTION OF COMPUTATIONALLY DESIGNED SHORT TRANSMEMBRANE SEGMENTS

r.2.1. Predicted insertion capacity for designed sequences

Using the data obtained in the previous chapter, we generated sets of sequences with different lengths and amino acid composition (See [Methodology](#) and [Figure 15](#) for details of the design procedure).

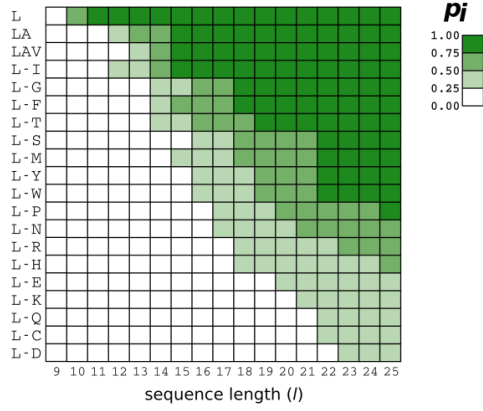
The membrane insertion efficiency of the computationally designed sequences was predicted using the experimentally based *ΔG Prediction Server* (<http://dgpred.cbr.su.se/>). In an initial screen, the insertion efficiency of poly-leucine segments of different lengths (*l*, from 9 to 25 residues) was calculated. Stretches of 11 or more leucine residues were predicted to be fully inserted, while 9 consecutive leucines was not enough to be inserted and 10 leucines resulted in a probability of insertion (p_i) of 0.59 ([Figure 26 A](#), first row).

As expected, these results were in excellent agreement with the previous experimental data (Jaud *et al.*, 2009) that was used to construct the *ΔG Predictor*. These extremely short sequences would likely provoke the adaptation of the surrounding bilayer to reduce the putative hydrophobic mismatch by changes in the lipid order parameters in the peptide neighborhood, according to molecular dynamics simulations (Jaud *et al.*, 2009).

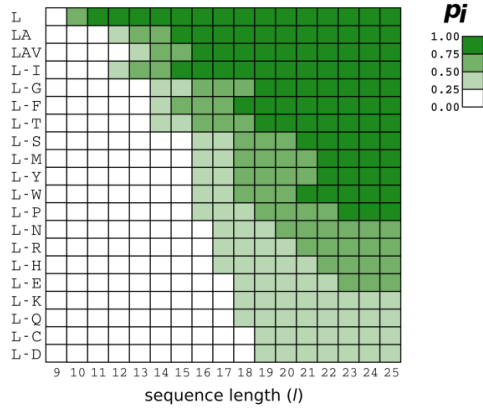
We thus began by testing computationally designed segments with more variable amino acid composition, initially formed by leucine and alanine residues, which are the two most prevalent residues in TM helices. Sets of 1,000 sequences from 9 to 25 residues long were computed with the leucine-alanine relative ratio (58.2% Leu-41.8% Ala) found in TM helices in membrane proteins of known structures. At least 13 residues were needed to obtain a

RESULTS

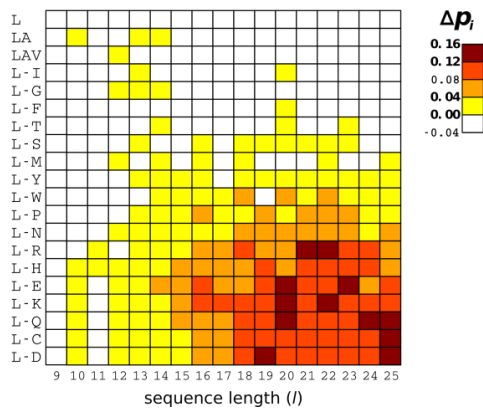
A



B



C



RESULTS

Figure 26. Predicted membrane insertion efficiencies for computationally designed sequences. (A) Probability of insertion (p_i) for a series of computationally designed TM segments of different lengths. First row corresponds to the predictions of p_i values for polyleucine stretches of different lengths (l). Each row in descending order represents the inclusion of a specific amino acid to the TM segment composition used in the previous row. The row order is derived from the prevalence for each amino acid type in TM helix composition in a previous structure-based statistical analysis. (B) Similar to (A) but including in the computational design information of the position-dependent distribution for each amino acid type in TM helices. (C) Differences between the position-defined (B) p_i values and those obtained for the computed sequences using only amino acid composition constraints (A).

significant level of predicted insertion ([Figure 26 A](#), second row), which was in good agreement with what it was found for alternating leucine and alanine peptides in lipid vesicles (Krishnakumar and London, 2007). Subsequently, we included one by one the rest of the 20 natural amino acids following the order of prevalence found in native membrane proteins and maintaining the ratios between them as obtained in the previous chapter. For instance, sets of 1,000 sequences formed by leucines, alanines and valines were generated using 42.6% Leu, 30.6% Ala and 26.8% Val residues, for each sequence length ([Table 2](#)). As expected, sequence sets including less hydrophobic residues needed longer segments to achieve insertion p_i values greater than 0.5 ([Figure 26 A](#)). Interestingly, when the less abundant lysine, glutamine, cysteine and aspartate residues were included at their naturally observed frequencies, the algorithm predicted that these computationally designed sequences were not expected to be efficiently inserted into biological membranes through the ER translocon. However, some natural sequences with these precise compositions are successfully inserted *in vivo*. Therefore, in line with previous findings (Hessa *et al.*, 2007), we hypothesized that not only the amino acid composition is relevant for TM insertion, but that the actual position of the amino acid residues with respect the center of

%	L	A	V	I	G	F	T	S	M	Y	W	P	N	R	H	E	K	Q	C	D
L	100.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LA	58.2	41.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LAV	42.6	30.6	26.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LVI	33.6	24.2	21.2	21.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LVI	28.5	20.5	18.0	17.8	15.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LVI	24.9	17.9	15.7	15.6	13.4	12.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LVI	23.1	16.6	14.5	14.4	12.4	11.5	7.5	-	-	-	-	-	-	-	-	-	-	-	-	-
LVI	21.5	15.5	13.6	13.4	11.5	10.7	7.0	6.8	-	-	-	-	-	-	-	-	-	-	-	-
LVI	20.5	14.7	12.9	12.8	11.0	10.2	6.6	6.5	4.8	-	-	-	-	-	-	-	-	-	-	-
LVI	19.6	14.1	12.4	12.3	10.5	9.8	6.4	6.2	4.6	4.1	-	-	-	-	-	-	-	-	-	-
LVI	19.0	13.6	12.0	11.8	10.2	9.5	6.1	6.0	4.5	4.0	3.4	-	-	-	-	-	-	-	-	-
LVI	18.4	13.2	11.6	11.5	9.8	9.2	5.9	5.8	4.3	3.9	3.3	3.1	-	-	-	-	-	-	-	-
LVI	18.0	12.9	11.3	11.2	9.6	9.0	5.8	5.7	4.3	3.8	3.2	3.0	2.1	-	-	-	-	-	-	-
LVI	17.6	12.7	11.1	11.0	9.4	8.8	5.7	5.6	4.2	3.7	3.1	2.9	2.1	2.1	-	-	-	-	-	-
LVI	17.3	12.4	10.9	10.8	9.3	8.6	5.6	5.5	4.1	3.6	3.1	2.9	2.0	2.0	1.9	-	-	-	-	-
LVI	17.0	12.2	10.7	10.6	9.1	8.5	5.5	5.4	4.0	3.6	3.0	2.8	2.0	2.0	1.8	1.7	-	-	-	-
LVI	16.7	12.0	10.5	10.4	9.0	8.3	5.4	5.3	4.0	3.5	3.0	2.8	2.0	2.0	1.8	1.6	1.6	-	-	-
LVI	16.5	11.8	10.4	10.3	8.8	8.2	5.3	5.2	3.9	3.5	2.9	2.7	1.9	1.9	1.8	1.6	1.6	1.5	-	-
LVI	16.3	11.7	10.2	10.2	8.7	8.1	5.3	5.1	3.8	3.4	2.9	2.7	1.9	1.9	1.8	1.6	1.6	1.5	1.2	-
LVI	16.1	11.6	10.1	10.0	8.6	8.0	5.2	5.1	3.8	3.4	2.9	2.7	1.9	1.9	1.7	1.6	1.6	1.5	1.2	1.1

Table 2. Amino acid relative ratio values used in the computational calculations. We included one by one the 20 natural amino acids following the order of prevalence found in native proteins and described in the previous chapter.

RESULTS

the TM segment could also have an important influence on the insertion process.

To address this question, we changed our computational design algorithm to reflect the amino acid frequency values stratified at different position within a TM segment, taking as a reference the TM center. Sets of 1,000 sequences were designed, using this position-dependent distribution of residues in natural TM helices, and their p_i values were predicted ([Figure 26 B](#)). The predicted insertion efficiency for sets with higher proportion of hydrophilic residues increased significantly when residue position constrains were included in the computational design. Hence, series of position-defined sequences including Arg, His, Glu, Lys, Gln, Cys and Asp were predicted to insert more efficiently at any given segment length than the previous sets based only on the global TM helix residue composition. The differences between the two data sets were most evident for TM sequences that include the less prevalent, more hydrophilic residues ([Figure 26 C](#)).

r.2.2. Determination of experimental insertion by glycosylation assay

To be able to directly compare the predicted insertion efficiencies to insertion into the mammalian ER, a total of 39 computationally designed sequences with different lengths and amino acid composition (31 from the non-position defined dataset and 8 from the position-defined dataset) were analysed using a well-established *in vitro* assay for quantifying the efficiency of membrane integration of designed TM sequences into dog pancreas rough microsomes (Sääf *et al.*, 1998). The experimentally tested sequences are listed in [Table 3](#).

ID	Sequence
LA11	LAALLLAAALL
LA13	LLALALALLAALL
LA15	LALLLALAALLLALAL
LA17	AALALLLALLLALALALL
LA19	LALALALAALALLLALLLA
LA21	LALLLAALLLALALLLAAAAAL
LA23	ALLAALLLALLLALAALALAALAL
L-F13	FFPIFGGVILIAA
L-F15	AATALLGLAVGLFLF
L-F17	AAAFATAGLFIAIVLIF
L-F19	AIIILFLLVLIGVGVLAAGV
L-F21	VGIFVGLIFLAVLGIALLLGG
L-F23	IILVIAGAFVVGVAFFVVLVLLF
L-W13	MLTVATIFISLGF
L-W15	GFVAYVAMIAWLLIG
L-W17	VSPVGITVAFVWLVPMT
L-W19	AMISVLVGVWVWVLLFFAGT
L-W21	MTLFMIIMLLAMYAWAGGLVG
L-W23	IVLLAGLALYSGIVAVVTIFMWM
L-R13	LMRVFAVVLGNVI
L-R15	LPLLWASVITAGTVL
L-R17	RWGTTAYFMLAVIAAPF
L-R19	PLLRPFPTLVTVLAYMAIV
L-R21	LFMAPVWVYLVGNLITALLIYL
L-R23	LMFRVATSAPLIILYGFMRLLTTF
L-K13	IFFEWLVLGMGI
L-K15	VIEIVLIYHLVPIWI
L-K17	LFFFAVSLKLAWTMFVP
L-K19	PFIWAASLSIGAKLSYAW
L-K21	MIFLLLLPMAILHRSPLLVGK
L-K23	FFHVLEFIWLSVVMPLPFAIEAYN
L-R17 position-defined	LIVLFVRMLLAVLGLNG
L-R19 position-defined	LAVITLAIAWFMSIFGAYP
L-R21 position-defined	RIWSILYIALTWTFFILAGASR
L-R23 position-defined	AAIYISINLFAFVTMVLFARLPA
L-K17 position-defined	SVALMFSLVGMYPGLH
L-K19 position-defined	WVEPTVYLIITFLALLRVK
L-K21 position-defined	YWIEVPSVVIITVAAPGVLSP
L-K23 position-defined	TWKIISGLVFLALFIWGMYPSEA

Table 3. List of experimentally tested sequences. All sequences were flanked by insulating glycine-proline regions (GGPG- X_n -GPGG), where X represents the designed sequences of n (11-23) residues length.

RESULTS

The host protein (Lep) consists of two TM segments (H1 and H2) connected by a cytoplasmic loop (P1) and a large C-terminal domain (P2), and inserts into ER-derived RMs with both termini located in the lumen ([Figure 16](#)). The designed sequence (“TM-tested”) was engineered into the luminal P2 domain and flanked by two acceptor sites (G1 and G2) for *N*-linked glycosylation. The engineered glycosylation sites can be used as membrane insertion reporters because G1 will always be glycosylated due to its native luminal localization, but G2 will be glycosylated only upon translocation of the analyzed region through the microsomal membrane. A singly glycosylated construct in which TM-tested is inserted into the membrane has a molecular mass ~2.5 kDa higher than the molecular mass of Lep expressed in the absence of microsomes; the molecular mass shifts by ~5 kDa upon double glycosylation (i.e. membrane translocation of the TM-tested).

We measured membrane insertion efficiencies of systematically designed sequences GGPG-X₁₁₋₂₃-GPGG, in which the flanking tetrapeptides are included to insulate the central computed stretches of different lengths (11 to 23 residues long) from the surrounding sequence in the Lep model protein. The insertion efficiency was calculated on the basis of the fractions of singly (f_{1g}) and doubly (f_{2g}) glycosylated forms by using $p_i = f_{1g}/(f_{1g} + f_{2g})$ determined from quantitative analyses of SDS-PAGE gels.

[Figure 27](#) includes representative examples of SDS-PAGE gels showing the translation products of non-position defined sequences composed of Leu and Ala residues (LA, *top*), of Leu, Ala, Val, Ile Gly and Phe (LAVIGF, *middle*) and of Leu, Ala, Val, Ile Gly, Phe, Thr, Met, Tyr and Trp (LAVIGFTSMYW, *bottom*). In all three cases, the insertion efficiency into the ER membrane of the translated products increased with the sequence length. Furthermore, the length needed to achieve an efficient insertion (>0.75) increases when we included a

RESULTS

wider variety of amino acids (i.e., 15 residues long for LA subset, 17 for LAVIGF and 19 for LAVIGTSMYW).

When less frequent polar amino acids as Pro, Asn and Arg (LAVIGFTSMYWPNR subset) or Pro, Asn, Arg, His, Glue and Lys (LAVIGFTSMYWPNRHEK subset) are added to the composition of sequences, complete insertion is not achieved even for long stretches of 23 residues ([Figure 28 A](#)). However, for the same length, insertion increased when the putative TM sequences were designed using position-defined constraints ([Figure 28 B](#)).

The results of the quantitative analysis for all sequences experimentally tested are shown in [Table 4](#), as well as their predicted values obtained with *ΔG Predictor*.

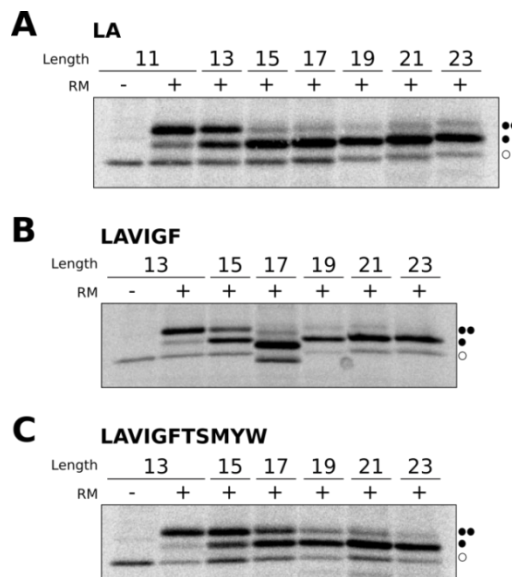


Figure 27. *In vitro* translation in the presence (+) or absence (-) of rough microsomes (RM) of computationally designed TM sequences of different length of subsets LA (A), LAVIGF (B) and LAVIGTSMYW (C) without position-defined constraints. Non-glycosylated protein bands are indicated by an empty dot; singly and doubly glycosylated proteins are indicated by one or two black dots, respectively.

RESULTS

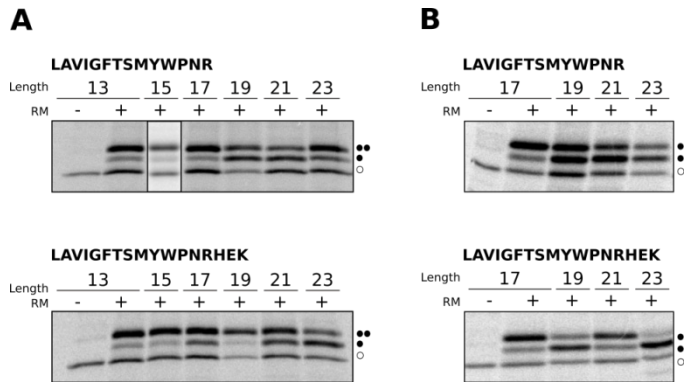


Figure 28. *In vitro* translation products of computationally designed TM sequences of different length of subsets LAVIGFTSMYWPNR and LAVIGFTSMYWPNRHEK designed without (A) and with position-defined constraints (B).

r.2.3. Correlation between predicted and experimentally determined insertion efficiencies

Next, we analyze the correlation between the predicted p_i values and the experimental p_i obtained with the glycosylation assay.

Figure 29 shows plots of the probability of membrane insertion as a function of TM length for series of sequences that included the more prevalent amino acid residues in TM helices. Hydrophobic Leu and Ala account for more than one quarter (27.7%) of all amino acids in TM helices, and together with Val, Ile, Gly and Phe constitute the bulk of the amino acids embedded into the hydrophobic core of the membrane, accounting for almost two thirds (64.4%) of all amino acids in this membrane region. The grey area corresponds to the 500 sequences between percentiles 0.25 and 0.75 of the 1,000 predicted p_i values (see Annex II). The figure also shows the probability of

RESULTS

ID	ΔG_{app}^{pred}	ΔG_{app}^{exp}	p_i^{pred}	p_i^{exp}
LA11	1,31	0,92	0,10	0,17
LA13	-0,02	0,05	0,51	0,48
LA15	-1,30	-1,15	0,90	0,88
LA17	-2,43	-1,36	0,98	0,91
LA19	-3,52	-4,23	1,00	1,00
LA21	-4,55	-1,38	1,00	0,91
LA23	-5,52	-2,69	1,00	0,99
L-F13	0,93	1,27	0,17	0,10
L-F15	0,04	-0,48	0,49	0,69
L-F17	-1,05	-1,43	0,86	0,92
L-F19	-1,99	-1,37	0,97	0,91
L-F21	-2,90	-1,55	0,99	0,93
L-F23	-3,63	-5,45	1,00	1,00
L-W13	2,07	1,29	0,03	0,10
L-W15	1,08	0,43	0,14	0,32
L-W17	0,37	-0,37	0,35	0,65
L-W19	-0,37	-0,87	0,65	0,81
L-W21	-1,38	-1,31	0,91	0,90
L-W23	-1,70	-1,46	0,95	0,92
L-R13	2,93	1,01	0,01	0,15
L-R15	2,12	1,11	0,03	0,13
L-R17	1,50	1,11	0,07	0,13
L-R19	0,45	-0,02	0,32	0,51
L-R21	0,31	-0,22	0,37	0,59
L-R23	-0,14	0,62	0,56	0,26
L-K13	3,64	1,09	0,00	0,14
L-K15	3,12	1,14	0,01	0,13
L-K17	1,96	0,66	0,04	0,25
L-K19	2,04	1,05	0,03	0,14
L-K21	0,97	0,27	0,16	0,39
L-K23	0,95	-0,36	0,17	0,65
L-R17 position-defined	1,05	1,29	0,15	0,10
L-R19 position-defined	0,31	0,62	0,37	0,26
L-R21 position-defined	-0,27	0,29	0,61	0,38
L-R23 position-defined	-0,71	0,29	0,77	0,38
L-K17 position-defined	1,27	1,20	0,10	0,12
L-K19 position-defined	0,11	0,25	0,45	0,40
L-K21 position-defined	0,67	0,99	0,25	0,16
L-K23 position-defined	-0,87	-0,02	0,81	0,51

Table 4. Predicted and experimentally measured apparent ΔG (kcal/mol) and p_i values for all sequences tested with the glycosylation assay. Experimental values are the mean of at least three independent experiments.

RESULTS

insertion for the experimentally measured sequences (orange line) as well as their particular predicted values (blue line).

As expected, the inclusion of less prevalent amino acid residues in the designed TM sequences increased the variability of the predictions (compare the grey areas in [Figure 29](#) and [Figure 30 A](#)) for all sets of sequence lengths. Moreover, the differences between the predictions and the experimental measurements were larger for some sequences ([Figure 30 A](#)). Next, we introduced constrains in our computational designs derived from the position-defined distributions found in native TM helices.

The introduction of the amino acid position constrains in our computational algorithm increased both the predicted as well as the experimentally measured p_i values for the TM sequences containing less abundant (polar) residues ([Figure 30 B](#)). Nevertheless, we noticed that in some cases the differences between the predicted and experimental values were large. In these cases, we re-ran the *ΔG Predictor* algorithm but this time the algorithm was allowed to identify subsequences (i.e., with lower ΔG estimated values). The new predictions, in all cases, approached the experimental values ([Figure 30 A and B](#), dashed lines), reinforcing that biological membranes can adapt to accommodate sequences harboring deviations from canonical hydrophobic regions.

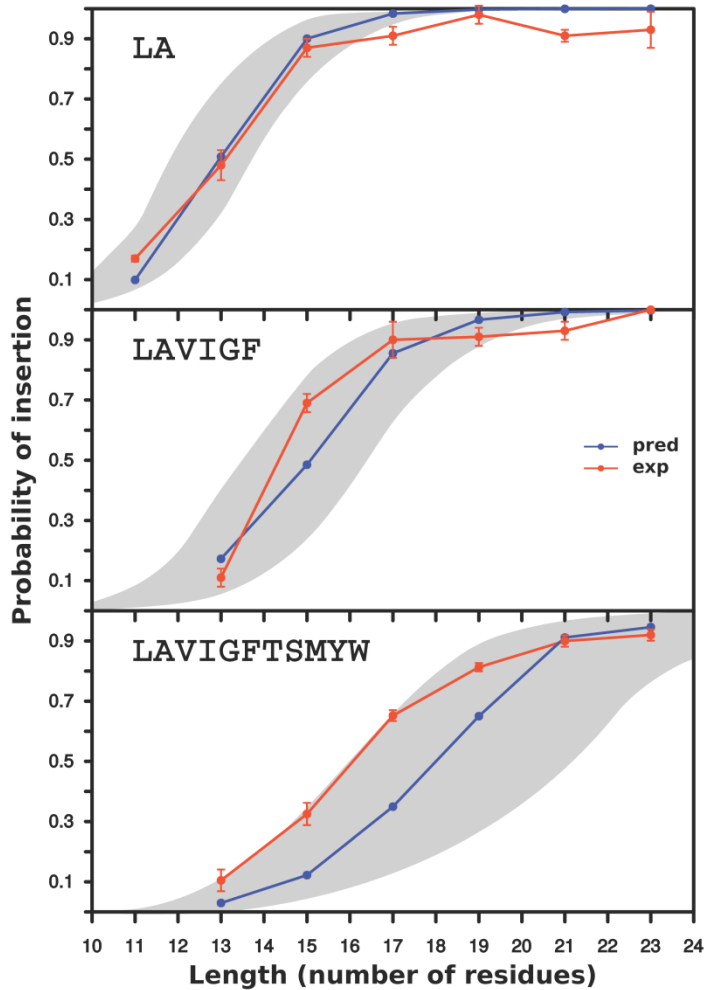


Figure 29. Predicted and experimental p_i values of sequences including the most prevalent amino acid residues in TM helices. Upper panel: Computationally designed Leu and Ala sequences of different lengths. The predicted values for each given sequence are shown in blue and the measured values obtained for at least three independent experiments in orange. The gray area represents the predictions of the p_i values for the 500 sequences between percentiles 0.25 and 0.75 of the total population (1,000 computed sequences) (see [Annex II](#)). Central panel: Similar to upper but the computationally designed TM sequences contained Leu, Ala, Val, Ile Gly and Phe. Bottom panel: Similar to upper but including, in addition to the hydrophobic, the more prevalent polar and aromatic residues in TM segments Met, Tyr and Trp.

RESULTS

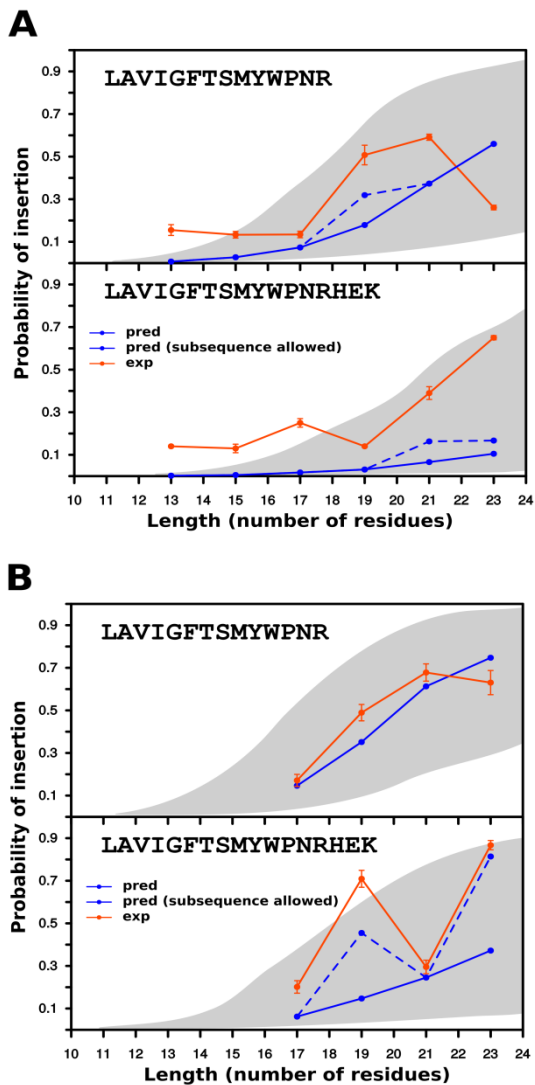


Figure 30. Predicted and experimental p_i values of computationally designed TM sequences including polar and charged amino acid residues. Experimental and predicted values are shown as in [Figure 29](#). (A) Computationally designed TM sequences were constrained only by amino acid composition constraints. (B) Computationally designed TM sequences were constrained both by amino acid composition and distribution along the helix.

r.2.4. Analysis and mutants of sequences with low correlation between predicted and experimental insertion.

Considering the complexity of the biological system, the two sets of p_i values are well correlated ([Figure 31](#)): the linear fit has a slope of 0.80 with an r value of 0.93. Interestingly, the only outliers (highlighted as empty dots) are the longer sequences designed without position-dependent constraints, which include polar/charged residues.

A closer look to the first of these sequences (L-K23, empty orange dot) revealed the presence of a histidine and glutamate residues ([Table 5](#)). This construct inserted experimentally more efficiently than predicted by the ΔG algorithm. Given the 3.6-residue periodicity of an ideal (canonical) α -helix, an intrahelical charge pair would be expected for this ($i, i+3$) His-Glu pair (see [Figure 32](#)).

To test this hypothesis, we swapped the histidine residue with its neighboring valine residue (L-K23 H3V/V4H, [Table 1](#)) generating an ($i, i+2$) periodicity for the His-Glu pair and likely precluding intrahelical pairing by orienting the two side-chains toward opposite faces of the helix. Noticeably, this mutant resulted in a slightly increased predicted p_i value but consistently diminished experimental insertion efficiency. The combined effect of these data improved the correlation between the experimental and prediction values for the mutant sequence ([Figure 31](#), arrow pointed dot), which has exactly the same amino acid composition as the L-K23 sequence.

RESULTS

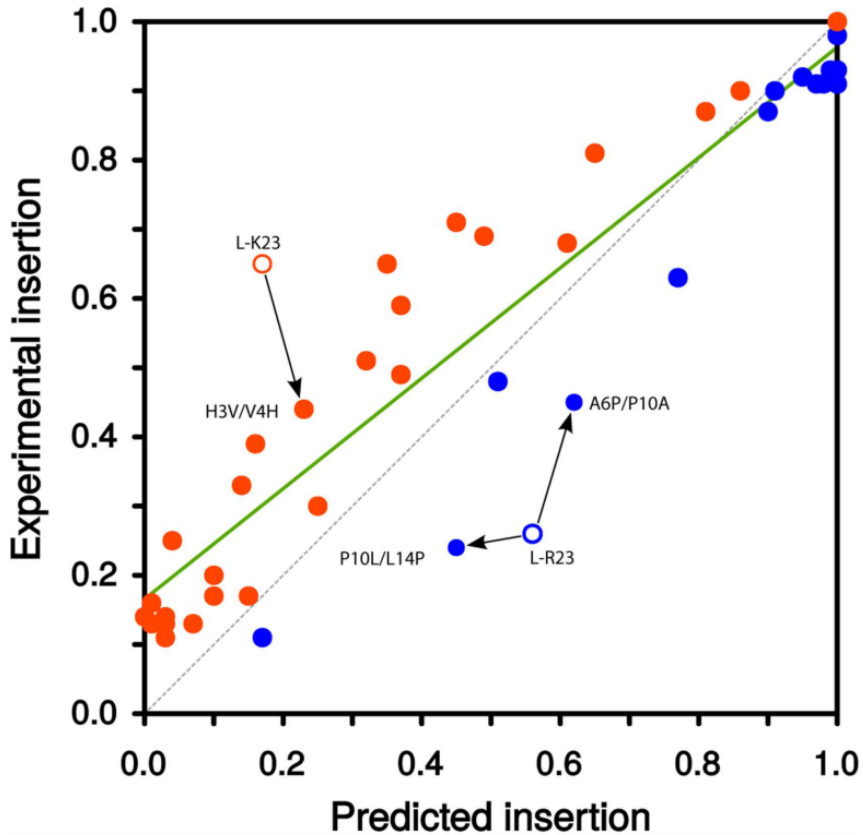


Figure 31. Correlation between experimentally measured and predicted p_i values. For each sequence analyzed, predicted values higher than the experimental ones (i.e., below the grey dashed line) are shown in blue, whereas experimental values higher than the predicted ones (i.e., above the grey dashed line) are shown in orange. The correlation between the experimental and predicted insertion probabilities is indicated by a green line. Outliers are shown as empty circles and the results of their mutated sequences (that is, P10L/L14P, A10P/P10A and H3V/V4H) are indicated by arrows.

ID	Sequence	ΔG_{app}^{pred}	ΔG_{app}^{exp}	P_i^{pred}	P_i^{exp}
L-K23	-10 -8 -6 -4 -2 0 2 4 6 8 10 FF H V L E F IWLSVVMLPFAIEAYN	0,95	-0,36	0,17	0,65
L-K23 H3V/V4H	FF V H N E F IWLSVVMLPFAIEAYN	0,71	0,14	0,23	0,44
L-R23	LMFRVATSAPLIILYGFMRLLTT	-0,14	0,62	0,56	0,26
L-R23 P10L/L14P	LMFRVATS A LLI I PYGFMRLLTT	0,13	0,70	0,45	0,24
L-R23 A6P/P10A	LMFRV P T S A A LIILYGFMRLLTT	-0,30	0,11	0,62	0,45

Table 5. Thermodynamic cost of L-K23- and L-R23-derived TM segments integration. The predicted and experimental (ΔG_{app}) energetic cost in kcal/mol of the computationally designed TM segments. Negative values are indicative of TM disposition, while positive values indicate non-TM disposition. Charged residues studied are highlighted in color and mutated residues are shown in bold.

These results support the idea that intra-helical salt-bridge formation between residues located on the same face of a TM helix (i , $i+3$ or i , $i+4$) may reduce the free energy of membrane partitioning (Chin and von Heijne, 2000; Bañó-Polo *et al.*, 2012), whereas the presence of His and Glu on opposite faces of the helix (i , $i+2$) is unfavorable and lowers the ER translocon membrane insertion efficiency.

Moreover, we analyzed the presence of charge pairs in the 792 TM helices of our database. We found that 169 helices (21.3%) contain at least one charge pair and a total of 184 charge pairs, of which 103 (56%) were in i , $i+3$ positions and 81 (44%) were in i , $i+4$ positions. More detailed results are shown in [Table 6](#). Then we look at the positions in which these charge pairs are located ([Figure 33](#)), and we found that charge pairs are preferably located at the transition regions between the hydrophobic core and the interphase of the membrane, and are more prevalent in the cytosolic side.

Next, we analyzed the L-R23 ([Figure 31](#), empty blue dot) sequence. In this case the predicted value suggests a higher propensity to insert than the measured experimental value. Inspection of L-R23 sequence ([Table 5](#)) highlighted the presence of a proline residue in a

RESULTS

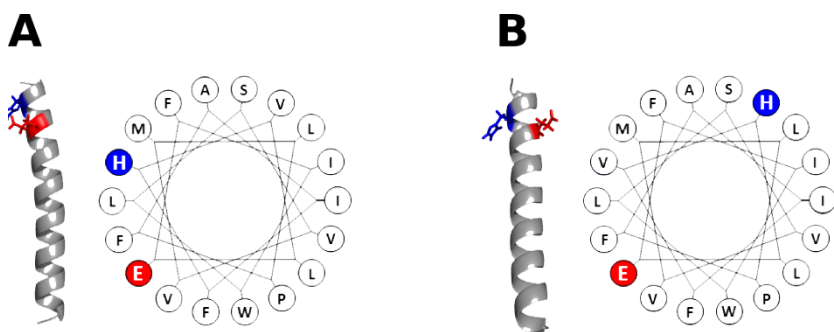


Figure 32. Canonic α -helix (left) and helical wheel (right) representations of L-K23 sequence (A) and L-K23-derived sequence (B). His residues are shown in blue and Glu residues are shown in red. In the canonic α -helix representation, both residues (His and Glu) side chains are shown in *sticks* mode.

central position within the hydrophobic region (-2 relative to the center, negatively labeled positions indicate extra-cytoplasmic face).

Based on the L-R23 construct, we made two different mutants with the proline residue placed at different positions. In particular, we compared the insertion efficiency of positioning the proline roughly one helical turn (4 residues) towards the N or C terminus by swapping mutagenesis. When the proline residue was moved towards the C-terminus (+2 position, mutant P10L/L14P) the experimental p_i value remained very similar to the one obtained for the original sequence (Table 5), whereas the construct in which the proline was moved four residues towards the N-terminus (-6 position, mutant A6P/P10A) was inserted more efficiently into the ER (Table 5).

The different effect observed for these two mutants can be explained by the different location of the proline residue in relation to the midpoint of the TM segment. Hence, in the case of P10L/L14P mutant (+2 position) the distortions produced by the proline occur around the center of the membrane plane where the system is probably

RESULTS

more sensitive to distortions. On the contrary, in the case of A6P/P10A mutant (-6 position) the presence of the proline closer to the interface would locate the unsatisfied carbonyl group in a less hydrophobic environment (White and Wimley, 1999; MacCallum *et al.*, 2008) probably reducing the free energy of membrane partitioning. It should be noted that all these mutants have exactly the same amino acid composition.

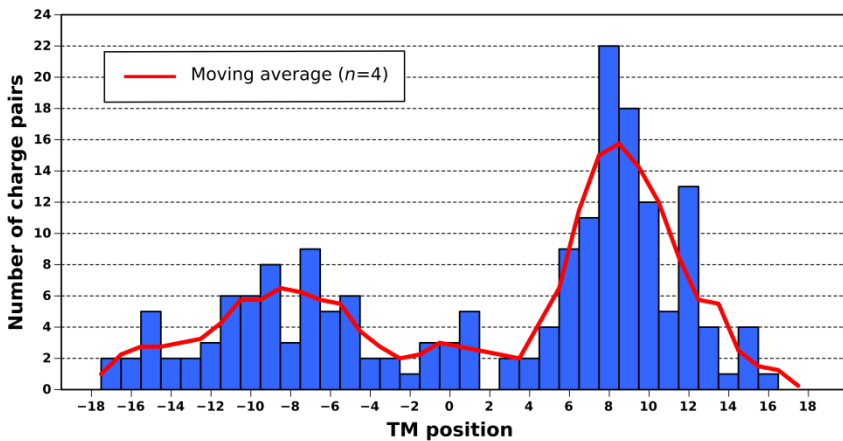


Figure 35. Position-dependent distribution of charge pairs. Blue bars represents the number of charge pairs in i , $i+3$ or i , $i+4$ positions centered at each TM position. Red line represents the moving average for a window of 4 positions. Positive positions indicate cytosolic side of the membrane; negative positions indicate extracytoplasmic/luminal side of the membrane.

	i , $i+3$		i , $i+4$		Total	
	D	E	D	E	D	E
H	10	15	9	7	19	22
K	13	17	11	19	24	36
R	21	27	19	16	40	43

Table 6. Charge pairs in TM helices. Number of each of the six different charge pairs (D-H, D-K, D-R, E-H, E-K and E-R) found in positions i , $i+3$ (left), i , $i+4$ (middle) and total (right).

DISCUSSION

D.1. STRUCTURE-BASED STATISTICAL ANALYSIS OF TRANSMEMBRANE HELICES

In this chapter, we revisited the differences between helices from water-soluble proteins and TM helices from integral membrane proteins in terms of length and amino acid composition. In addition, we analyzed the distribution of amino acid residues in TM segments, which have to energetically accommodate into the highly heterogeneous media of biological membranes by interacting favourably with its local environment ([Figure 3](#)). The present study involved 170 helical membrane proteins with known three-dimensional structure and topology, containing a total of 792 TM segments. These sequences were also compared with 7,348 helices from 930 water-soluble protein structures.

We first analyzed the length distribution for TM and water-soluble helices and we found that TM helices are longer than soluble helices. This is most likely due to that TM helices do suffer the restrictions imposed by the low dielectric constant at the hydrocarbon core of biological membranes, which forces the polypeptide backbone to adopt, on average, larger secondary structures than water soluble helices. On the other hand, water-soluble helices do not have to satisfy the demanding restrictions imposed by the complexity of the membrane environment, which allows them to adopt shorter structures. Moreover, as the translation per residue in a canonical helix is 1.5Å, a stretch of about 20 consecutive hydrophobic residues can span the 30 Å of the hydrocarbon core of biological membranes, which is in good agreement with the most prevalent length of TM helices (21 residues) found in our analysis. Longer helices can span the bilayer by different mechanisms to avoid the hydrophobic mismatch ([Figure 7](#)), such as tilting the helix axis respect to the

DISCUSSION

membrane plane, polypeptide backbone deformation or lipid accommodation (Holt and Killian, 2010).

The comparison of amino acid composition showed differences between the two types of helix ([Figure 18](#)), as noted previously by studies using smaller datasets (Bywater *et al.*, 2001). As expected, hydrophobic amino acids are more prevalent in TM helices, due to that their prevalence in the membrane environment depends primarily on their side chain hydrophobicity and on the local polypeptide region in which the amino acids reside spanning the membrane (Li and Deber, 1994). However, Ala is not over-represented in TM helices; probably because its greater tendency to participate in an helical structure in a aqueous environment (Blaber *et al.*, 1993) than in membrane-mimetic environments (Li and Deber, 1994). In fact, both biological (Nilsson *et al.*, 2003; Hessa *et al.*, 2005) and biophysical (Jayasinghe *et al.*, 2001) measurements have placed Ala at the threshold between those amino acids that promote membrane integration of TM helices and those residues that preclude membrane insertion. Additionally, although Gly and Pro residues are commonly considered as “helix breakers”, we found that both residues are also more frequent in TM helices. The explanation to this finding could be that Gly residues occur frequently in TM helix-helix interactions, especially in association with β -branched residues at neighbouring positions (Senes *et al.*, 2000), and that Pro, in addition to its function in signal transduction and gating across the membrane, may also be significantly involved in these processes (Orzáez *et al.*, 2004).

To clarify which amino acids have position constraints in TM helices, we analyzed the position-dependent distribution ([Figure 22](#)). Surprisingly, not only hydrophobic amino acids, but also Gly, Ser and Thr are equally distributed along the hydrophobic core of the membrane. It is important to note that Gly is a residue type that is

normally regarded as being conducive to turn (Williams *et al.*, 1987). Nevertheless, there are important folding reasons for incorporating Gly into TM helices: the absence of a side-chain of the Gly enables bulkier groups to be accommodated close to the polypeptide backbone of the TM helices and this might be important for intramolecular helix-helix packing, for homo-oligomerization, or for recognition of other membrane proteins, among other factors. Indeed, it has been observed that Gly has the highest overall packing value in membrane proteins (Eilers *et al.*, 2002). On the other hand, Ser or Thr residues within TM helices participate in hydrogen-bonding networks through by hydrogen bonding of the side chain oxygen atom to the acceptor side chain or peptide bond groups. These effects, intimate packing (Gly) and hydrogen bonding (Ser and Thr), can be relevant at any position along the TM region, which would explain the absence of position prevalence for these residues in TM helices.

In contrast, the biophysical reason for the observed distribution of Trp and Tyr residues, avoiding the center of the membrane ([Figure 22](#)), could rely on the relatively amphipathic nature of their side chains, which can form hydrogen bonds as well as exhibit hydrophobic character. Actually, this preferred location has previously been observed not only for α -helical but also β -barrel membrane proteins (Ulmschneider and Sansom, 2001). Pro has a similar distribution, with an increased prevalence in the center of the TM, which might be associated with the fundamental and subtle function of Pro in the dynamics, structure and function of many membrane proteins (Cordes *et al.*, 2002). Indeed, about 30% (4) of the 13 proteins with a Pro residue located at position 0 in a TM helix are rhodopsins. Rhodopsins are members of the G protein-coupled receptor family, in which a highly conserved Pro residue located in the center of helix 6 serves as a molecular hinge for the bending that

DISCUSSION

allows the creation of an opening in the proteins structure, essential to their function (Zhou *et al.*, 2012).

Charged residues (Asp, Glu, Arg and Lys) have an even more strongly position-restricted distribution (Figure 22), located preferentially in the interface regions of the membrane. This is likely because these amino acids should generally be excluded from the hydrophobic core of the TM helices since the energetic cost of inserting an ionizable group in the hydrophobic environment of the membrane is very high (White and Wimley, 1999). Indeed, within charged amino acids, basic residues (Arg and Lys) distribution is strongly biased towards the cytoplasmic side of the membrane, clearly reflecting the positive-inside rule (von Heijne, 1986a). Moreover, it has been demonstrated experimentally that basic residues act as stronger topological signals than acidic residues (Nilsson and von Heijne, 1990; Saurí *et al.*, 2009), which is reflected by their different statistical preferences on either end of the TM segments. The effect of positively charged amino acids located near the cytoplasmic end of hydrophobic segments has been in fact estimated to be approximately -0.5 kcal/mol to the apparent free energy of membrane insertion (Lerch-Bader *et al.*, 2008). This energetic contribution can be relevant for precise anchoring of hydrophobic regions to biological membranes.

Finally, the different transition from hydrophobic to charged, polar and aromatic amino acids observed between in both sides of the membrane (Figure 24) may reflect the important effect of the different lipid composition between the two lipid leaflets in biological membranes and the strong electrochemical potential over the prokaryotic inner cell membranes. For instance, asymmetry in the distribution of amino acid within TM segments from plasma membrane proteins has been reported (Sharpe *et al.*, 2010), and has

been attributed to asymmetry in the state of lipid order in the membrane. Such asymmetry is likely to be because of the enrichment of lipids, for example sterols and sphingolipids, in the extra-cytoplasmic leaflet, where a more gradual amino acid distribution can be expected.

The results on TM helix architecture obtained in this chapter should prove useful for constructing models of membrane proteins with desired properties, which could help filling in some of the many gaps in our knowledge in this field.

DISCUSSION

D.2. BIOLOGICAL INSERTION OF COMPUTATIONALLY DESIGNED SHORT TRANSMEMBRANE SEGMENTS

In this chapter, we systematically explored how the amino acid composition and positioning affect the efficiency of membrane insertion capacity of computationally designed TM segments, using both prediction and experimental measurement. We used a microsomal *in vitro* expression system to examine the translocon insertion efficiency of chosen examples from the designed sequences. To generate the sequences we took advantage of the calculated distributions of amino acids from our previous structure-based statistical analysis of TM helices (previous chapter).

Prior work showed that polyleucine sequences of 9-10 residues were efficiently inserted by the ER translocon into microsomal membranes (Kuroiwa *et al.*, 1991; Jaud *et al.*, 2009) and by the *E. coli* translocon (Chen and Kendall, 1995). However, TM segments in natural membrane proteins are not made exclusively of leucines. To expand our knowledge towards natural membrane proteins, we have designed and analyzed large sets of sequences with amino acid compositions that become more and more like the natural ones.

Using the ΔG Prediction Server we found that 12-14 consecutive hydrophobic residues is about the minimum required for insertion into biological membranes through the ER translocon for highly hydrophobic sequences composed by leucine, alanine, valine and isoleucine (L-I series), which account for almost half (47.8%) of amino acid residue composition in TM helices. Sequences containing less prevalent (more hydrophilic) amino acid residues in TM segments have to be longer to efficiently insert into the membrane.

Not surprisingly, there is a correlation between amino acid abundance in TM helices and hydrophobicity ([Figure 34](#)), which

explains the need for an increased sequence length compensating the lower hydrophobicity. However, this effect can be partly balanced by taking into account amino acid position-dependent contributions. Thus, when this last parameter was included in our computational designs, the predictions for the insertion efficiency of sequences harboring polar and charged residues significantly improved ([Figure 26 C](#)).

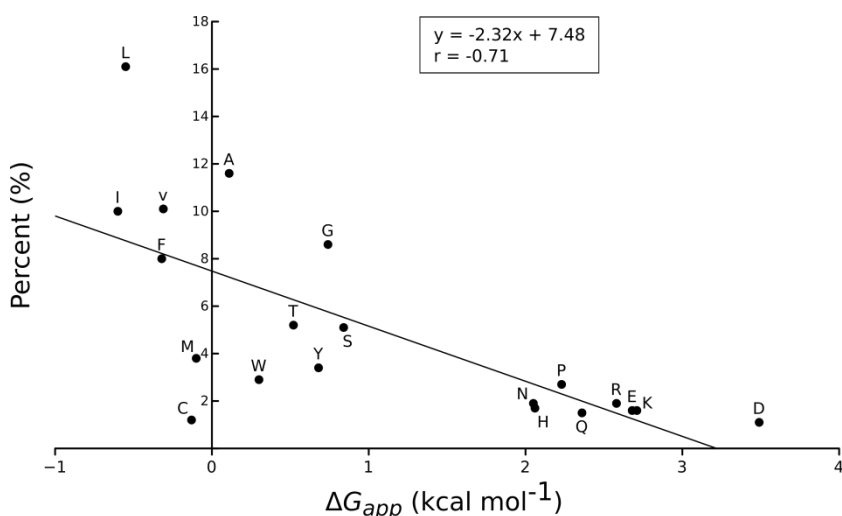


Figure 34. Correlation between the amino acid type prevalence found in TM helices from our data set and the biological free energy scale determined by (Hessa *et al.*, 2005).

Indeed, the results of our experimental assay using microsomal membranes ([Figure 29](#) and [Figure 30](#)) mirrored the ΔG algorithm predictions. Our data reinforces the accepted idea that there is an unfavorable free energy associated with locating hydrophilic residues in the hydrophobic core of the membrane. However, this effect can be reduced by allocating non-hydrophobic residues close to the polar

DISCUSSION

headgroup region of the lipids at the membrane interface (White and von Heijne, 2005), as well as by engaging polar residues in salt-bridge pair formation (Jayasinghe *et al.*, 2001; Baño-Polo *et al.*, 2013). This is in line with our analysis of specific sequences with experimental/predicted p_i values that deviate from a linear correlation (Figure 31).

One of these outlier sequences (L-K23) was inserted by the translocon surprisingly efficiently compared with its ΔG Predictor value (Table 5). We suggest that the stability of L-K23 TM helix within the lipid membrane is derived from intra-helical electrostatic and/or salt-bridge interaction between the histidine and the glutamic acid side chains positioned at $(i, i+3)$ periodicity. In TM helices, intra-helical charge pairs within the same helix have been reported for appropriately spaced $(i, i+3)$ and $(i, i+4)$, oppositely charged residues (Chin and von Heijne, 2000; Baño-Polo *et al.*, 2012). We support this hypothesis by locating the His-Glu pair at $(i, i+2)$ periodicity, which is non-compatible with intra-helical electrostatic and/or salt-bridge interaction. This mutation strongly reduced the experimental insertion efficiency (Table 5).

As expected, ionizable histidine and glutamic acid residues are present in TM helices at a low frequency level (1.7% and 1.6%, respectively). Nevertheless, among the 792 TM helices included in our database, 84 helices (10.6%) contained both amino acid residues in their sequence, and 15 of these helices present the His-Glu pair at $(i, i+3)$ periodicity (Table 6). Interestingly, only one fourth of these His-Glu pairs (4 from 15) are partly exposed to the lipid face, whereas the rest (11) are buried in the protein interior, emphasizing the necessity to shield the polarity of this interaction from the hydrophobic environment of the membrane core. Moreover, although charged and polar residues can face a high energetic barrier when

inserting into a biological membrane, positive and negative charges within the same (Chin and von Heijne, 2000; Bañó-Polo *et al.*, 2012) and different (Johnson and Parson, 2002; Bañó-Polo *et al.*, 2013) TM helices can interact with each other, thereby drastically reducing this barrier. In fact, these interactions can occur in the proximity of the translocon (Sadlish *et al.*, 2005; Saurí *et al.*, 2007), where a strict coupling of correct tertiary structure formation and membrane insertion can be achieved (Cymer *et al.*, 2015).

The other outlier sequence (L-R23) displayed a completely different behaviour, since in this case the experimentally measured p_i value was lower than the predicted value (Figure 31). This sequence contains a proline residue at a nearly central position in the TM segment. Proline is rarely found in the middle of helices from soluble proteins because it results in distortion of the canonical helical geometry and loss of at least one backbone hydrogen bond (Barlow and Thornton, 1988; von Heijne, 1991). However, proline residues are relatively common in TM helices (Figure 18) (Cordes *et al.*, 2002). This suggests that proline residues may be of particular structural and/or functional significance in membrane proteins, even though they invariably produce deviations from canonical helical structure (Yohannan *et al.*, 2004). To learn how the proline present in the L-R23 mutant reduces membrane insertion, we analyzed the insertion of a mutant with the proline residue positioned near the middle of the helix (P10L/L14P), and a mutant with the proline residue located near the N-terminus of the helix (A6P/P10A), both with the same amino acid composition. The A6P/P10A mutant inserted more efficiently than the original sequence, while the P10L/L14P sequence resulted in only minor changes in terms of membrane insertion.

These results indicate that proline residues are not easily accommodated in the center of the helix, despite their prevalence

DISCUSSION

found in this position ([Figure 22](#)). In general, the presence of proline residues in an α -helix generates a constrained Φ rotamer at the position of the proline, the loss of a hydrogen bond donor and the appearance of steric clashes between the proline cyclic side chain and the peptide backbone. In the case of TM helices, all these effects may eventually increase the polarity of the carbonyl groups of the TM helix at the positions three and four residues N-terminal of the proline location (Cordes *et al.*, 2002), reducing insertion efficiency (Orzáez *et al.*, 2004). Moreover, structural studies have shown that proline substitutions at the end of a TM helix can be accommodated by movement of a small part of the helix, while proline substitutions in the middle can require more complex and difficult to accommodate structural changes (Yohannan *et al.*, 2004). Statistical analyses of TM helices show a similar pattern for proline residue distribution ([Figure 35](#)). Altogether, these results indicate that the proline in the original sequence can diminish membrane insertion efficiency depending on the position along the TM segment.

In summary, our analysis of the membrane insertion of computationally designed TM sequences resulted in a good correlation between the values predicted by the ΔG Predictor and experimentally measured values. Nevertheless, our data indicate that some extra attention has to be paid to accommodate intra-helical salt-bridge formation and proline residues when designing short TM helices as building blocks of membrane proteins, a major challenge when engineering new membrane proteins to perform biomimetic functions.

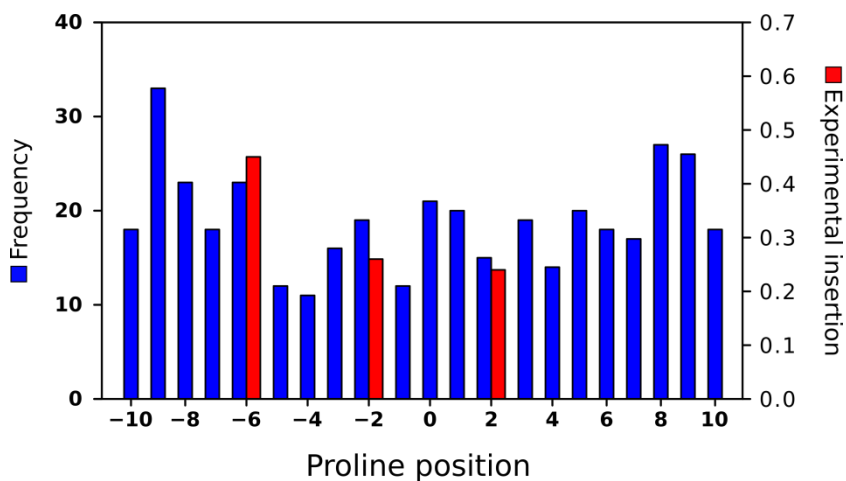


Figure 35. Proline positions along TM helices. The frequency of proline residues in TM helices is shown as a function of their position within the helices. Positively labelled positions indicate the cytoplasmic side of the membrane whereas negatively labelled positions are indicative of extra-cytoplasmic regions. The measured p_i values for the L-R23-derived sequences containing single proline residues are shown in red.

CONCLUSIONS

CONCLUSIONS

1. TM helices adapt their length to the dimensions and constraints of biological membranes. The observed differences relative to helices from soluble proteins highlight that in the lipid bilayer, where the environment forces secondary structure formation, amino acid side chain hydrophobicity prevails to helicity.
2. Half of the natural amino acid residues are equally distributed along the TM helices, whilst aromatic, polar and charged residues plus Pro are biased towards the ends of the TM helices. Specially, the distribution of charged residues is asymmetric occurring more frequently on the cytoplasmic side of the membrane. In addition to this asymmetry, Trp, Tyr and Pro residues are found to be more frequent at the extra-cytoplasmic interface of the membrane and the polar residues (Gln, His, and Asn) at the extra-cytoplasmic flanking region of the TM helices.
3. Insertion efficiency predictions for computed sequences increase with length, and for those sequences harbouring polar and charged residues insertion predictions increase significantly when residue position-dependent distribution is taken into account.
4. The experimental insertion efficiency observed correlates well with predicted values. However, for long sequences designed without position-dependent constraints, some differences between predicted and experimental insertions are observed.
5. Our data demonstrate that to the *de novo* design of TM segments as building blocks of membrane proteins, special attention has to be paid to accommodate intra-helical salt-bridges and proline residues.

RESUMEN

R.1. INTRODUCCIÓN

La membrana biológica

Las membranas biológicas constituyen los límites que rodean y mantienen la integridad celular, actuando como una barrera que separa el interior de la célula del ambiente exterior. La hidrofobicidad de la membrana permite el paso de moléculas apolares al tiempo que evita que compuestos polares o iones la atraviesen por difusión libre, permitiendo así la creación de un potencial eléctrico y de gradientes de concentración a ambos lados de la membrana. Además, la membrana celular eucariótica forma orgánulos, resultando en la compartimentación necesaria para el correcto funcionamiento de las rutas metabólicas. Además de su papel como barrera semipermeable, las membranas biológicas participan en la división celular, la comunicación con el exterior, el tráfico molecular y la reproducción biológica.

Los dos componentes principales de las membranas son los lípidos y las proteínas, moléculas caracterizadas por ser anfifílicas, una propiedad que determina la estructura de una membrana en una solución acuosa. Los lípidos están organizados en una doble capa que favorece y estabiliza termodinámicamente la membrana por la interacción de cadenas alifáticas hidrofóbicas que se enfrentan en el interior de la membrana excluyendo moléculas de agua, con los grupos polares orientados hacia medio acuoso externo.

Uno de los primeros modelos de membrana celular que explica cómo se relacionan los lípidos y las proteínas que la forman, fue el propuesto por Singer y Nicolson (Singer and Nicolson, 1972), conocido como el “modelo del mosaico fluido” (Fig. 1). En este modelo, la membrana se asemeja a un océano de lípidos con unas pocas proteínas flotando en él, y en el que lípidos y proteínas se

RESUMEN

encuentran en constante movimiento lateral. Sin embargo, aproximaciones experimentales más recientes nos proporcionan una visión más compacta de la bicapa, con un elevado número de proteínas insertadas en la membrana, limitando así la libre difusión lateral de las moléculas. Esto sugiere, por tanto, que las membranas son más “mosaico que fluidas” (Engelman, 2005; Goñi, 2014).

Proteínas de membrana

Las proteínas de membrana son aquellas que residen y ejercen su función en membranas biológicas. Éstas desempeñan papeles cruciales y específicos, y están implicadas en muchas funciones biológicas distintas. Además de mantener la forma de la bicapa lipídica, las proteínas de membrana funcionan como receptores, transductores de señales, transportadores, canales, motores o anclajes y participan en señalización, transporte, procesos enzimáticos y adhesión celular. Alrededor del 25% de los genes del genoma de organismos procariotas y eucariotas codifican proteínas de membrana (Krogh *et al.*, 2001), y representan más del 50% de las dianas terapéuticas de los fármacos actuales (Overington *et al.*, 2006).

En función de la interacción que establecen con la bicapa lipídica, las proteínas de membrana pueden diferenciarse entre proteínas integrales de membrana y proteínas periféricas. Las proteínas integrales de membrana se encuentran insertadas dentro de la bicapa, rodeadas de lípidos. Están fuertemente unidas a la membrana y sólo pueden separarse con tratamientos agresivos tales como el uso de detergentes, disolventes orgánicos o de agentes desnaturizantes. Las proteínas periféricas no interactúan directamente con el núcleo hidrofóbico de la membrana, sino que se asocian a la interfase de la bicapa a través de otras proteínas o de grupos lipídicos unidos covalentemente a la proteína.

La determinación de la estructura tridimensional de proteínas de membrana ha revelado dos tipos principales de motivos estructurales: los haces de hélices α y los barriles β (Heijne, 1994; Vinothkumar and Henderson, 2010) (Fig. 4). Mientras que la mayoría de las proteínas en las membranas eucariotas son haces de hélices α , los barriles β se encuentran casi exclusivamente en la membrana externa de bacterias, mitocondrias y cloroplastos. Se ha estimado que el 2-3% de los genes de las bacterias Gram-positivas codifican proteínas con estructura basada en barriles β (Wimley, 2003). En comparación, las proteínas helicoidales representan aproximadamente el 25% de todas las pautas de lectura abierta en los genomas completamente secuenciados (Krogh *et al.*, 2001). Estos dos tipos de estructura secundaria permiten la inserción en la membrana del esqueleto polipeptídico de la proteína maximizando el número de enlaces de hidrógeno intramoleculares, y reduciendo así la polaridad intrínseca de los grupos CO y NH del enlace peptídico.

A partir de aquí, en esta Tesis Doctoral nos referiremos exclusivamente a proteínas de membrana helicoidales, ya que, como se ha descrito, éstas constituyen el grupo más numeroso y relevante de proteínas de membrana.

Plegamiento y estabilidad de proteínas de membrana helicoidales

El conocimiento actual sobre el plegamiento de proteínas de membrana helicoidales se basa en el modelo propuesto a principios de los años 90 por Popot y Engelman, conocido como modelo de dos etapas (Popot and Engelman, 1990) (Fig. 5). En la primera etapa, las secuencias hidrofóbicas adoptan una conformación α -helicoidal cuando se insertan en la membrana. En presencia de agua y una bicapa lipídica, una hélice α transmembrana (TM) representa el estado termodinámicamente más estable para una secuencia de aminoácidos

RESUMEN

no polares. Dado que el número de enlaces de hidrógeno del esqueleto polipeptídico no es crítico en un medio acuoso, se considera que la transición entre una secuencia no estructurada y una hélice α estará gobernada por la naturaleza de las cadenas laterales de los aminoácidos. En la segunda etapa, que probablemente ocurre al mismo tiempo que la inserción en la bicapa y la adquisición de la estructura secundaria, nuevos segmentos TM se asocian con los que ya están insertados en la membrana, dando lugar a la formación de la estructura terciaria.

El principal determinante para la inserción de una secuencia en la membrana es su valor global de hidrofobicidad. En otras palabras, la tendencia a la inserción aumenta con el grado de hidrofobicidad de la secuencia polipeptídica. Este concepto entra en conflicto con la visión clásica de los residuos cargados dentro de la membrana, cuya presencia en el núcleo hidrofóbico se pensó prohibido durante muchos años. Estudios más recientes han mostrado que este tipo de aminoácidos no impide la inserción en la bicapa de una secuencia determinada, sino que la inserción de un segmento TM depende de la hidrofobicidad global de la secuencia (MacCallum *et al.*, 2008; Martínez-Gil *et al.*, 2008). Un efecto similar se observa con aquellos aminoácidos que alteran la formación de la estructura α -helicoidal, tales como la prolina (Nilsson *et al.*, 1998) o la glicina (Dong *et al.*, 2012).

Aunque la hidrofobicidad es la característica principal de las secuencias TM, una fracción considerable ($> 30\%$) de las hélices TM de proteínas integrales de membrana que atraviesan la bicapa más de una vez no son lo suficientemente hidrofóbicas como para ser insertadas de manera eficiente por sí mismas (Hessa *et al.*, 2007). Este hallazgo sugiere que la inserción en la membrana de segmentos TM puede depender en muchos casos de las características de la secuencia

extrínsecas al propio segmento hidrofóbico (Hedin *et al.*, 2010; Öjemalm *et al.*, 2012).

Además de los requisitos de hidrofobicidad y helicidad, una hélice TM debe ser lo suficientemente larga (en número de residuos de aminoácidos) como para poder atravesar la bicapa lipídica. Teniendo en cuenta una translación de 1.5 Å por residuo en una hélice α canónica, se requieren aproximadamente unos 20 residuos para abarcar el núcleo hidrofóbico (~ 30 Å) de la membrana (Fig. 3). Sin embargo, la longitud mínima necesaria para constituir una hélice TM ha sido investigada usando péptidos hidrofóbicos modelo insertados en membranas (Krishnakumar and London, 2007). Estos resultados muestran que en el caso de péptidos compuestos por residuos de leucina y alanina alternos, la longitud mínima necesaria para adoptar una disposición predominantemente transmembrana en bicapas lipídicas modelo es de 13 residuos. Más recientemente, la disposición en la membrana de secuencias de poli-leucina se analizó utilizando péptidos sintéticos y bicapas fosfolipídicas orientadas, inserción *in vitro* en membranas biológicas y simulaciones de dinámica molecular (Jaud *et al.*, 2009). La conclusión que emerge de estos estudios es que las bicapas lipídicas se adaptan a hélices TM tan cortas como 10-12 leucinas.

Biogénesis de proteínas de membrana helicoidales

La gran mayoría de las proteínas se sintetizan en el citosol, donde el ribosoma traduce los codones de RNA mensajeros en aminoácidos cataliza la formación de enlaces peptídicos. Las proteínas citosólicas (solubles) se pliegan cuando emergen del ribosoma. Sin embargo, para las proteínas de secreción y las de membrana el proceso es más complicado, ya que tienen que atravesar la membrana, total o parcialmente. Para ello, utilizan una maquinaria celular denominada

RESUMEN

translocón. El translocón es un complejo multiproteico situado en la membrana del retículo endoplasmático (ER), que consiste básicamente en un canal que atraviesa la bicapa lipídica (Fig. 8). El translocón permite que las proteínas solubles pasen completamente a través de la membrana ER, y los fragmentos TM de las proteínas integrales se introduzcan lateralmente en la bicapa (Panzner *et al.*, 1995), lo que le convierte en el único canal conocido que debe permitir el paso de moléculas en dos direcciones, perpendicular (de un lado a otro de la membrana) para la translocación de dominios y proteínas solubles, y lateral para la inserción de las regiones TM.

El mecanismo en el que la translocación o inserción de la proteína en la membrana se produce a la vez que la traducción es el utilizado por las proteínas de secreción y por la mayoría de las proteínas de membrana. Esta vía ‘co-traducciona’ comienza cuando una secuencia señal (SS), o en su ausencia el primer fragmento TM de la cadena nascente, emerge del ribosoma y es reconocida por la partícula de reconocimiento de señal (SRP) (Fig. 9). La SS tiene una región N-terminal cargada positivamente seguida por un dominio hidrófobo compuesto de 7-15 aminoácidos y una región C-terminal polar. Cuando la SRP se une a la SS, interrumpe la elongación de la cadena nascente (Walter and Blobel, 1981; Mary *et al.*, 2010) y lleva todo el complejo ribosoma-cadena nascente (RNC) a la membrana del ER interactuando con el receptor de la SRP (SR) (Akopian *et al.*, 2013). Cuando el complejo formado por SRP, SR y RNC se acopla al translocón, el ribosoma puede reiniciar la elongación de la proteína que está sintetizando liberando la cadena nascente al canal de translocón. El translocón permitirá que los dominios solubles atraviesen la membrana y que segmentos TM hidrofóbicos salgan lateralmente a la fase lipídica de la membrana (Fig. 9) (Nyathi *et al.*, 2013; Whitley and Mingarro, 2014).

Topología de las proteínas de membrana

La correcta topología de una proteína de membrana (determinación del número de segmentos TM y su orientación relativa en la bicapa) es fundamental para poder llevar a cabo su función biológica. En general, una proteína puede adoptar una única topología en la membrana, aunque en la última década se han identificado casos en los que la misma secuencia es capaz de insertarse en la bicapa con dos topologías opuestas (Rapp *et al.*, 2006, 2007; Seppälä *et al.*, 2010).

El principal determinante de la topología de una proteína de membrana, la denominada '*positive-inside rule*' (von Heijne, 1986a), establece que las regiones extramembranas no translocadas (orientadas al citosol) contienen de dos a cuatro veces más residuos cargados positivamente (Arg y Lys) que las regiones extramembranas translocadas. Las cargas en el lado N-terminal de la SS pueden influir en la orientación del segmento hidrofóbico N-terminal de una proteína de membrana. Si tiene cargas positivas, la parte N-terminal tendrá una mayor probabilidad de permanecer en el citosol y el extremo C-terminal estará expuesto al periplasma bacteriano o al lumen del ER en el caso de organismos eucarióticos (von Heijne, 1986b).

La longitud de la secuencia hidrofóbica también puede determinar la orientación del fragmento TM. En eucariotas las secuencias más largas muestran una preferencia por localizar el extremo N-terminal en el lumen (Sakaguchi *et al.*, 1992; Wahlberg and Spiess, 1997; Eusebio *et al.*, 1998). De hecho, las SS tienen segmentos hidrofóbicos más cortos que los fragmentos TM y orientan la parte N-terminal al citosol (Nilsson *et al.*, 1994), confirmando la longitud de la secuencia TM como un determinante en la definición de la orientación de la hélice hidrofóbica.

RESUMEN

Además de la longitud de la secuencia, tanto su grado como su gradiente de hidrofobicidad pueden influir en la orientación en la bicapa de un dominio TM. Cuanto más hidrofóbico es el segmento TM, más tiende a insertarse con una orientación con el extremo N-terminal hacia el exterior celular y el extremo C-terminal hacia el citosol, independientemente de las cargas de la región flanqueante (Wahlberg and Spiess, 1997; Goder and Spiess, 2003). Por el contrario, las SS naturales, que presentan una orientación inversa, son menos hidrofóbicas y permanecen más tiempo en el translocón, lo que les permite reorientarse durante la traducción (Whitley and Mingarro, 2014).

Esta Tesis se centra en el estudio de los dominios TM de proteínas de membrana helicoidales, un tipo de moléculas que comprenden el 20-30% de todos los genes en los genomas secuenciados. Sin embargo, debido a la complejidad del entorno en el que se encuentran, nuestro conocimiento acerca de su biogénesis y plegamiento queda todavía muy lejos del que tenemos de las proteínas globulares.

R.2. OBJETIVOS

El objetivo general de esta tesis es la caracterización de la inserción y el ensamblaje en la membrana lipídica de hélices α procedentes de proteínas de membrana, y cómo la inserción se ve afectada por la longitud de la secuencia y la composición de aminoácidos. Durante el estudio de estas hélices TM, en esta Tesis se han abordado los siguientes objetivos concretos:

- Describir las diferencias entre hélices TM y hélices solubles en términos de longitud y composición de aminoácidos.
- Estudiar los patrones de distribución de aminoácidos en función de la posición en la hélice TM.
- Analizar la inserción teórica de secuencias generadas computacionalmente con diferente longitud y composición de aminoácidos.
- Determinar la inserción experimental en membranas microsomales de secuencias generadas computacionalmente mediante un sistema de traducción/glicosilación *in vitro*.

R.3. METODOLOGÍA

r.3.1. Métodos computacionales

Bases de datos de α -hélices

Los conjuntos de datos de hélices solubles y de hélices TM fueron obtenidos del PDB (Berman *et al.*, 2000) y de la base de datos MPTOPO (Jayasinghe *et al.*, 2001), respectivamente.

En primer lugar, el conjunto de datos de hélices solubles fue construido seleccionando un total de 4.405 proteínas globulares depositadas en el PDB (hasta el 17 de noviembre de 2011) que pasaron los siguientes criterios: (i) su estructura secundaria total tenía más del 60% de hélices α y no contenía hojas β ; (ii) su resolución cristalográfica era de 2.0 Å o superior; y (iii) la palabra MEMBRANE no aparecía en los campos "TITLE" ni "DESCRIPTION" del archivo PDB. Además, para eliminar redundancia, las 4.405 secuencias se compararon entre sí con el programa CD-HIT (Huang *et al.*, 2010) y los pares de secuencia alineados que obtuvieron una identidad del 80% o superior fueron descartados. El conjunto final no redundante de cadenas polipeptídicas del PDB fue analizado para identificar un total de 7.348 hélices mediante el campo "HELIX" de cada entrada del PDB. Así, el conjunto de datos de hélices solubles contenía 930 estructuras de proteínas no redundantes y de alta resolución, 7.348 hélices α y 108.277 aminoácidos.

En segundo lugar, se seleccionaron todas las proteínas de membrana helicoidales depositadas en la base de datos MPTOPO (actualizada al 19 de enero de 2010) (Jayasinghe *et al.*, 2001) y, por tanto, con topología definida. El conjunto inicial se filtró más a fondo: (i) eliminando cualquier entrada de estructura desconocida basada en la clasificación de entrada MPTOPO (es decir, manteniendo sólo las

entradas descritas como "3D_helix" y "1D_helix"); (ii) eliminando pares de secuencias redundantes con una identidad de secuencia del 80% o superior mediante el programa CD-HIT (Huang *et al.*, 2010). El conjunto final de datos de hélices TM contenía 170 estructuras polipeptídicas no redundantes, 837 hélices y 20.079 aminoácidos. Además, para analizar adecuadamente la distribución de aminoácidos en hélices de proteínas de membrana monotópicas (que atraviesan la bicapa una sola vez), descartamos cualquier hélice de menos de 17 o más de 38 aminoácidos de longitud. El subconjunto de datos de hélices transmembrana resultante contenía 792 hélices y 19.356 aminoácidos.

Distribución de aminoácidos

Se calcularon tres diferentes medidas para cada tipo de aminoácidos: (i) probabilidad y porcentaje, (ii) Odds, y (iii) LogOdds. La probabilidad (p_i) de un aminoácido i se define como:

$$p_i = \frac{n_i}{N} \quad [1]$$

donde i es el tipo de aminoácido (uno de los 20 aminoácidos), n_i es el número de residuos del aminoácido i , y N es el total de aminoácidos presentes en el conjunto de datos. De forma similar, el porcentaje de un aminoácido dado i se define como su probabilidad multiplicada por 100. Las Odds (O_i) de un aminoácido i se definen como:

$$O_i = \frac{p_{i,c}}{(1-p_{i,c})} / \frac{p_{i,r}}{(1-p_{i,r})} \quad [2]$$

donde $p_{i,c}$ es la probabilidad del aminoácido i en la clase c (por ejemplo, hélice transmembrana) y $p_{i,r}$ es la probabilidad del aminoácido i en la clase r (por ejemplo, hélice soluble). Del mismo

RESUMEN

modo, las LogOdds de un aminoácido dado i se definen como el logaritmo en base 10 de sus Odds. En pocas palabras, un Odds mayor que 1 (o LogOdds positivos) indican una sobre-representación de ese tipo de aminoácido en esa clase. Por el contrario, Odds menores que 1 (o LogOdds negativos) indican una menor tendencia de ese tipo de aminoácido en esa clase.

Diseño computacional de secuencias

Usando el conjunto de datos de hélices TM mencionado anteriormente (792 hélices y 19.356 aminoácidos), se generaron una serie de secuencias de distinta longitud usando en su composición un número creciente de tipo de aminoácidos, siguiendo el orden de frecuencia encontrado en nuestro conjunto de datos de hélices TM.

En primer lugar, se generaron secuencias formadas sólo por leucina (el aminoácido más común en hélices TM) de distintas longitudes, desde 9 a 25 residuos. A continuación, se incluyó el segundo aminoácido más común (es decir, alanina) con su probabilidad relativa en comparación con Leu. Se generaron 1.000 secuencias de cada longitud (de 9 a 25 residuos), donde cada secuencia se obtuvo mezclando una cadena de letras de Leu y Ala con las proporciones entre ellos encontradas en nuestro conjunto de datos (Leu 58,2% y Ala 41,8%). Este procedimiento se repitió incluyendo en la composición de las secuencias cada vez un nuevo aminoácido, siguiendo el orden de presencia encontrado en hélices TM, hasta incluir los 20 aminoácidos naturales ([Fig. 15](#)).

A continuación, se generaron conjuntos de secuencias incluyendo no solo la frecuencia sino también la distribución por posición de aminoácidos. Para ello, se cogieron los residuos de segmentos TM y se anotó su posición relativa con respecto al centro de la hélice. En este caso, se calculó la probabilidad de encontrar cada tipo de

aminoácido en cada una de las posiciones de un segmento TM empezando por el residuo central de la hélice (posición 0) e incrementando valores positivos conforme se aproximaba a la parte citoplasmática de la hélice, o valores negativos conforme se aproximaba a la parte extracelular de la secuencia. De esta forma, se crearon subconjuntos con los aminoácidos encontrados en cada una de las posiciones relativas de las hélices TM, y se usaron dichos subconjuntos para diseñar secuencias teniendo en cuenta la posición de los aminoácidos. De forma similar al procedimiento seguido con las secuencias creadas sin tener en cuenta la posición, se generaron 1.000 secuencias de cada longitud (de 9 a 25 residuos) para cada composición de aminoácidos distinta (Fig. 15).

Predicción de los valores de ΔG_{app} y probabilidad de inserción

Para cada uno de los subconjuntos de 1.000 secuencias con determinada longitud y composición de aminoácidos (tanto del conjunto de secuencias creadas sin tener en cuenta la posición como del conjunto creado incluyendo este factor) se generó una serie de valores teóricos de inserción, utilizando el programa *ΔG Predictor* (<http://dgpred.cbr.su.se/>) (Hessa *et al.*, 2005, 2007). Este programa asigna a cada secuencia un valor teórico de variación de energía libre (ΔG); valores negativos indican una mayor tendencia a la inserción en la membrana mientras que valores positivos indican mayor probabilidad de translocación. La variación de energía libre aparente (ΔG_{app}) se describe como:

$$\Delta G_{app} = -RT \ln K_{app} \quad [3]$$

RESUMEN

donde $T = 298 \text{ K}$, $R = 0.0019872 \text{ kcal}/(\text{K}\cdot\text{mol})$ y K_{app} es la constante de equilibrio entre la fracción de moléculas insertadas (f_i) y la fracción de moléculas translocadas (f_t):

$$K_{app} = \frac{f_i}{f_t} \quad [4]$$

Dado que

$$f_i + f_t = 1 \quad [5]$$

y que la fracción de moléculas insertadas (f_i) es igual que la probabilidad de inserción (p_i), la ecuación 3 se puede reescribir de la siguiente manera:

$$\Delta G_{app} = -RT \ln\left(\frac{p_i}{1-p_i}\right) \quad [6]$$

De esta manera, reorganizando la ecuación 6 podemos calcular la probabilidad de inserción (p_i) a partir del valor de ΔG_{app} obtenido:

$$p_i = \frac{e^{\left(\frac{\Delta G_{app}}{-RT}\right)}}{1 + e^{\left(\frac{\Delta G_{app}}{-RT}\right)}} \quad [7]$$

Todos los análisis descritos anteriormente se realizaron en Python (Python Software Foundation, version 2.7, <https://www.python.org/>), R (The R Foundation, <https://www.r-project.org/>) y RStudio (<https://www.rstudio.com/>).

r.3.2. Métodos experimentales

Material biológico

La cepa utilizada para la extracción rutinaria de DNA plasmídico fue *E. coli* DH5 α (genotipo: Δ lacZ Δ M15 Δ (lacZYA-argF) U169 recA1 endA1 hsdR17(rK-mK+) supE44 thi-1 gyrA96 relA1) (Taylor *et al.*, 1993). Las células competentes fueron preparadas en el laboratorio siguiendo el protocolo establecidos (Sambrook and Russell, 2001).

Para el crecimiento de las bacterias se usó medio LB (*Luria Bertani*, compuesto de extracto de levadura 0.5% (p/v), triptona 1% (p/v) y NaCl 1%) líquido o sólido (añadiendo 2% (p/v) de agar bacteriológico). El medio se autoclavó para su esterilización durante 20 minutos a 1 atmósfera de presión y 121°C, y fue suplementado con el antibiótico apropiado en cada caso. Las células se cultivaron a 37°C con agitación (250 rpm) durante al menos 12 horas.

Para introducir los plásmidos de interés dentro de las cepas de *E. coli* se incubaron en hielo 10 ng de DNA junto con 50 μ L de células competentes durante 30 minutos. Tras este periodo las células se sometieron a un choque térmico de 42°C durante 45 segundos y posteriormente se incubaron 5 minutos en hielo. Seguidamente se añadieron 500 μ L de LB y se recuperaron las células a 37°C durante 45 minutos en agitación suave. Las células se recogieron por sedimentación a 3.000 rpm durante 5 minutos y se eliminaron 450 μ L del sobrenadante. Las células fueron resuspendidas con el resto del sobrenadante (aproximadamente 50 μ L) y sembradas en una placa *Petri* de LB, suplementada con el antibiótico requerido en cada caso. Dichas placas se mantuvieron en una incubadora a 37°C durante al menos 12 horas.

En aquellas ocasiones en las que se requería una alta eficiencia de transformación del DNA se recurrió a la transformación por

RESUMEN

electroporación. En este caso, 50 μL de células electrocompetentes se incubaron durante 5 minutos con 5 ng de DNA plasmídico. A continuación, las células fueron sometidas a un choque eléctrico de $1,8 \text{ kV}/\text{cm}^2$ durante 5 milisegundos, para lo cual se emplearon cubetas de electroporación con una distancia entre los electrodos de 0,1 cm. Para permitir la recuperación de las células, a la mezcla de transformación se le añadió 500 μL de medio LB y se incubaron a 37°C con agitación durante 45 minutos, pasados los cuales las células se recogieron y sembraron en placas como se explica anteriormente.

Manipulación del DNA

Tanto el aislamiento de DNA plasmídico de *E. coli* como la purificación en gel de agarosa se realizaron utilizando kits comerciales de la casa Thermo (Ulm, Alemania), siguiendo las instrucciones del fabricante.

Las secuencias ensayadas experimentalmente se construyeron mediante el método de hibridación de oligonucleótidos, utilizando oligonucleótidos bicatenarios con extremos 5' protuberantes y solapantes. Primero, se hibridaron pares de oligonucleótidos complementarios incubándolos a 85°C durante 10 minutos tras los cuales se dejó enfriar lentamente hasta una temperatura de 30°C . A continuación, se mezclaron los dos pares de oligonucleótidos bicatenarios recién hibridados y se incubaron a 65°C durante 5 minutos, dejándolos posteriormente enfriar lentamente hasta temperatura ambiente (unos 25°C). Los oligonucleótidos resultantes, conteniendo la secuencia nucleotídica codificante de la secuencia a ensayar, se purificaron en gel de agarosa al 2% (p/v), tras lo cual se fosforilaron y ligaron en el vector previamente preparado. El vector se preparó mediante digestión con las enzimas de restricción adecuadas, defosforilación y purificación en gel de agarosa al 1% (p/v).

Los mutantes realizados sobre las secuencias construidas se obtuvieron mediante mutagénesis dirigida empleando el kit comercial *QuikChange* de Agilent Technologies (Santa Clara, CA, EE.UU.). La mezcla de reacción (2,5 μ L tampón de reacción 10x, 1 μ L mezcla de dNTPs 25 mM, 250 ng de cada uno de los dos oligos y 2,5 U de polimerasa Pfu Turbo en 25 μ L de reacción ajustados con agua miliQ) se sometió a 25 ciclos de amplificación (desnaturalización a 95°C durante 50 segundos, hibridación a 58-60°C durante 1 minuto y elongación a 68°C durante 12 minutos). El DNA resultante fue sometido a digestión con *DpnI* para eliminar el DNA molde (metilado), y seguidamente transformado en *E. coli* utilizando 1 μ L del producto de PCR.

Todos los oligos usados tanto en la realización de las construcciones como en las mutagénesis dirigidas se muestran en el Anexo I.

Ensayos de glicosilación

Para analizar la inserción experimental en membranas de las secuencias generadas computacionalmente se utilizó el plásmido pGEM-Lep. En este plásmido, la peptidasa de la secuencia señal de *E. coli* (Lep) se encuentra bajo un promotor SP6. Lep está formada por dos segmentos TM (H1 y H2) unidos por un lazo citoplasmático (P1) y un dominio C-terminal grande (P2), y se inserta en microsomas derivados del retículo endoplasmático con ambos extremos localizados en el lumen (Figure 16). La secuencia a ensayar (“TM-tested”) se insertó mediante ingeniería genética en el dominio luminal P2, y flanqueada por dos dianas de *N*-glicosilación (G1 y G2). Los sitios de glicosilación pueden servir para detectar de la inserción en la membrana ya que G1 siempre será glicosilado por el complejo OST (Fig. 12) debido a su localización luminal, mientras que G2 sólo será

RESUMEN

glicosilado si la región analizada es translocada a través de la membrana. Una construcción en la que la secuencia a ensayar se inserta en la membrana tendrá una única glicosilación, lo que aumentará su masa molecular en aproximadamente 2,5 kDa con respecto a la masa molecular de la proteína expresada en ausencia de membranas microsomales. En cambio, si la secuencia es translocada, ambas dianas G1 y G2 serán glicosiladas, lo que aumentará la masa molecular de la proteína en unos 5 kDa.

Transcripción y traducción in vitro

Las construcciones realizadas en el plásmido pGEM-Lep fueron transcritas y traducidas *in vitro* utilizando el sistema TNT SP6 Quick Coupled System (Promega). A 5 μ L de lisado de reticulocitos se le añadió un total de 75 ng de DNA molde, 0,5 μ L de metionina marcada radiactivamente [35 S-Met] (5 μ Ci) y 0,25 μ L de membranas microsomales (tRNA Probes), y las mezclas fueron incubadas a 30°C durante 90 minutos. Los productos de traducción se diluyeron en 50 μ L de tampón de carga (concentración final Tris-HCl 625 mM pH 6.8, glicerol 10% (v/v), SDS 2 % (p/v), β -mercaptoetanol 4% (v/v) y azul de bromofenol 0,025% (p/v)) y fueron analizados mediante electroforesis desnaturante en gel de poliacrilamida en presencia de SDS (dodecilsulfato sódico). A continuación los geles se analizaron y cuantificaron usando un *Fujifilm Fluorescent Image Analyzer* modelo FLA-300R (Tokyo, Japón) y el software *Image Gauge* v4.0 de Fuji Photo Film.

La probabilidad de inserción en la membrana (p_i) de una secuencia dada fue calculada como el cociente entre la intensidad de la banda mono-glicosilada (f_{1g} , insertado) y la suma de las bandas mono- y doble-glicosiladas (f_{2g} , translocado):

$$p_i = \frac{f_{1g}}{f_{1g}+f_{2g}} \quad [8]$$

El valor experimental de ΔG fue calculado con la siguiente fórmula:

$$\Delta G_{app}^{exp} = -RT \ln K_{app}^{exp} \quad [9]$$

donde, en este caso, la constante de equilibrio experimental (K_{app}^{exp}) se define como el cociente entre la intensidad de la banda monoglicosilada (f_{1g} , insertado) y la banda doble-glicosiladas (f_{2g} , translocado):

$$K_{app}^{exp} = \frac{f_{1g}}{f_{2g}} \quad [10]$$

Productos químicos y enzimas

Todas las enzimas utilizadas, así como el plásmido pGEM1, el sistema de transcripción/traducción *in vitro* TNT SP6 Quick Coupled System y los lisados de reticulocitos de conejo fueron adquiridos de Promega (Madison, WI, EE.UU.). Las membranas microsomaes derivadas de ER de páncreas de perro se adquirieron a tRNA Probes (College Station, TX, EE.UU.). La metionina marcada radiactivamente se compró a Perkin Elmer (Waltham, MA, EE.UU.). Las enzimas de restricción se adquirieron de Roche Molecular Systems (Pleasanton, CA, EE.UU.). Los kits de purificación de DNA y de mutagénesis dirigidas se compraron a Thermo (Ulm, Alemania). Los oligonucleótidos fueron adquiridos de Sigma-Aldrich (Suiza). Todas

RESUMEN

las construcciones realizadas, así como los mutantes, fueron confirmadas por secuenciación del DNA mediante el servicio de secuenciación de MacroGen Europe (Ámsterdam, Países Bajos).

R.4. CONCLUSIONES

1. Las hélices TM adaptan su longitud a las dimensiones y restricciones de las membranas biológicas. Las diferencias observadas con respecto a las hélices de proteínas solubles destacan que en la bicapa lipídica, donde el medio ambiente fuerza la formación de estructura secundaria, la hidrofobicidad de la cadena lateral de aminoácidos prevalece a la helicidad.
2. La mitad de los aminoácidos naturales se distribuyen uniformemente a lo largo de las hélices TM, mientras que los residuos aromáticos, polares y cargados, además de Pro, presentan una distribución sesgada hacia los extremos de las hélices TM. Especialmente, la distribución de residuos cargados es asimétrica, encontrándose con mayor frecuencia en el lado citoplasmático de la membrana. Además, los residuos de Trp, Tyr y Pro resultan más frecuentes en la interfase extra-citoplasmática de la membrana y los residuos polares (Gln, His y Asn) en la región flanqueante extra-citoplasmática de las hélices TM.
3. La inserción teórica de las secuencias generadas computacionalmente incrementa con la longitud, y para las secuencias que albergan residuos polares y cargados la inserción teórica aumenta significativamente cuando se incluye en su diseño computacional la distribución por posición de los aminoácidos.

RESUMEN

4. La inserción experimental de las secuencias diseñadas computacionalmente, medida mediante ensayos de glicosilación, presenta una alta correlación con los valores teóricos. Sin embargo, en secuencias largas diseñadas sin tener en cuenta la posición que incluyen residuos polares y/o cargados, las diferencias entre la inserción teórica y experimental son mayores.
5. Nuestros datos demuestran que en el diseño *de novo* de segmentos TM como bloques de construcción de proteínas de membrana se debe prestar especial atención a la disposición de puentes salinos intra-helicoidales y de residuos de prolina.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Akopian, D., Shen, K., Zhang, X. and Shan, S. (2013) 'Signal Recognition Particle: An Essential Protein-Targeting Machine', *Annual Review of Biochemistry*, **82(1)**, pp. 693–721.
- Andersson, H. and von Heijne, G. (1993) 'Sec dependent and sec independent assembly of E. coli inner membrane proteins: the topological rules depend on chain length.', *The EMBO journal*, **12(2)**, pp. 683–691.
- Bañó-Polo, M., Baeza-Delgado, C., Orzáez, M., Marti-Renom, M. A., Abad, C. and Mingarro, I. (2012) 'Polar/Ionizable Residues in Transmembrane Segments: Effects on Helix-Helix Packing', *PLoS ONE*, **7(9)**, pp. 1–8.
- Bañó-Polo, M., Martínez-Garay, C. A., Grau, B., Martínez-Gil, L. and Mingarro, I. (2017) 'Membrane insertion and topology of the translocon-associated protein (TRAP) gamma subunit', *Biochimica et Biophysica Acta (BBA) - Biomembranes*. Elsevier B.V., **1859(5)**, pp. 903–909.
- Bañó-Polo, M., Martínez-Gil, L., Wallner, B., Nieva, J. L., Elofsson, A. and Mingarro, I. (2013) 'Charge pair interactions in transmembrane helices and turn propensity of the connecting sequence promote helical hairpin insertion', *Journal of Molecular Biology*. Elsevier B.V., **425(4)**, pp. 830–840.
- Barlow, D. J. and Thornton, J. M. (1988) 'Helix geometry in proteins', *Journal of Molecular Biology*, **201(3)**, pp. 601–619.
- Berg, B. van den, Jr, W. M. C., Collinson, I., Modis, Y., Hartmann, E., Harrison, S. C. and Tom A. Rapoport (2004) 'X-ray structure of a protein-conducting channel', *Nature*, **427(6969)**, pp. 36–44.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) 'The protein data bank.', *Nucleic acids research*, **28(1)**, pp. 235–242.
- Bernsel, A., Viklund, H., Falk, J., Lindahl, E., von Heijne, G. and Elofsson, A. (2008) 'Prediction of membrane-protein topology from first principles.', *Proceedings of the National Academy of Sciences of the United States of America*, **105(20)**, pp. 7177–7181.
- Bernsel, A., Viklund, H., Hennerdal, A. and Elofsson, A. (2009) 'TOPCONS: Consensus prediction of membrane protein topology', *Nucleic Acids Research*, **37(SUPPL. 2)**, pp. 465–468.
- Blaber, M., Zhang, X.-J. and Matthews, B. W. (1993) 'Structural basis of amino acid α -helix propensity', *Science*, **260(5114)**, pp. 1637–1640.
- Bogdanov, M., Dowhan, W. and Vitrac, H. (2014) 'Lipids and topological rules governing membrane protein assembly', *Biochimica et Biophysica Acta - Molecular Cell Research*. Elsevier B.V., **1843(8)**, pp. 1475–1488.
- Bogdanov, M., Heacock, P. N. and Dowhan, W. (2002) 'A polytopic membrane protein displays a reversible topology dependent on membrane lipid composition', *EMBO Journal*, **21(9)**, pp. 2107–2116.
- Bogdanov, M., Sun, J., Kaback, H. R. and Dowhan, W. (1996) 'A phospholipid acts as a chaperone in assembly of a membrane transport protein', *Journal of Biological Chemistry*, **271(20)**, pp. 11615–11618.

BIBLIOGRAPHY

- Bowie, J. U. (1997) 'Helix packing in membrane proteins', *Journal of Molecular Biology*, **272(5)**, pp. 780–789.
- Bywater, R. P., Thomas, D. and Vriend, G. (2001) 'A sequence and structural study of transmembrane helices', *J Comput Aided Mol Des*, **15(6)**, pp. 533–552.
- Chavan, M., Yan, A. and Lennarz, W. J. (2005) 'Subunits of the translocon interact with components of the oligosaccharyl transferase complex', *Journal of Biological Chemistry*, **280(24)**, pp. 22917–22924.
- Chen, H. and Kendall, D. A. (1995) 'Artificial transmembrane segments. Requirements for stop transfer and polypeptide orientation', *Journal of Biological Chemistry*, **270(23)**, pp. 14115–14122.
- Chin, C. N. and von Heijne, G. (2000) 'Charge pair interactions in a model transmembrane helix in the ER membrane.', *Journal of molecular biology*, **303(1)**, pp. 1–5.
- Choma, C., Gratkowski, H., Lear, J. D. and DeGrado, W. F. (2000) 'Asparagine-mediated self-association of a model transmembrane helix.', *Nature structural biology*, **7(2)**, pp. 161–6.
- Claros, M. G. and von Heijne, G. (1994) 'Toppred II: An improved software for membrane protein structure predictions', *Bioinformatics*, **10(6)**, pp. 685–686.
- Cordes, F. S., Bright, J. N. and Sansom, M. S. P. (2002) 'Proline-induced distortions of transmembrane helices', *Journal of Molecular Biology*, **323(5)**, pp. 951–960.
- Cserzo, M., Eisenhaber, F., Eisenhaber, B. and Simon, I. (2002) 'On filtering false positive transmembrane protein predictions', *Protein Engineering Design and Selection*, **15(9)**, pp. 745–752.
- Cserző, M., Wallin, E., Simon, I., von Heijne, G. and Elofsson, A. (1997) 'Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method.', *Protein engineering*, **10(6)**, pp. 673–676.
- Cymer, F., von Heijne, G. and White, S. H. (2015) 'Mechanisms of integral membrane protein insertion and folding', *Journal of Molecular Biology*. Elsevier Ltd, **427(5)**, pp. 999–1022.
- Dalbey, R. E. and von Heijne, G. (1992) 'Signal peptidases in prokaryotes and eukaryotes - a new protease family', *Trends in Biochemical Sciences*, pp. 474–478.
- Daleke, D. L. (2003) 'Regulation of transbilayer plasma membrane phospholipid asymmetry.', *Journal of lipid research*, **44(2)**, pp. 233–42.
- Denzer, A. J., Nabholz, C. E. and Spiess, M. (1995) 'Transmembrane orientation of signal-anchor proteins is affected by the folding state but not the size of the N-terminal domain.', *The EMBO journal*, **14(24)**, pp. 6311–6317.
- Donald, J. E., Kulp, D. W. and DeGrado, W. F. (2011) 'Salt bridges: Geometrically specific, designable interactions', *Proteins: Structure, Function and Bioinformatics*, **79(3)**, pp. 898–915.
- Dong, H., Sharma, M., Zhou, H. X. and Cross, T. A. (2012) 'Glycines: Role in α -helical membrane protein structures and a potential indicator of native

BIBLIOGRAPHY

- conformation', *Biochemistry*, **51(24)**, pp. 4779–4789.
- Donohue, P. J., Sainz, E., Akeson, M., Kroog, G. S., Mantey, S. A., Battey, J. F., Jensen, R. T. and Northup, J. K. (1999) 'An aspartate residue at the extracellular boundary of TMII and an arginine residue in TMVII of the gastrin-releasing peptide receptor interact to facilitate heterotrimeric G protein coupling', *Biochemistry*, **38(29)**, pp. 9366–9372.
- Eilers, M., Patel, A. B., Liu, W. and Smith, S. O. (2002) 'Comparison of helix interactions in membrane and soluble alpha-bundle proteins.', *Biophysical journal*, **82(5)**, pp. 2720–36.
- Engelman, D. M. (2005) 'Membranes are more mosaic than fluid.', *Nature*, **438(7068)**, pp. 578–580.
- Engelman, D. M., Chen, Y., Chin, C.-N. C.-N., Curran, A. R. R., Dixon, A. M., Dupuy, A. D., Lee, A. S., Lehnert, U., Matthews, E. E., Reshetnyak, Y. K., Senes, A. and Popot, J.-L. J.-L. (2003) 'Membrane protein folding: beyond the two stage model', *FEBS Letters*, **555(1)**, pp. 122–125.
- Engelman, D. M., Steitz, T. A. and Goldman, A. (1986) 'Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins.', *Annual review of biophysics and biophysical chemistry*, **15**, pp. 321–353.
- Eusebio, A., Friedberg, T. and Spiess, M. (1998) 'The role of the hydrophobic domain in orienting natural signal sequences within the ER membrane.', *Experimental cell research*, **241(1)**, pp. 181–5.
- Fons, R. D., Bogert, B. A. and Hegde, R. S. (2003) 'Substrate-specific function of the translocon-associated protein complex during translocation across the ER membrane', *Journal of Cell Biology*, **160(4)**, pp. 529–539.
- Fujita, H., Yamagishi, M., Kida, Y. and Sakaguchi, M. (2011) 'Positive charges on the translocating polypeptide chain arrest movement through the translocon.', *Journal of cell science*, **124(Pt 24)**, pp. 4184–93.
- Goder, V., Junne, T. and Spiess, M. (2004) 'Sec61p Contributes to Signal Sequence Orientation According to the Positive-Inside Rule', *Molecular Biology of the Cell*, **15**, pp. 1470–1478.
- Goder, V. and Spiess, M. (2001) 'Topogenesis of membrane proteins: Determinants and dynamics', in *FEBS Letters*, pp. 87–93.
- Goder, V. and Spiess, M. (2003) 'Molecular mechanism of signal sequence orientation in the endoplasmic reticulum', *EMBO Journal*, **22(14)**, pp. 3645–3653.
- Goerlich, D., Hartmann, E., Prehn, S. and Rapoport, T. A. (1992) 'A protein of the endoplasmic reticulum involved early in polypeptide translocation', *Nature*, **357(6373)**, pp. 47–52.
- Gogala, M., Becker, T., Beatrix, B., Armache, J.-P., Barrio-Garcia, C., Berninghausen, O. and Beckmann, R. (2014) 'Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion.', *Nature*. Nature Publishing Group, **506(7486)**, pp. 107–10.

BIBLIOGRAPHY

- Goñi, F. M. (2014) 'The basic structure and dynamics of cell membranes: An update of the Singer–Nicolson model', *Biochimica et Biophysica Acta (BBA) - Biomembranes*. Elsevier B.V., **1838(6)**, pp. 1467–1476.
- Granseth, E., Von Heijne, G. and Elofsson, A. (2005) 'A study of the membrane-water interface region of membrane proteins', *Journal of Molecular Biology*, **346(1)**, pp. 377–385.
- Hall, J. A., Fann, M. C. and Maloney, P. C. (1999) 'Altered substrate selectivity in a mutant of an intrahelical salt bridge in UhpT, the sugar phosphate carrier of *Escherichia coli*', *Journal of Biological Chemistry*, **274(10)**, pp. 6148–6153.
- Hartmann, E., Gorlich, D., Kostka, S., Otto, A., Kraft, R., Knespel, S., Burger, E., Rapoport, T. A. and Prehn, S. (1993) 'A tetrameric complex of membrane proteins in the endoplasmic reticulum', *Eur J Biochem*, **214(2)**, p. 375–81.
- Hebb, D. O. (1949) *The Organization of Behavior*, New York: Wiley.
- Hedin, L. E., Öjemalm, K., Bernsel, A., Hennerdal, A., Illergård, K., Enquist, K., Kauko, A., Cristobal, S., von Heijne, G., Lerch-Bader, M., Nilsson, I. and Elofsson, A. (2010) 'Membrane Insertion of Marginally Hydrophobic Transmembrane Helices Depends on Sequence Context', *Journal of Molecular Biology*. Elsevier Ltd, **396(1)**, pp. 221–229.
- von Heijne, G. (1986a) 'The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology.', *The EMBO Journal*, **5(11)**, pp. 3021–3027.
- von Heijne, G. (1986b) 'Towards a comparative anatomy of N-terminal topogenic protein sequences', *Journal of Molecular Biology*, **189(1)**, pp. 239–242.
- von Heijne, G. (1991) 'Proline kinks in transmembrane α -helices', *Journal of Molecular Biology*, **218(3)**, pp. 499–503.
- von Heijne, G. (1992) 'Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule', *Journal of Molecular Biology*, **225(2)**, pp. 487–494.
- von Heijne, G. (1994) 'MEMBRANE PROTEINS : from sequence to structure', *Ann. Rev. Biophys. Biomol. Struct.*, **23**, pp. 167–92.
- Heinrich, S. U., Mothes, W., Brunner, J. and Rapoport, T. A. (2000) 'The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain', *Cell*, **102(2)**, pp. 233–244.
- Heinrich, S. U. and Rapoport, T. A. (2003) 'Cooperation of transmembrane segments during the integration of a double-spanning protein into the ER membrane', *EMBO Journal*, **22(14)**, pp. 3654–3663.
- Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S. H. and von Heijne, G. (2005) 'Recognition of transmembrane helices by the endoplasmic reticulum translocon.', *Nature*, **433(7024)**, pp. 377–81.
- Hessa, T., Meindl-Beinker, N. M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, I., White, S. H. and von Heijne, G. (2007) 'Molecular code for transmembrane-helix recognition by the Sec61 translocon.', *Nature*, **450(7172)**,

BIBLIOGRAPHY

pp. 1026–1030.

- Hirokawa, T., Boon-Chieng, S. and Mitaku, S. (1998) 'SOSUI: classification and secondary structure prediction system for membrane proteins.', *Bioinformatics*, **14(4)**, pp. 378–379.
- Holt, A. and Killian, J. A. (2010) 'Orientation and dynamics of transmembrane peptides: The power of simple models', *European Biophysics Journal*, **39(4)**, pp. 609–621.
- Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W. (2010) 'CD-HIT Suite: A web server for clustering and comparing biological sequences', *Bioinformatics*, **26(5)**, pp. 680–682.
- Hunter, C. A. and Sanders, J. K. M. (1990) 'The nature of .pi.-pi. interactions', *Journal of the American Chemical Society*, **112(14)**, pp. 5525–5534.
- Igura, M., Maita, N., Kamishikiryo, J., Yamada, M., Obita, T., Maenaka, K. and Kohda, D. (2008) 'Structure-guided identification of a new catalytic motif of oligosaccharyltransferase.', *The EMBO journal*, **27(1)**, pp. 234–43.
- Jacobs, R. E. and White, S. H. (1989) 'The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices.', *Biochemistry*, **28**, pp. 3421–3437.
- Jaud, S., Fernández-Vidal, M., Nilsson, I., Meindl-Beinker, N. M., Hübner, N. C., Tobias, D. J., von Heijne, G. and White, S. H. (2009) 'Insertion of short transmembrane helices by the Sec61 translocon.', *Proceedings of the National Academy of Sciences of the United States of America*, **106(28)**, pp. 11588–11593.
- Jayasinghe, S., Hristova, K. and White, S. H. (2001a) 'Energetics, stability, and prediction of transmembrane helices', *J Mol Biol*, **312(5)**, pp. 927–934.
- Jayasinghe, S., Hristova, K. and White, S. H. (2001b) 'MPtopo: A database of membrane protein topology.', *Protein Science*, **10(2)**, pp. 455–458.
- Jiang, L. and Lai, L. (2002) 'CH??O hydrogen bonds at protein-protein interfaces', *Journal of Biological Chemistry*, **277(40)**, pp. 37732–37740.
- Johansson, M., Nilsson, I. and von Heijne, G. (1993) 'Positively charged amino acids placed next to a signal sequence block protein translocation more efficiently in Escherichia coli than in mammalian microsomes', *MGG Molecular & General Genetics*, **239(1–2)**, pp. 251–256.
- Johnson, E. T. and Parson, W. W. (2002) 'Electrostatic interactions in an integral membrane protein', *Biochemistry*, **41(20)**, pp. 6483–6494.
- Jones, D. T. (2007) 'Improving the accuracy of transmembrane protein topology prediction using evolutionary information', *Bioinformatics*, **23(5)**, pp. 538–544.
- Jones, D. T., Taylor, W. R. and Thornton, J. M. (1994) 'A Model Recognition Approach to the Prediction of All-Helical Membrane Protein Structure and Topology', *Biochemistry*, **33(10)**, pp. 3038–3049.
- Käll, L., Krogh, A. and Sonnhammer, E. L. L. (2004) 'A combined transmembrane topology and signal peptide prediction method', *Journal of Molecular Biology*,

BIBLIOGRAPHY

338(5), pp. 1027–1036.

- Käll, L., Krogh, A. and Sonnhammer, E. L. L. (2005) 'An HMM posterior decoder for sequence feature prediction that includes homology information', *Bioinformatics*, **21(SUPPL. 1)**.
- Killian, J. A. (1998) 'Hydrophobic mismatch between proteins and lipids in membranes', *Biochimica et Biophysica Acta - Reviews on Biomembranes*, **1376(3)**, pp. 401–415.
- Killian, J. A. and von Heijne, G. (2000) 'How proteins adapt to a membrane-water interface', *Trends in Biochemical Sciences*, **25(9)**, pp. 429–434.
- Kim, S., Jeon, T.-J., Oberai, A., Yang, D., Schmidt, J. J. and Bowie, J. U. (2005) 'Transmembrane glycine zippers: physiological and pathological roles in membrane proteins.', *Proceedings of the National Academy of Sciences of the United States of America*, **102(40)**, pp. 14278–83.
- Kleinfeld, A. M., Chu, P. and Romero, C. (1997) 'Transport of long-chain native fatty acids across lipid bilayer membranes indicates that transbilayer flip-flop is rate limiting', *Biochemistry*, **36(46)**, pp. 14146–14158.
- Van Klompenburg, W., Nilsson, I., von Heijne, G. and De Kruijff, B. (1997) 'Anionic phospholipids are determinants of membrane protein topology', *EMBO Journal*, **16(14)**, pp. 4261–4266.
- Kol, M. A., De Kroon, A. I. P. M., Killian, J. A. and De Kruijff, B. (2004) 'Transbilayer Movement of Phospholipids in Biogenic Membranes', *Biochemistry*, pp. 2673–2681.
- Kozma, D., Simon, I. and Tusnady, G. E. (2013) 'PDBTM: Protein Data Bank of transmembrane proteins after 8 years', *Nucleic Acids Research*, **41(D1)**, pp. D524–D529.
- Krishnakumar, S. S. and London, E. (2007) 'Effect of Sequence Hydrophobicity and Bilayer Width upon the Minimum Length Required for the Formation of Transmembrane Helices in Membranes', *Journal of Molecular Biology*, **374(3)**, pp. 671–687.
- Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. L. (2001) 'Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes', *Journal of molecular biology*, **305(3)**, pp. 567–580.
- Kuroiwa, T., Sakaguchi, M., Mihara, K. and Omura, T. (1991) 'Systematic analysis of stop-transfer sequence for microsomal membrane', *Journal of Biological Chemistry*, **266(14)**, pp. 9251–9255.
- Kyte, J. and Doolittle, R. F. (1982) 'A simple method for displaying the hydropathic character of a protein', *Journal of Molecular Biology*, **157(1)**, pp. 105–132.
- Lemmon, M. A., Flanagan, J. M., Hunt, J. F., Adair, B. D., Bormann, B. J., Dempsey, C. E. and Engelman, D. M. (1992) 'Glycophorin A dimerization is driven by specific interactions between transmembrane ??-helices', *Journal of Biological Chemistry*, **267(11)**, pp. 7683–7689.
- Lemmon, M. A., Treutlein, H. R., Adams, P. D., Brünger, A. T. and Engelman, D. M.

BIBLIOGRAPHY

- (1994) 'A dimerization motif for transmembrane α -helices', *Nature Structural Biology*, **1**(3), pp. 157–63.
- Lerch-Bader, M., Lundin, C., Kim, H., Nilsson, I. and von Heijne, G. (2008) 'Contribution of positively charged flanking residues to the insertion of transmembrane helices into the endoplasmic reticulum.', *Proceedings of the National Academy of Sciences of the United States of America*, **105**(11), pp. 4127–4132.
- Li, S. C. and Deber, C. M. (1994) 'A measure of helical propensity for amino acids in membrane environments.', *Nature structural biology*, **1**(8), p. 558.
- Liam J. McGuffin, Bryson, K. and Jones, D. T. (2000) 'The PSIPRED protein structure prediction server', *Bioinformatics (Oxford, England)*, **16**(4), pp. 404–405.
- Liu, Y., Engelman, D. M. and Gerstein, M. (2002) 'Genomic analysis of membrane protein families: abundance and conserved motifs.', *Genome biology*, **3**, p. research0054.
- Lomize, A. L., Pogozheva, I. D., Lomize, M. a. and Mosberg, H. I. (2006) 'Positioning of proteins in membranes: A computational approach', *Protein Science*, **15**(6), pp. 1318–1333.
- Lomize, A. L., Pogozheva, I. D. and Mosberg, H. I. (2011) 'Anisotropic solvent model of the lipid bilayer. 1. Parameterization of long-range electrostatics and first solvation shell effects', *Journal of Chemical Information and Modeling*, **51**(4), pp. 918–929.
- Lomize, M. A., Lomize, A. L., Pogozheva, I. D. and Mosberg, H. I. (2006) 'OPM: Orientations of proteins in membranes database', *Bioinformatics*, **22**(5), pp. 623–625.
- Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. and Lomize, A. L. (2012) 'OPM database and PPM web server: Resources for positioning of proteins in membranes', *Nucleic Acids Research*, **40**(D1), pp. 370–376.
- Lycklama A Nijeholt, J. A., De Keyzer, J., Prabudiansyah, I. and Driessen, A. J. M. (2013) 'Characterization of the supporting role of SecE in protein translocation', *FEBS Letters*, **587**(18), pp. 3083–3088.
- MacCallum, J. L., Bennett, W. F. D. and Tieleman, D. P. (2008) 'Distribution of amino acids in a lipid bilayer from computer simulations.', *Biophysical journal*, **94**(9), pp. 3393–3404.
- Martínez-Gil, L., Pérez-Gil, J. and Mingarro, I. (2008) 'The surfactant peptide KL4 sequence is inserted with a transmembrane orientation into the endoplasmic reticulum membrane.', *Biophysical journal*, **95**(6), pp. L36-8.
- Mary, C., Scherrer, A., Huck, L., Lakkaraju, A. K. K., Thomas, Y., Johnson, A. E. and Strub, K. (2010) 'Residues in SRP9/14 essential for elongation arrest activity of the signal recognition particle define a positively charged functional domain on one side of the protein.', *RNA (New York, N.Y.)*, **16**(5), pp. 969–979.
- McCulloch, W. S. and Pitts, W. H. (1943) 'A logical calculus of ideas imminent in nervous activity', *Bulletin of Mathematics Biophysics*, **5**, pp. 115–133.

BIBLIOGRAPHY

- van Meer, G. (2011) 'Dynamic transbilayer lipid asymmetry', *Cold Spring Harbor Perspectives in Biology*, **3(5)**, pp. 1–11.
- van Meer, G., Voelker, D. R. and Feigenson, G. W. (2008) 'Membrane lipids: where they are and how they behave.', *Nature reviews. Molecular cell biology*, **9(2)**, pp. 112–124.
- Meindl-Beinker, N. M., Lundin, C., Nilsson, I., White, S. H. and von Heijne, G. (2006) 'Asn- and Asp-mediated interactions between transmembrane helices during translocon-mediated membrane protein assembly.', *EMBO reports*, **7(11)**, pp. 1111–1116.
- Ménétrét, J. F., Hegde, R. S., Aguiar, M., Gygi, S. P., Park, E., Rapoport, T. A. and Akey, C. W. (2008) 'Single Copies of Sec61 and TRAP Associate with a Nontranslating Mammalian Ribosome', *Structure*, **16(7)**, pp. 1126–1137.
- Monné, M., Nilsson, I., Johansson, M., Elmhed, N. and von Heijne, G. (1998) 'Positively and negatively charged residues have different effects on the position in the membrane of a model transmembrane helix.', *Journal of molecular biology*, **284(4)**, pp. 1177–83.
- Mottamal, M. and Lazaridis, T. (2005) 'The contribution of C alpha-H...O hydrogen bonds to membrane protein stability depends on the position of the amide.', *Biochemistry*, **44(5)**, pp. 1607–13.
- Nilsson, I. and von Heijne, G. (1990) 'Fine-tuning the topology of a polytopic membrane protein: Role of positively and negatively charged amino acids', *Cell*, **62(6)**, pp. 1135–1141.
- Nilsson, I., Johnson, A. E. and von Heijne, G. (2003) 'How hydrophobic is alanine?', *Journal of Biological Chemistry*, **278(32)**, pp. 29389–29393.
- Nilsson, I., Saaf, A., Whitley, P., Gafvelin, G., Waller, C. and von Heijne, G. (1998) 'Proline-induced disruption of a transmembrane alpha-helix in its natural environment', *J Mol Biol*, **284(4)**, pp. 1165–1175.
- Nilsson, I., Whitley, P. and von Heijne, G. (1994) 'The COOH-terminal ends of internal signal and signal-anchor sequences are positioned differently in the ER translocase', *Journal of Cell Biology*, **126(5)**, pp. 1127–1132.
- Nilsson, J., Persson, B. and von Heijne, G. (2005) 'Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes', *Proteins*, **60(4)**, pp. 606–616.
- Nouwen, N., Berrelkamp, G. and Driessen, A. J. M. (2009) 'Charged Amino Acids in a Preprotein Inhibit SecA-Dependent Protein Translocation', *Journal of Molecular Biology*, **386(4)**, pp. 1000–1010.
- Nyathi, Y., Wilkinson, B. M. and Pool, M. R. (2013) 'Co-translational targeting and translocation of proteins to the endoplasmic reticulum', *Biochimica et Biophysica Acta - Molecular Cell Research*, pp. 2392–2402.
- Oberai, A., Joh, N. H., Pettit, F. K. and Bowie, J. U. (2009) 'Structural imperatives impose diverse evolutionary constraints on helical membrane proteins.', *Proceedings of the National Academy of Sciences of the United States of America*, **106(42)**, pp. 17747–50.

BIBLIOGRAPHY

- Öjemalm, K., Halling, K. K., Nilsson, I. and von Heijne, G. (2012) 'Orientational Preferences of Neighboring Helices Can Drive ER Insertion of a Marginally Hydrophobic Transmembrane Helix', *Molecular Cell*, **45(4)**, pp. 529–540.
- Orzáez, M., Lukovic, D., Abad, C., Pérez-Payá, E. and Mingarro, I. (2005) 'Influence of hydrophobic matching on association of model transmembrane fragments containing a minimised glycoporphin A dimerisation motif', *FEBS Letters*, **579(7)**, pp. 1633–1638.
- Orzáez, M., Salgado, J., Giménez-Giner, A., Pérez-Payá, E. and Mingarro, I. (2004) 'Influence of Proline Residues in Transmembrane Helix Packing', *Journal of Molecular Biology*, **335(2)**, pp. 631–640.
- Overington, J. P., Al-Lazikani, B. and Hopkins, A. L. (2006) 'How many drug targets are there?', *Nature reviews. Drug discovery*, **5(12)**, pp. 993–6.
- Paetzel, M., Dalbey, R. E. and Strynadka, N. C. J. (2002) 'Crystal structure of a bacterial signal peptidase apoenzyme. Implications for signal peptide binding and the Ser-Lys dyad mechanism', *Journal of Biological Chemistry*, **277(11)**, pp. 9512–9519.
- Panzner, S., Dreier, L., Hartmann, E., Kostka, S. and Rapoport, T. A. (1995) 'Posttranslational protein transport in yeast reconstituted with a purified complex of Sec proteins and Kar2p', *Cell*, **81(4)**, pp. 561–570.
- Park, E. and Rapoport, T. a. (2012) 'Mechanisms of Sec61/SecY-Mediated Protein Translocation Across Membranes', *Annual Review of Biophysics*, **41(1)**, pp. 21–40.
- Park, E. and Rapoport, T. A. (2011) 'Preserving the membrane barrier for small molecules during bacterial protein translocation.', *Nature*, **473(7346)**, pp. 239–42.
- Peters, C., Tsirigos, K. D., Shu, N. and Elofsson, A. (2016) 'Improved topology prediction using the terminal hydrophobic helices rule', *Bioinformatics*, **32(8)**, pp. 1158–1162.
- Pfeffer, S., Burbaum, L., Unverdorben, P., Pech, M., Chen, Y., Zimmermann, R., Beckmann, R. and Förster, F. (2015) 'Structure of the native Sec61 protein-conducting channel', *Nature Communications*, **6**, p. 8403.
- Pfeffer, S., Dudek, J., Gogala, M., Schorr, S., Linxweiler, J., Lang, S., Becker, T., Beckmann, R., Zimmermann, R. and Förster, F. (2014) 'Structure of the mammalian oligosaccharyl-transferase complex in the native ER protein translocon.', *Nature communications*, **5**, p. 3072.
- Pfeffer, S., Dudek, J., Schaffer, M., Ng, B. G., Albert, S., Plitzko, J. M., Baumeister, W., Zimmermann, R., Freeze, H. H., Engel, B. D. and Förster, F. (2017) 'Dissecting the molecular organization of the translocon-associated protein complex', *Nature Communications*, **8**, p. 14516.
- Plath, K. and Rapoport, T. A. (2000) 'Spontaneous release of cytosolic proteins from posttranslational substrates before their transport into the endoplasmic reticulum', *Journal of Cell Biology*, **151(1)**, pp. 167–178.
- Popot, J. L. and Engelman, D. M. (1990) 'Membrane protein folding and

BIBLIOGRAPHY

- oligomerization: the two-stage model', *Biochemistry*, **29**(17), pp. 4031–4037.
- Rabiner, L. . and Juang, B. H. (1986) 'An introduction to hidden Markov models.', *IEEE ASSP Magazine*, **3**, pp. 4–16.
- Raetz, C. R. H. and Dowhan, W. (1990) 'Biosynthesis and function of phospholipids in *Escherichia coli*', *Journal of Biological Chemistry*, **265**(3), pp. 1235–1238.
- Rapoport, T. A. (2007) 'Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes.', *Nature*, **450**(7170), pp. 663–9.
- Rapp, M., Granseth, E., Seppälä, S. and von Heijne, G. (2006) 'Identification and evolution of dual-topology membrane proteins', *Nature Structural & Molecular Biology*, **13**(2), pp. 112–116.
- Rapp, M., Seppala, S., Granseth, E. and von Heijne, G. (2007) 'Emulating membrane protein evolution by rational design', *Science*, **315**(5816), pp. 1282–1284.
- Reynolds, S. M., Käll, L., Riffle, M. E., Bilmes, J. A. and Noble, W. S. (2008) 'Transmembrane topology and signal peptide prediction using dynamic Bayesian networks', *PLoS Computational Biology*, **4**(11).
- Ridder, A., Skupjen, P., Unterreitmeier, S. and Langosch, D. (2005) 'Tryptophan supports interaction of transmembrane helices', *Journal of Molecular Biology*, **354**(4), pp. 894–902.
- Rosenblatt, F. (1958) 'The perceptron: A probabilistic model for information storage and organization in the brain.', *Psychological Review*, **65**(6), pp. 386–408.
- Rost, B., Fariselli, P. and Casadio, R. (1996) 'Topology prediction for helical transmembrane proteins at 86% accuracy-Topology prediction at 86% accuracy', *Protein Science*, **5**(8), pp. 1704–1718.
- Rost, B., Sander, C., Casadio, R. and Fariselli, P. (1995) 'Transmembrane helices predicted at 95% accuracy', *Protein Science*, **4**(3), pp. 521–533.
- Russ, W. P. and Engelman, D. M. (2000) 'The GxxxG motif: a framework for transmembrane helix-helix association.', *Journal of molecular biology*, **296**(3), pp. 911–9.
- Sääf, A., Wallin, E. and von Heijne, G. (1998) 'Stop-transfer function of pseudo-random amino acid segments during translocation across prokaryotic and eukaryotic membranes.', *European journal of biochemistry / FEBS*, **251**(3), pp. 821–9.
- Sadlish, H., Pitonzo, D., Johnson, A. E. and Skach, W. R. (2005) 'Sequential triage of transmembrane segments by Sec61alpha during biogenesis of a native multispansing membrane protein', *Nature Structural & Molecular Biology*, **12**(10), pp. 870–878.
- Sahin-Toth, M., Dunten, R. L., Gonzalez, A. and Kaback, H. R. (1992) 'Functional interactions between putative intramembrane charged residues in the lactose permease of *Escherichia coli*.', *Proceedings of the National Academy of Sciences*, **89**(21), pp. 10547–10551.
- Sakaguchi, M., Tomiyoshi, R., Kuroiwa, T., Mihara, K. and Omura, T. (1992) 'Functions of signal and signal-anchor sequences are determined by the balance

BIBLIOGRAPHY

- between the hydrophobic segment and the N-terminal charge', *Proceedings of the National Academy of Sciences*, **89(1)**, pp. 16–19.
- Sal-Man, N., Gerber, D., Bloch, I. and Shai, Y. (2007) 'Specificity in transmembrane helix-helix interactions mediated by aromatic residues', *Journal of Biological Chemistry*, **282(27)**, pp. 19753–19761.
- Sambrook, J. and Russell, D. W. (2001) 'Molecular Cloning: A Laboratory Manual, Third Edition', *Cold Spring Harbour Laboratory Press*, **1–3**.
- Saurí, A., McCormick, P. J., Johnson, A. E. and Mingarro, I. (2007) 'Sec61 α and TRAM are Sequentially Adjacent to a Nascent Viral Membrane Protein during its ER Integration', *Journal of Molecular Biology*, **366(2)**, pp. 366–374.
- Saurí, A., Saksena, S., Salgado, J., Johnson, A. E. and Mingarro, I. (2005) 'Double-spanning plant viral movement protein integration into the endoplasmic reticulum membrane is signal recognition particle-dependent, translocon-mediated, and concerted', *Journal of Biological Chemistry*, **280(27)**, pp. 25907–25912.
- Saurí, A., Tamborero, S., Martínez-Gil, L., Johnson, A. E. and Mingarro, I. (2009) 'Viral Membrane Protein Topology Is Dictated by Multiple Determinants in Its Sequence', *Journal of Molecular Biology*, **387(1)**, pp. 113–128.
- Senes, A., Gerstein, M. and Engelman, D. M. (2000) 'Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with β -branched residues at neighboring positions', *Journal of Molecular Biology*, **296(3)**, pp. 921–936.
- Seppälä, S., Slusky, J. S., Lloris-Garcerá, P., Rapp, M. and von Heijne, G. (2010) 'Control of membrane protein topology by a single C-terminal residue.', *Science (New York, N.Y.)*, **328(5986)**, pp. 1698–1700.
- Sharpe, H. J., Stevens, T. J. and Munro, S. (2010) 'A Comprehensive Comparison of Transmembrane Domains Reveals Organelle-Specific Properties', *Cell*. Elsevier Ltd, **142(1)**, pp. 158–169.
- Singer, S. J. and Nicolson, G. L. (1972) 'The fluid mosaic model of the structure of cell membranes', *Science*, **175(4023)**, pp. 720–31.
- Skach, W. R. (2009) 'Cellular mechanisms of membrane protein folding', *Nature structural & molecular biology*, **16(6)**, pp. 606–612.
- Song, W., Raden, D., Mandon, E. C. and Gilmore, R. (2000) 'Role of Sec61 α in the Regulated Transfer of the Ribosome–Nascent Chain Complex from the Signal Recognition Particle to the Translocation Channel', *Cell*, **100(3)**, pp. 333–343.
- Sonnhammer, E. L., von Heijne, G. and Krogh, A. (1998) 'A hidden Markov model for predicting transmembrane helices in protein sequences', *Proc.Int.Conf.Intell.Syst.Mol.Biol.*, **6**, pp. 175–182.
- Sonnino, S. and Prinetti, A. (2013) 'Membrane domains and the "lipid raft" concept.', *Current medicinal chemistry*, **20(1)**, pp. 4–21.
- Stansfeld, P. J., Goose, J. E., Caffrey, M., Carpenter, E. P., Parker, J. L., Newstead, S. and Sansom, M. S. P. (2015) 'MemProtMD: Automated Insertion of Membrane

BIBLIOGRAPHY

- Protein Structures into Explicit Lipid Membranes', *Structure*. Elsevier, **23(7)**, pp. 1–12.
- Stewart, R. S., Drisaldi, B. and Harris, D. A. (2001) 'A transmembrane form of the prion protein contains an uncleaved signal peptide and is retained in the endoplasmic Reticulum', *Mol Biol Cell*, **12(4)**, pp. 881–889.
- Stone, T. A., Schiller, N., von Heijne, G. and Deber, C. M. (2015) 'Hydrophobic blocks facilitate lipid compatibility and translocon recognition of transmembrane protein sequences', *Biochemistry*, **54(7)**, pp. 1465–1473.
- Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. A., Glass, C. K., Merrill, A. H., Murphy, R. C., Raetz, C. R. H., Russell, D. W. and Subramaniam, S. (2007) 'LMSD: LIPID MAPS structure database', *Nucleic Acids Research*, **35(SUPPL. 1)**.
- Suetsugu, S., Kurisu, S. and Takenawa, T. (2014) 'Dynamic shaping of cellular membranes by phospholipids and membrane-deforming proteins', *Physiological reviews*, **94(4)**, pp. 1219–1248.
- Tamborero, S., Vilar, M., Martínez-Gil, L., Johnson, A. E. and Mingarro, I. (2011) 'Membrane insertion and topology of the translocating chain-associating membrane protein (TRAM)', *Journal of Molecular Biology*, **406(4)**, pp. 571–582.
- Taylor, R. G., Walker, D. C. and McInnes, R. R. (1993) 'E.coli host strains', **21(7)**, pp. 1677–1678.
- Tsirigos, K. D., Peters, C., Shu, N., Kuhl, L. and Elofsson, A. (2015) 'The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides', *Nucleic Acids Research*, **43(W1)**, pp. W401–W407.
- Tusnády, G. E., Dosztányi, Z. and Simon, I. (2004) 'Transmembrane proteins in the Protein Data Bank: Identification and classification', *Bioinformatics*, **20(17)**, pp. 2964–2972.
- Tusnády, G. E., Dosztányi, Z. and Simon, I. (2005a) 'PDB_TM: Selection and membrane localization of transmembrane proteins in the protein data bank', *Nucleic Acids Research*, **33(DATABASE ISS.)**, pp. 275–278.
- Tusnády, G. E., Dosztányi, Z. and Simon, I. (2005b) 'TMDET: Web server for detecting transmembrane regions of proteins by using their 3D coordinates', *Bioinformatics*, **21(7)**, pp. 1276–1277.
- Tusnády, G. E. and Simon, I. (1998) 'Principles governing amino acid composition of integral membrane proteins: application to topology prediction.', *Journal of molecular biology*, **283(2)**, pp. 489–506.
- Ulmschneider, M. B. and Sansom, M. S. P. (2001) 'Amino acid distributions in integral membrane protein structures', *Biochim Biophys Acta*, **1512(1)**, pp. 1–14.
- Ulmschneider, M. B., Sansom, M. S. P. and Di Nola, A. (2005) 'Properties of integral membrane protein structures: Derivation of an implicit membrane potential', *Proteins: Structure, Function and Genetics*, **59(2)**, pp. 252–265.

BIBLIOGRAPHY

- Ulmschneider, M. B., Ulmschneider, J. P., Freites, J. A., von Heijne, G., Tobias, D. J. and White, S. H. (2017) 'Transmembrane helices containing a charged arginine are thermodynamically stable', *European Biophysics Journal*, pp. 1–11.
- Viklund, H., Bernsel, A., Skwark, M. and Elofsson, A. (2008) 'SPOCTOPUS: A combined predictor of signal peptides and membrane protein topology', *Bioinformatics*, **24(24)**, pp. 2928–2929.
- Viklund, H. and Elofsson, A. (2004) 'Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information.', *Protein science: a publication of the Protein Society*, **13(7)**, pp. 1908–1917.
- Viklund, H. and Elofsson, A. (2008) 'OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar', *Bioinformatics*, **24(15)**, pp. 1662–1668.
- Vinothkumar, K. R. and Henderson, R. (2010) *Structures of membrane proteins., Quarterly reviews of biophysics.*
- Wahlberg, J. M. and Spiess, M. (1997) 'Multiple determinants direct the orientation of signal-anchor proteins: The topogenic role of the hydrophobic signal domain', *Journal of Cell Biology*, pp. 555–562.
- Walter, P. and Blobel, G. (1981) 'Translocation of proteins across the endoplasmic reticulum. III. Signal recognition protein (SRP) causes signal sequence-dependent and site-specific arrest of chain elongation that is released by microsomal membranes', *Journal of Cell Biology*, **91(2 I)**, pp. 557–561.
- Weihofen, A., Binns, K., Lemberg, M. K., Ashman, K. and Martoglio, B. (2002) 'Identification of Signal Peptide Peptidase, a Presenilin-Type Aspartic Protease', *Science*, **296(5576)**, pp. 2215–2218.
- White, S. H. (2004) 'The progress of membrane protein structure determination', *Protein Science*, **13(7)**, pp. 1948–1949.
- White, S. H. and von Heijne, G. (2005) 'Do protein-lipid interactions determine the recognition of transmembrane helices at the ER translocon?', *Biochemical Society Transactions*, **33(Pt 5)**, pp. 1012–1015.
- White, S. H. and von Heijne, G. (2008) 'How translocons select transmembrane helices.', *Annual review of biophysics*, **37**, pp. 23–42.
- White, S. H. and Wimley, W. C. (1999) 'Membrane Protein Folding and Stability: Physical Principles', *Ann. Rev. Biophys. Biomol. Struct.*, **28**, pp. 319–365.
- Whitley, P. and Mingarro, I. (2014) 'Stitching proteins into membranes, not sew simple', *Biological Chemistry*, **395(12)**, pp. 1417–1424.
- Williams, R. W., Chang, A., Juretić, D. and Loughran, S. (1987) 'Secondary structure predictions and medium range interactions.', *Biochimica et biophysica acta*, **916(2)**, pp. 200–4.
- Williamson, P. (2015) 'Phospholipid Scramblases.', *Lipid insights*, **8(Suppl 1)**, pp. 41–4.
- Wimley, W. C. (2003) 'The versatile beta-barrel membrane protein.', *Current opinion*

BIBLIOGRAPHY

- in structural biology*, **13(4)**, pp. 404–11.
- Wimley, W. C., Creamer, T. P. and White, S. H. (1996) ‘Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides.’, *Biochemistry*, **35(16)**, pp. 5109–5124.
- Wimley, W. C. and White, S. H. (1996) ‘Experimentally determined hydrophobicity scale for proteins at membrane interfaces’, *Nature Structural Biology*, **3(10)**, pp. 842–848.
- Yamagishi, M., Onishi, Y., Yoshimura, S., Fujita, H., Imai, K., Kida, Y. and Sakaguchi, M. (2014) ‘A few positively charged residues slow movement of a polypeptide chain across the endoplasmic reticulum membrane’, *Biochemistry*, **53(33)**, pp. 5375–5383.
- Yohannan, S., Yang, D., Faham, S., Boulting, G., Whitelegge, J. and Bowie, J. U. (2004) ‘Proline substitutions are not easily accommodated in a membrane protein’, *Journal of Molecular Biology*, **341(1)**, pp. 1–6.
- Zhang, W., Bogdanov, M., Pi, J., Pittard, A. J. and Dowhan, W. (2003) ‘Reversible Topological Organization within a Polytopic Membrane Protein is Governed by a Change in Membrane Phospholipid Composition’, *Journal of Biological Chemistry*, **278(50)**, pp. 50128–50135.
- Zhang, W., Campbell, H. A., King, S. C. and Dowhan, W. (2005) ‘Phospholipids as Determinants of Membrane Protein Topology’, *Journal of Biological Chemistry*, **280(28)**, pp. 26032–26038.
- Zhao, G. and London, E. (2006) ‘An amino acid “transmembrane tendency” scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity.’, *Protein science: a publication of the Protein Society*, **15(8)**, pp. 1987–2001.
- Zhou, F. X., Cocco, M. J., Russ, W. P., Brunger, a T. and Engelman, D. M. (2000) ‘Interhelical hydrogen bonding drives strong interactions in membrane proteins.’, *Nature structural biology*, **7(2)**, pp. 154–160.
- Zhou, X. E., Melcher, K. and XU, H. E. (2012) ‘Structure and activation of rhodopsin’, *Acta Pharmacologica Sinica*. Nature Publishing Group, **33(0)**, pp. 291–299.

ANNEX I

List of oligonucleotides used to generate the constructs and mutants analysed in this work by annealing method and site-directed mutagenesis, respectively.

ID	Sequence (5'-3')
LF19-NP F1	CTAGTGGAGGTCCTGGAGCCATCCTCTGCTGCTGCTGATCGGCGTG
LF19-NP R1	GAGCAGGAACAGGATGATGGCTCCAGGACCTCCA
LF19-NP F2	GGGTCTGGCCGGCCGGCTGGACCTGGAGGGGTAC
LF19-NP R2	CCCTCAGTCCACCGCCGGCCAGACCCACCGCCGATCAGCAC
LF21-NP F1	CTAGTGGAGTCTGAGTGGCATCTCTGGCCCTGATCTCTGGCCGCTGGGG
LF21-NP R1	GATCAGGCCACGAAGATGCCCACTCCAGGACCTCCA
LF21-NP F2	ATCATGCCCTGCTCTGGCGGACCTGGAGGGGTAC
LF21-NP R2	CCCTCAGTCCGCCAGGACGAGGGGATGCCAGCAGCGGCCAGGAA
LF23-NP F1	CTAGTGGAGTCTGGAATCATCTGTGTGATGCCGGGCCCTTGTGGTGGCGGTGTTCCGCC
LF23-NP R1	GAAAGGCCCGGATCACCAAGGATGATTCACCTCCA
LF23-NP F2	GTGTTCTGTGGTCTCTCGGACCTGGAGGGGTAC
LF23-NP R2	CCCTCAGTCCGAAGACAGCAGCACCAAGCAACACCGGGAACACCGCCGACCAC
LW13-NP F	CTAGTGGAGGTCCTGGAATGCTGCAATGCTGCGGACCTTCATCTCCCTGGGATTCGGACCTGGAGGGGTAC
LW13-NP R	CCCTCAGTCCGAAGCCAGGAGATGAAGATGGTGCCACCGTCAAGCATTCCAGGACCTCCA
LW15-NP F	CTAGTGGAGTCTGGAGTTCTGGCTTACGTCCCAATGATCGCTTGGCTGATCGGTGGACCTGGAGGGGTAC
LW15-NP R	CCCTCAGTCCACCGATCAGCAGCAAGCATTTGGACGTAAAGCCAGAAACCTCCAGGACCTCCA
LW17-NP F	CTAGTGGAGTCTGGAGTGTCTTCTGTGGTATCACCGTGCCCTTCTTCTGGCTGGTTCACCATGGACCTGGAGGGGTAC
LW17-NP R	CCCTCAGTCCCATGTTGAACACAGCCAGAAAGCCAGGTGATCCCGAAGGACACTCCAGGACCTCCA
LW19-NP F1	CTAGTGGAGGTCCTGGTGTATGATCTCCGTGGTGGGTGATG
LW19-NP F2	CTGGTGGTGTCTTCGCTGGTACCGGACCTGGAGGGGTAC
LW19-NP R1	CACCAGCAGGGAGATAGCACCCAGGACCTCCA
LW19-NP R2	CCCTCAGTCCGGTACCAGCGAAGAACCCACCAAGCATCACCCAGCC
LW21-NP F1	CTAGTGGAGGTCCTGGTATGACCCCTTTCATGATCATGCTGCTCGCTATGTAC
LW21-NP F2	GCAATGGCTGGTGGACTGGTGGTGGACTGGAGGGGGTAC
LW21-NP R1	CATGATGATCATGAACAGGGTATACCAAGGACCTCCA
LW21-NP R2	CCCTCAGTCCACCCACCAAGTCCACCAAGCCCATGCGTACATAGCGAGCAG
LW23-NP F1	CTAGTGGAGTCTGGAATCGTGTGCTCGCCGGCTGGCCCTGTACTCCGGCATCGTGGCC
LW23-NP R1	CAGGGCCAGCCCGGAGCAGCACCATCCAGGACCTCCA
LW23-NP F2	GTGGTACCATCTTCATGTGGATGGGACCTGGAGGGGTAC
LW23-NP R2	CCCTCAGGTCCTCCATGAAGATGGTGACCAAGGACCGGACCGGACCGGATGCGCCGAGTA

ID	Sequence (5'-3')
LR13-NP F	CTAGTGGAGGTCCTGGACTGATGGCGGTGTTCCCGTGGTCTGGGCAACGCTGATCGGACCTGGAGGGGTAC
LR13-NP R	CCCTCAGGTCGGATACGTTGCCAGGACACCGCGAACACGGCGCATCAGTCCAGGACCTCCA
LR15-NP F	CTAGTGGAGGTCCTGGACTGCCCTCTGTGGCCCTCGGTGATCACCCGGCACCCGCTGGGACCTGGAGGGGTAC
LR15-NP R	CCCTCAGTCCAGCACCGGTGCCGGGTGATCAGGAGGCCACAGAGGGGCGAGTCCAGGACCTCCA
LR17-NP F1	CTAGTGGAGGTCCTGGACGCTGGGGCACAAACCGCTACTTCATGCTG
LR17-NP F2	GCCGTGATCGCGCTCCCTTCGGACCTGGAGGGGTAC
LR17-NP R1	GGTTGCCCCAGGGTCCAGGACCTCCA
LR17-NP R2	CCCTCAGTCCGAGGGAGCGGCGATCACGGCCAGCATGAAGTAGGC
LR19-NP F1	CTAGTGGAGGTCCTGGACCCCTCTCAGGCCCTCTCACCCCTGGTGACCCGTG
LR19-NP F2	CTGGCCTACATGGCCATCTGGGACCTGGAGGGGTAC
LR19-NP R1	GAAGAAGGCCCTGAGCAGGGGACACAGGACCTCCA
LR19-NP R2	CCCTCAGTCCACGATGGCCATGTAGGCCAGCACACGGTCCACCAGGGT
LR21-NP F1	CTAGTGGAGGTCCTGGTCTGTTCAITGGTCTGTCTGGTACCTGGTGGGGAACCTG
LR21-NP F2	ATCACCCCTGCTCATCTACCTGGGACCTGGAGGGGTA
LR21-NP R1	GTACCAGACAGGACCATGAACAGACCAGGACCTCCA
LR21-NP R2	CCCTCAGTCCAGGTAGATGAGCAGGGCGGTGATCAGGTTCCCCACCAG
LR23-NP F1	CTAGTGGAGGTCCTGGTCTGATGTTCCGGTGGCCACCTCCG
LR23-NP F2	CTCCTGTATCATCTGTACGGGTTTCATGCGCCTGACCAACCGGACCTGGAGGGGTAC
LR23-NP R1	CAGGATGATCAGAGGCGGAGGTGGCCACCGGAAACATCAGACCAGGACCTCCA
LR23-NP R2	CCCTCAGTCCGGTGTGTCAGGGCGATGAACCCGTA
LK13-NP F	CTAGTGGAGGTCCTGGAATCTTCTCGAGTGGCTGGTGGCCCTGGCATGGCACCTGGAGGGGTAC
LK13-NP R	CCCTCAGTCCGATGCCATGCCAGGCCACCAGCCACTCGAAAGATTCAGGACCTCCA
LK15-NP F	CTAGTGGAGGTCCTGGAGTATGAGATCGTGTGATCACCACCTGGTCCCATCTGGATCGGACCTGGAGGGGTAC
LK15-NP R	CCCTCAGTCCGGTGTGTCAGGGCGATGAACCCGTA
LK17-NP F1	CTAGTGGAGGTCCTGGACTGTTCTTTTTCGGTGTCCCTGAAAGTGGCC
LK17-NP R1	CAGGCAAGAAGAAGTCCAGGACCTCCA
LK17-NP F2	TGGACCATGTCGTCGCCGGACTGGAGGGGTAC
LK17-NP R2	CCCTCAGTCCGGCACGAACATGTTCCAGGCCAGCTTCAGGGGA
LK19-NP F1	CTAGTGGAGGTCCTGGACCCTTCATCTATCTGGCCCGCTCCCTGTCCATCGGC
LK19-NP R1	GGCGGCCAGATGATGAAGGGTCCAGGACCTCCA

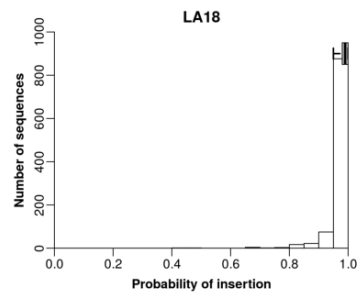
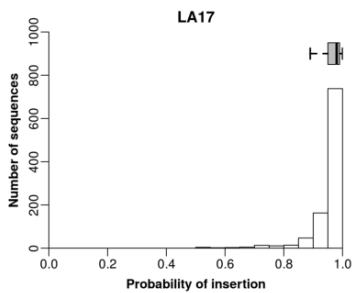
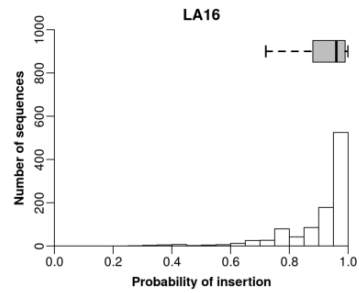
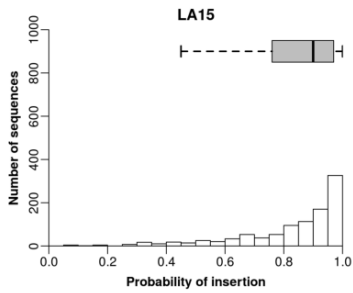
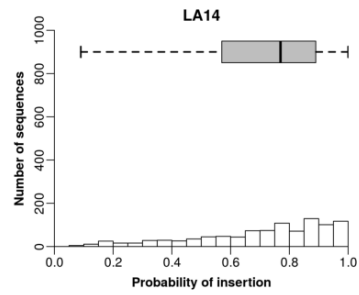
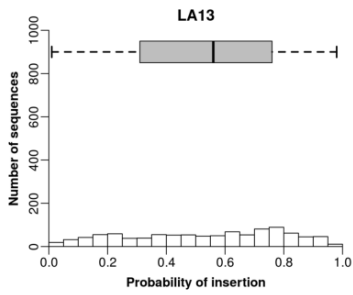
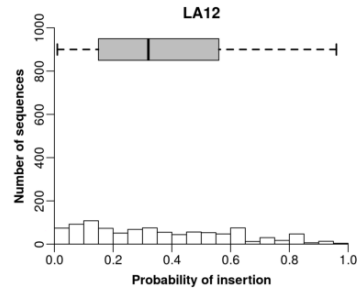
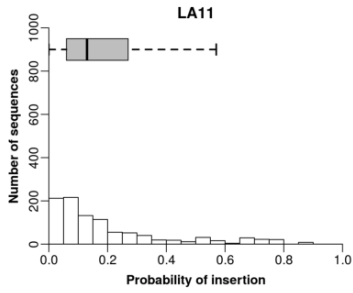
ANNEX I

ID	Sequence (5'-3')
LK19-NP F2	GCCAAGCTGTCTACGCCTGGGACCTGGAGGGGTAC
LK19-NP R2	CCCTCAGGTCCCAGGCGTAGGACAGCTTGGCGCGGATGGACAGGGA
LK21-NP F1	CTAGTGGAGGTCTGGAATGATCTTCTGCTCCTCTGCCATGGCCATCTGCACCCGGCTCC
LK21-NP R1	CATGGGACAGGAGGAGCAGGAAGATCATTCCAGGACCTCCA
LK21-NP F2	CCCTGCTGGTGGCAAGGGACCTGGAGGGGTAC
LK21-NP R2	CCCTCAGGTCCCTTGGCCACCAGCAGAGGGGAGCGGTGCAGGATGGC
LK23-NP F1	CTAGTGGAGGTCTGGATCTTCCACGTGCTGGAGTTTCATCTGGCTGTCCGTGGTTCATGCTG
LK23-NP R1	CCAGATGAACCTCCAGCACGTGGAGAATCCAGGACCTCCA
LK23-NP F2	CCCTCGCCATCAGGSCCTACACGGACCTGGAGGGGTAC
LK23-NP R2	CCCTCAGGTCCGTTGTAGGCTCGATGGCGAAGGCGCATGACCACGGACAG
LK23-NP Mut H3V/V4H F	AGTGGAGTCTGGATTCTTGTCCATCTGGAGTTTCATCTGGCTGTCC
LK23-NP Mut H3V/V4H R	GGACAGCCAGATGAACCTCCAGATGGACGAGAATCCAGGACCTCCACT
LR23-NP Mut A6P/P10A F	CCTGGTCTGATGTTCCGGGTCCCTACCTCCGCTGCCCTGATCCTGTACGGGTTTC
LR23-NP Mut A6P/P10A R	GAACCCGTACAGGATGATCAGGGCAGCGGAGGTAGGCAACCGGAACATCAGACCAGG
LR23-NP Mut P10L/L14P F	TTCCGGTGGCCACCTCCGCTCTGCTGATCCTCCCTACGGGTTTCATGGCCTGACC
LR23-NP Mut P10L/L14P R	GGTCAGGCGCATGAACCCGTTAGGGGATGATCAGCAGAGCGGGAGGTGGCCACCGCGGAA

ANNEX II

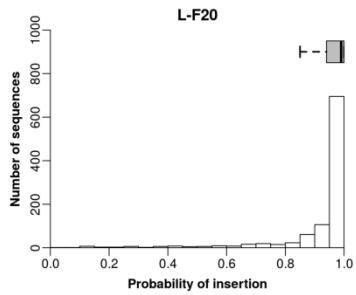
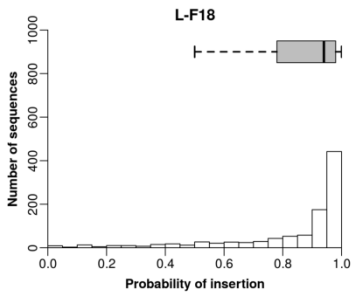
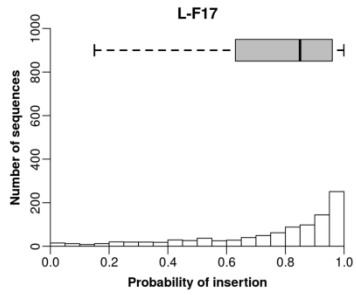
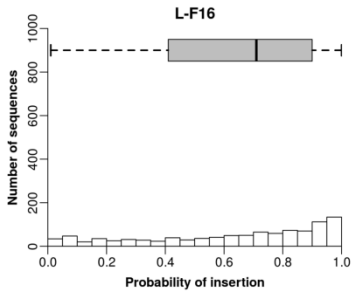
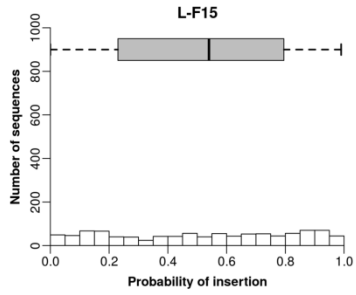
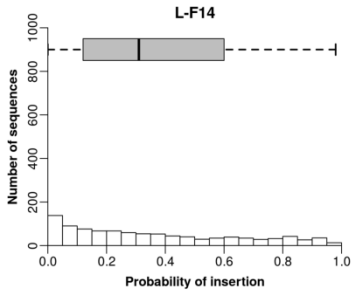
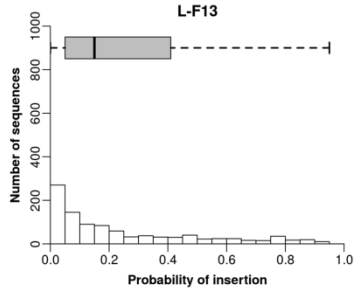
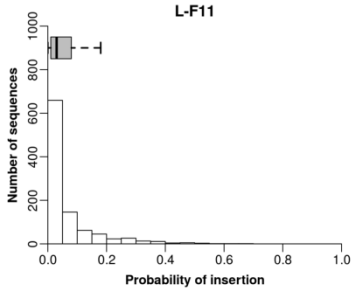
Distribution of predicted p_i values for the 1,000 sequences generated for each length and amino acid composition. Sequences are grouped in 0.05 p_i window range. The gray boxes correspond to the predicted p_i values for the 500 sequences between percentiles 0.25 and 0.75, used in [Figure 29](#) and [Figure 30](#) (gray areas). Dotted lines correspond to the predicted p_i values for the 250 sequences with higher standard deviations on both sides.

L/A-derived sequences

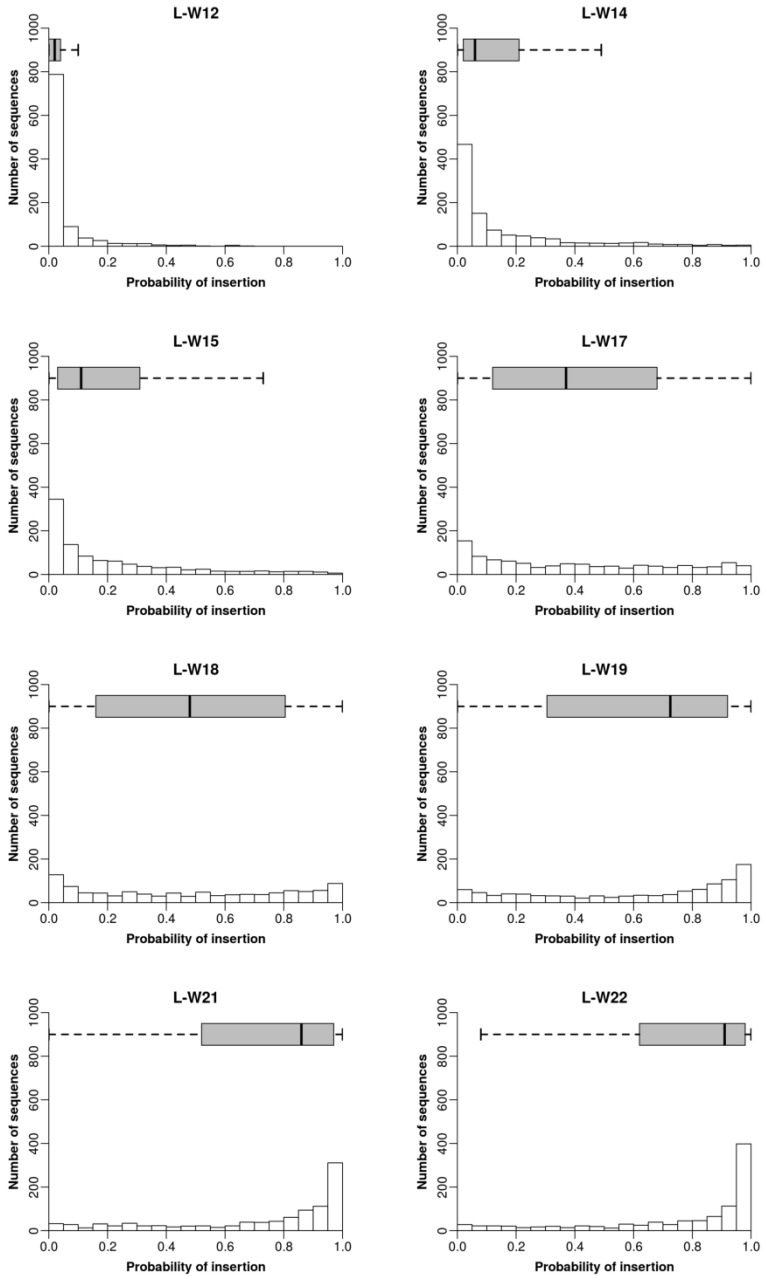


ANNEX II

L/A/V/I/G/F-derived sequences

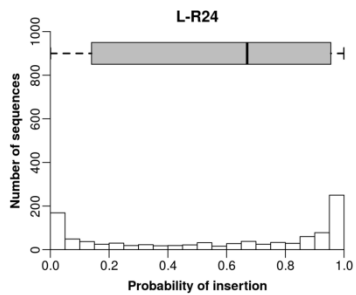
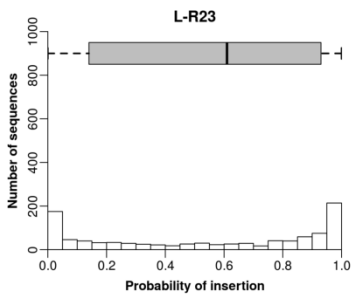
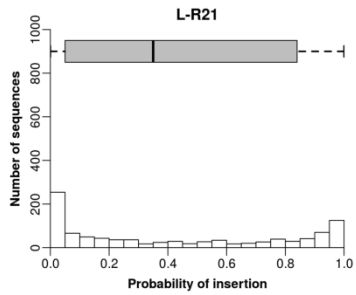
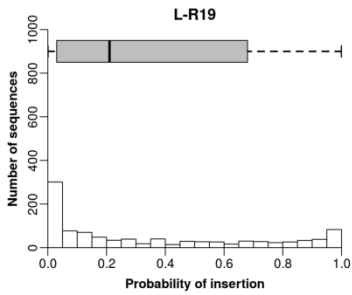
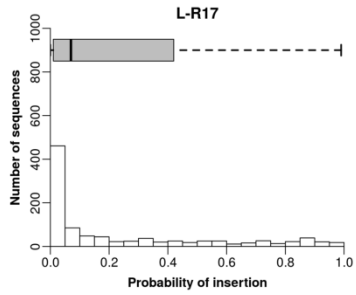
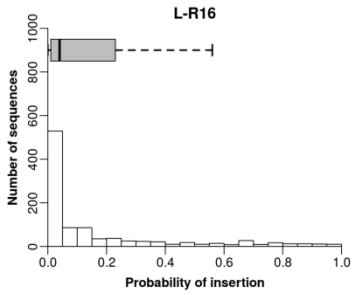
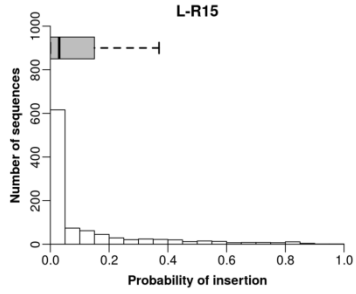
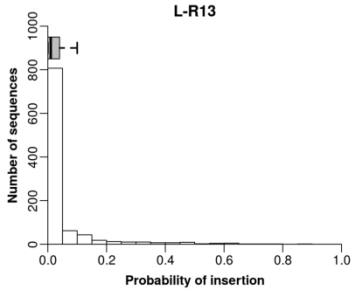


L/A/V/I/G/F/T/S/M/Y/W-derived sequences

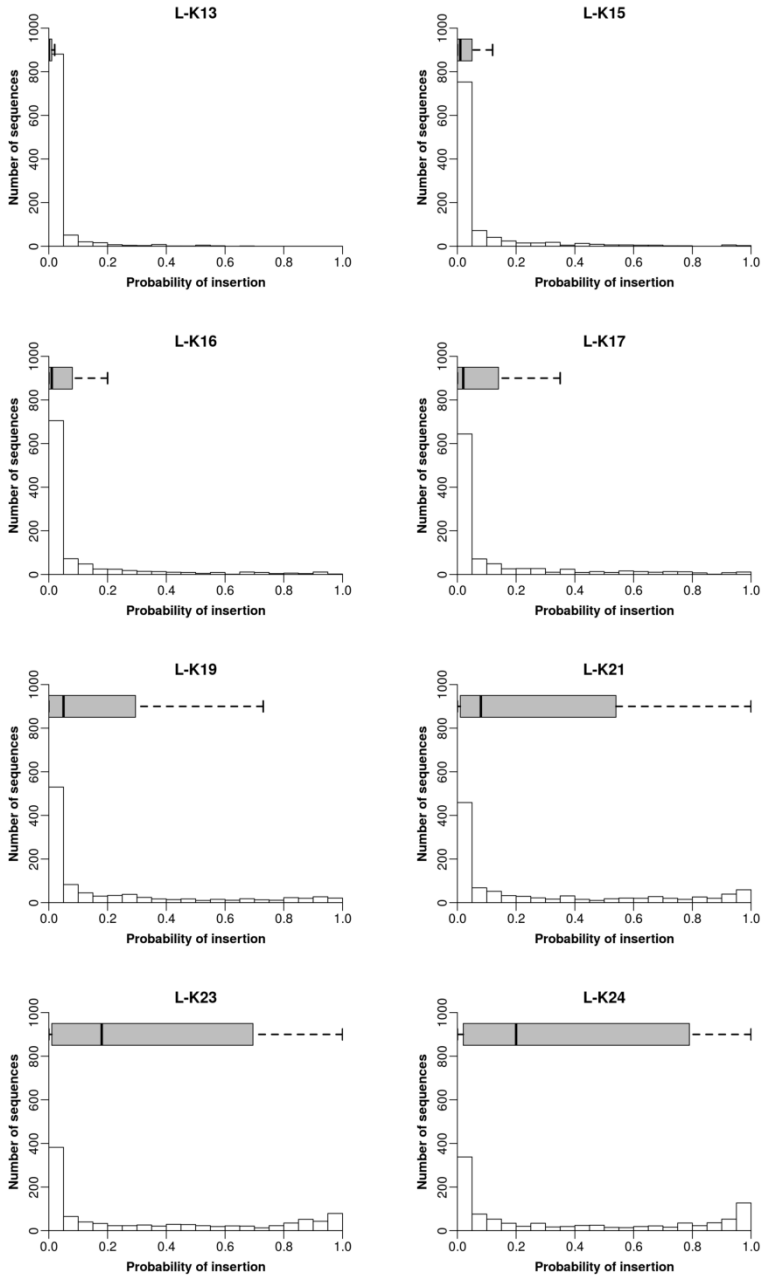


ANNEX II

L/A/V/I/G/F/T/S/M/Y/W/P/N/R-derived sequences

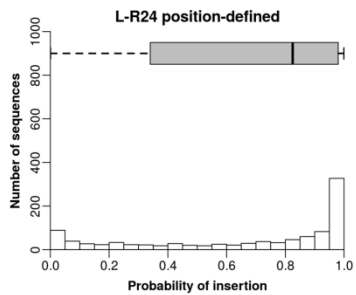
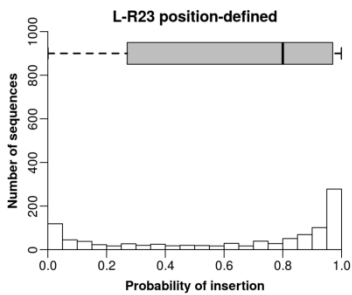
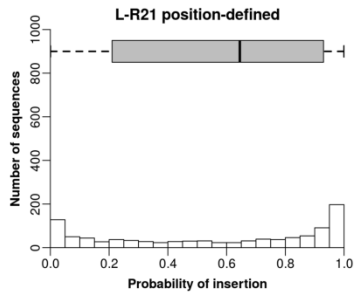
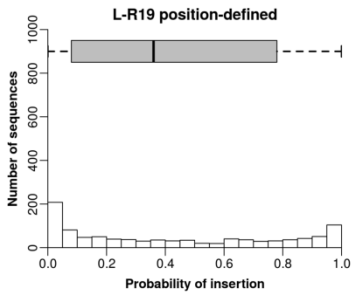
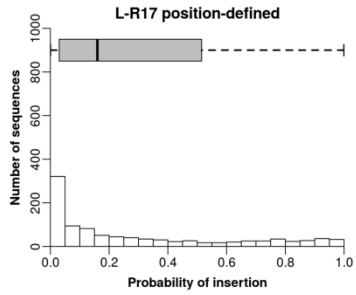
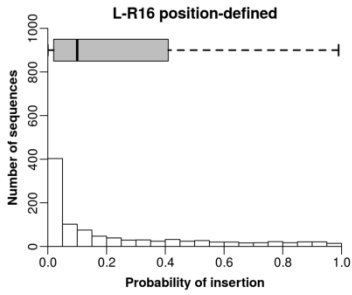
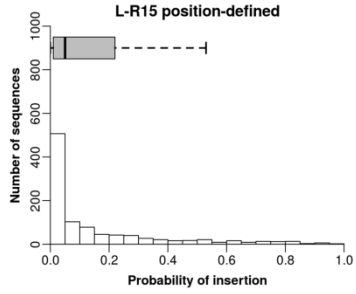
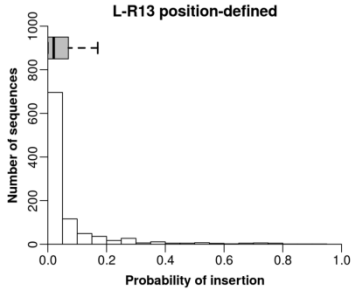


L/A/V/I/G/F/T/S/M/Y/W/P/N/R/H/E/K-derived sequences

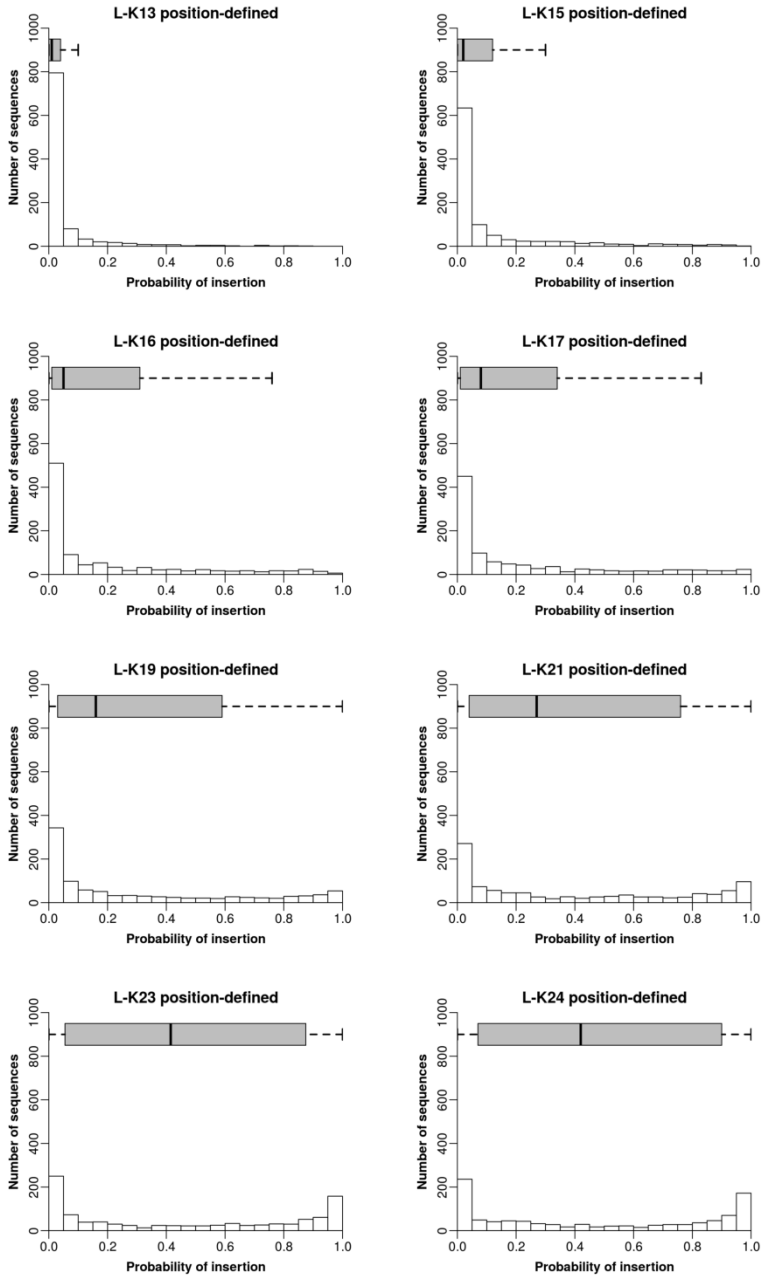


ANNEX II

LR-derived position-defined sequences



LK-derived position-defined sequences



ANNEX III

Baeza-Delgado, C., Marti-Renom, M. A. and Mingarro, I. (2013)
‘Structure-based statistical analysis of transmembrane helices’,
European Biophysics Journal, 42(2–3), pp. 199–207.

All the results of this publication are contained within this Thesis.
Carlos Baeza has performed the experiments and contributed to their
design and to the writing of the manuscript.

Article reproduced with permission of *Springer International
Publishing AG* ©.



Structure-based statistical analysis of transmembrane helices

Carlos Baeza-Delgado · Marc A. Martí-Renom ·
Ismael Mingarro

Received: 16 January 2012/Revised: 21 March 2012/Accepted: 14 April 2012
© European Biophysical Societies' Association 2012

Abstract Recent advances in determination of the high-resolution structure of membrane proteins now enable analysis of the main features of amino acids in transmembrane (TM) segments in comparison with amino acids in water-soluble helices. In this work, we conducted a large-scale analysis of the prevalent locations of amino acids by using a data set of 170 structures of integral membrane proteins obtained from the MPTopo database and 930 structures of water-soluble helical proteins obtained from the protein data bank. Large hydrophobic amino acids (Leu, Val, Ile, and Phe) plus Gly were clearly prevalent in TM helices whereas polar amino acids (Glu, Lys, Asp, Arg, and Gln) were less frequent in this type of helix. The distribution of amino acids along TM helices was also examined. As expected, hydrophobic and slightly polar amino acids are commonly found in the hydrophobic core of the membrane whereas aromatic (Trp and Tyr), Pro, and the hydrophilic amino acids (Asn, His, and Gln) occur more frequently in the interface regions. Charged amino

acids are also statistically prevalent outside the hydrophobic core of the membrane, and whereas acidic amino acids are frequently found at both cytoplasmic and extra-cytoplasmic interfaces, basic amino acids cluster at the cytoplasmic interface. These results strongly support the experimentally demonstrated biased distribution of positively charged amino acids (that is, the so-called the positive-inside rule) with structural data.

Keywords Membrane protein · Transmembrane helices · Amino acid distribution · Statistical analysis

Introduction

Although helical membrane proteins constitute approximately one quarter of all proteins in living organisms (Wallin and von Heijne 1998), the rules governing their folding are still not completely established. The hydrophobic effect is a dominant force driving folding of water-soluble proteins, but its contribution to the folding of membrane proteins is more complex, given that these proteins “live” in a biophysical environment—the membrane—which is clearly different from aqueous media. The cell membrane is a very heterogeneous medium, composed mainly of phospholipids that are self-organized into two leaflets giving rise to the formation of a bilayer. The hydrocarbon core, of dimension approximately 30 Å, is the hydrophobic part of the membrane. The polar head groups of the phospholipids define the lipid/water interface and add approximately 15 Å to the thickness of each leaflet (White and Wimley 1999). It is in this complex environment that membrane proteins must fold into their native conformations.

The hydrocarbon core of biological membranes and the interior of folded water-soluble proteins are hydrophobic.

Special issue: Structure, function, folding and assembly of membrane proteins—Insight from Biophysics.

C. Baeza-Delgado · I. Mingarro (✉)
Departament de Bioquímica i Biologia Molecular,
Universitat de València, Burjassot, Spain
e-mail: Ismael.Mingarro@uv.es

M. A. Martí-Renom
Structural Genomics Team, Genome Biology Group,
National Center for Genomic Analysis (CNAG),
Barcelona, Spain

M. A. Martí-Renom (✉)
Structural Genomics Group, Center for Genomic Regulation
(CRG), Barcelona, Spain
e-mail: mmarti@cpb.ub.cat

Published online: 16 May 2012

Springer

In such a hydrophobic environment, the polarity of the polypeptide backbone is energetically unfavorable. Thus, in protein structures, nearly all the polar groups of the peptide bond (carbonyl and amide groups) tend to hydrogen bond with one another, leading to secondary structure that stabilizes the folded state. Alpha-helices are the commonest secondary structures found in water-soluble and membrane protein structures. However, the distribution of the helices in these two groups of proteins is very different. Whereas helices in water-soluble proteins can be exposed to both the hydrophobic core and the water-accessible surface, transmembrane (TM) helices in membrane proteins are surrounded by a hydrophobic lipid phase in which water is essentially absent. Therefore, for structural stabilization of helical membrane proteins that reside in this apolar (low dielectric) environment, hydrogen bonding and van der Waals packing forces are highly important.

Although the vast majority of membrane proteins integrate into biological membranes through the translocon (recently reviewed by Martínez-Gil et al. 2011), our current biophysical understanding of its folding and function is hampered by the scarcity of structural information. Fortunately, the number of high-resolution structures of membrane proteins has increased exponentially in recent years (White 2004, 2009). Consequently, a new statistical survey of the properties of TM helices is timely.

In this paper, we revisit the differences between helices from water-soluble proteins and TM helices in terms of length and amino acid composition. In addition, we analyze the distribution of amino acids in TM segments, which are energetically accommodated in the highly heterogeneous media of biological membranes as a result of favorable interaction with the local environment. This study involved 170 helical membrane proteins with known three-dimensional structure and topology, containing a total of 792 TM segments, which were compared with 7,348 helices from 930 water-soluble protein structures. Approximately half of all amino acids are randomly distributed when allocated to the membrane, but the others correlate strongly with amino acid positions along the TM regions.

Methods

Helix data sets

Two data sets for water-soluble and TM helices were obtained from the protein data bank (PDB) (Berman et al. 2000) and the MPtopo database (Jayasinghe et al. 2001b), respectively.

First, a total of 4,405 structural chains deposited in the PDB (as of November 17th, 2011) that passed the following criteria were selected:

- 1 their total secondary structure had more than 60 % α -helices and no β -strands;
- 2 their crystallographic resolution was 2.0 Å or higher; and
- 3 the word *MEMBRANE* did not appear in either the "TITLE" or "DESCRIPTION" fields of the PDB file.

Furthermore, to remove redundancy, the 4,405 chain sequences were compared with each other by use of *cd-hit* software (Huang et al. 2010) and pairs resulting in sequence alignments with 80 % or higher identity were discarded. The final set of 930 non-redundant PDB chains was parsed to identify a total of 7,348 helices from the "HELIX" fields of each PDB chain entry. Thus, the data set of water-soluble helices contained 930 non-redundant and high-resolution protein structures, 7,348 α -helices, and 108,277 amino acids.

Second, all α -helical membrane proteins deposited in the MPtopo database (last updated on January 19th, 2010) (Jayasinghe et al. 2001b), and thus with known membrane insertion topology, were selected. The initial set was filtered by:

- 1 removing any entry of unknown structure as based on the MPtopo entry classification (i.e., keeping only entries described as "3D_helix" and "1D_helix"); and
- 2 removing redundant pairs at 80 % sequence identity by use of *cd-hit* software (Huang et al. 2010).

The final data set of TM helices contained 170 non-redundant structures, 837 TM helices, and 20,079 amino acids. Furthermore, to properly analyze the prevalent locations of amino acids in single membrane-spanning TM helices, we discarded any helix shorter than 17 amino acids or larger than 38 amino acids. The resulting TM data subset contained 792 TM helices, and 19,356 amino acids.

Measurement of the prevalent locations of amino acids

We calculated three different measures:

- 1 probability and percent,
- 2 Odds, and
- 3 LogOdds.

The probability (p_i) of an amino acid i is defined as:

$$P_i = \frac{n_i}{N}$$

where i is the amino acid type (one of the 20 amino acids), n_i is the observation count of amino acid i , and N is all amino acids in the data set. Similarly, the percentage of a given amino acid i is defined as its probability multiplied by 100. The Odds (O_i) of an amino acid i is defined as:

$$O_i = \frac{p_{i,c}}{(1 - p_{i,c})} \bigg/ \frac{p_{i,r}}{(1 - p_{i,r})}$$

where $p_{i,c}$ is the probability of amino acid i in class c (for example, TM helix) and $p_{i,r}$ is the probability of the amino acid i in class r (for example, water-soluble helix).

Similarly, the LogOdds for a given amino acid i is defined as the logarithm to the base 10 of its Odds. Briefly, Odds higher than 1 (or positive LogOdds) indicate over-occurrence of the amino acid type in the class. Odds smaller than 1 (or negative LogOdds) indicate under-representation of the amino acid type in the class.

Results and discussion

Helix length in membrane and water-soluble proteins

Length distributions for helices found in high-resolution structures deposited in the PDB (Berman et al. 2000) are very different for TM and water-soluble proteins (Fig. 1).

Helices in TM proteins are, on average, 24.0 (± 5.6) amino acids long; this result differs slightly from previous data obtained by using databases with 45 (Bowie 1997) and 129 (Ulmschneider and Sansom 2001) TM helices, for which average helix length was 26.4 and 27.1 amino acids, respectively. Because the translation per amino acid in a canonical helix is 1.5 Å, a stretch of approximately 20

consecutive hydrophobic amino acids can span the 30 Å of the hydrocarbon core of a biological membrane. Indeed, the most prevalent ($\sim 12\%$) length of TM helices in our data set was 21 amino acids (Fig. 1). Longer helices can span the bilayer with concomitant tilting of the helix axis relative to the membrane plane. Other options are also feasible, ranging from lipid accommodation to polypeptide backbone deformation (Holt and Killian 2009).

Helices from water-soluble proteins have an average length of 14.7 (± 8.7) amino acids, which agrees with previous studies in which the most prevalent helix length was 10–11 amino acids (Engel and DeGrado 2004; Pal et al. 2003). The shorter length of helices in water-soluble proteins is because of the absence of the restrictions imposed by the low dielectric constant at the hydrocarbon core of biological membranes, which forces the polypeptide backbone to adopt, on average, larger secondary structures.

Amino acid composition of α -helices

Amino acid composition was examined for both TM and water-soluble helices (Fig. 2). TM helices of lengths

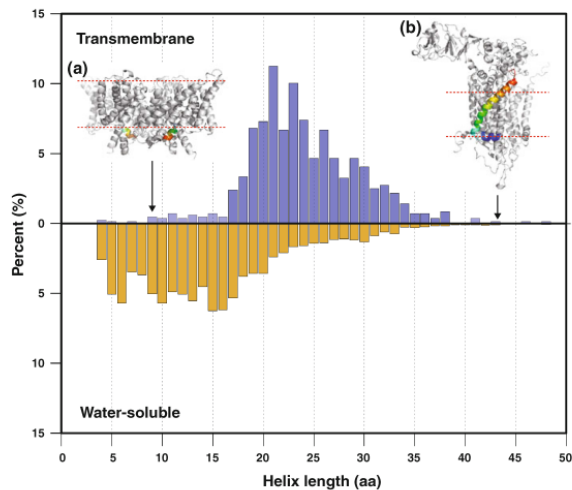


Fig. 1 Length distributions for 837 TM and 7,348 water-soluble helices from a set of non-redundant proteins of known structure (see the “Methods” section). Transmembrane helices are shown in blue (pale blue corresponds to discarded lengths) and water-soluble helices are shown in orange. **a** Example of a short nine-amino-acid-length helix in the CIC chloride channel from *E. coli* (1KPK entry in PDB).

Membrane boundaries were obtained from the PPM server (Lomize et al. 2012). The selected membrane is shown in rainbow coloring from the N-terminal (blue) end to the C-terminal (red) end. **b** Example of a large 43-amino-acid-length helix in the chicken cytochrome BCl complex (1BCC entry in the PDB); the N-terminus of the helix (blue) lies at the membrane/water interface. Representation as in inset **a**

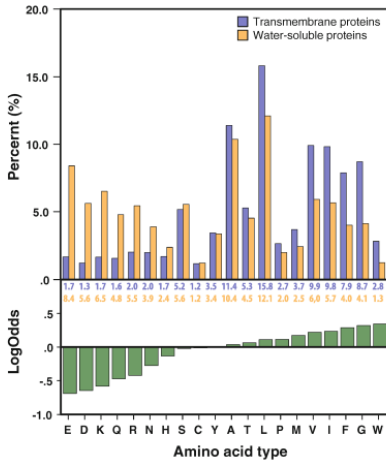


Fig. 2 Amino acid type distribution from 792 TM and 7,348 water-soluble helices from a set of non-redundant proteins of known structure (see the “Methods” section). (Upper plot) Amino acid type distribution for TM helices in blue and for water-soluble helices in orange. (Lower plot) LogOdds values for comparison of the relative abundance of each amino acid type in TM and water-soluble helices. Amino acid types are ordered by LogOdds values

between 17 and 38 amino acids were selected from the MPtopo database (Jayasinghe et al. 2001b); these included helical segments that do completely span the hydrophobic core of the membrane. TM helices shorter than 17 amino acids and larger than 38 amino acids were excluded, because they may not cross the membrane entirely (Fig. 1, inset a) or may contain segments parallel to the membrane (Fig. 1, inset b). Note that for water-soluble helices all lengths were included in our analysis because no restrictions in terms of length can be assumed for water-soluble proteins in an aqueous environment.

As expected, hydrophobic amino acids Leu, Ala, Val, and Ile constitute the bulk of the amino acids in the TM region accounting for almost half (47.0 %) of all amino acids. Similarly, these amino acids are also frequently found in the helices of water-soluble proteins (34.1 %). However, there are, as noted previously using smaller datasets (Bywater et al. 2001), differences between the composition of the two types of helix. Despite sharing the same structural features, the differences between the two types of helix are reflected by their preferential occurrences, as measured by the logarithm of the Odds of finding a given amino acid in a TM helix compared with its

frequency in a water-soluble helix (Fig. 2 bottom panel). For example, whereas charged and polar amino acids are much more frequently found in helices of water-soluble proteins, Trp, Gly, and Phe are more likely to occur in TM helices. Interestingly, in contrast with their prevalent conformations in water, the likelihood of amino acids such as Val, Ile, Phe, and Met occurring in a helical structure are notably increased in the membrane environment, and it has been suggested that their prevalence in helices depends primarily on their side chain hydrophobicity and on the hydrophobicity of the local polypeptide region in which the amino acids reside spanning the membrane (Li and Deber 1994). Significantly, Gly and Pro are more frequent in TM helices than in water-soluble helices. Although commonly regarded as “helix breakers” it has been reported that Gly occurs frequently in TM helix-helix interactions, especially in association with β -branched residues at neighboring positions (Senes et al. 2000), and that Pro, in addition to its function in signal transduction and gating across the membrane, may also be significantly involved in these processes (Orzáez et al. 2004).

Comparison of amino acid frequency between TM and water-soluble helices confirmed that strongly polar amino acids (Glu, Lys, Asp, Arg, and Gln) are more prevalent in water-soluble helices (Fig. 3). These amino acids constitute only 8.2 % of the amino acids within TM helices compared with 30.9 % of those in water-soluble helices. Despite their lower occurrence, polar amino acids are evolutionary

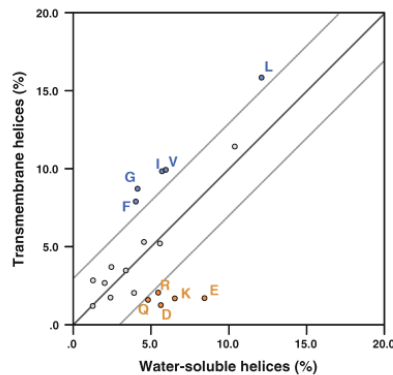


Fig. 3 Amino acid type percentage comparison between TM and water-soluble helices. Blue colored amino acids are over-represented (difference >3 % points) in TM helices compared with water-soluble helices. Orange colored amino acids are over-represented (difference >3 % points) in water-soluble helices compared with TM helices. Dashed grey lines indicate a cut-off of 3 % difference points

conserved in TM proteins, which has been partially explained by their tendency to be buried in the protein interior and, in many cases, because of their direct involvement in the function of the protein (Illergård et al. 2011). Conversely, hydrophobic amino acids (Leu, Val, Ile, Gly, and Phe) are over-represented in TM helices (Fig. 3). Interestingly, Ala, although the second most abundant amino acid in TM helices (Fig. 2), it is not over-represented in this type of helix; this is probably because its greater tendency to participate in a helical structure in aqueous environments (Blaber et al. 1993) than in membrane-mimetic environments (Li and Deber 1994). In fact, both biological (Nilsson et al. 2003; Hessa et al. 2005) and biophysical (Jayasinghe et al. 2001a) measurements have placed Ala at the threshold between those amino acids that promote membrane integration of TM helices and those that preclude membrane insertion.

Position-dependent distribution of amino acids in TM helices

Comparison of amino acid frequency at different positions in a TM segment, taking as reference the TM center, confirmed that approximately half of the natural amino acids have similar distributions in positive positions (toward the inside of the cell) than at negative positions (toward the outside of the cell) (Fig. 4). It was found that not only the strongly hydrophobic amino acids but also Gly and the hydroxylated amino acids Ser and Thr are equally distributed along the hydrophobic core of the membrane. It is important to note that Gly is normally regarded as being conducive to turn (Williams et al. 1987), yet it is a common amino acid in TM helices (Fig. 2). There are important folding reasons for incorporating Gly into TM helices. The absence of a side-chain from Gly enables bulkier groups to

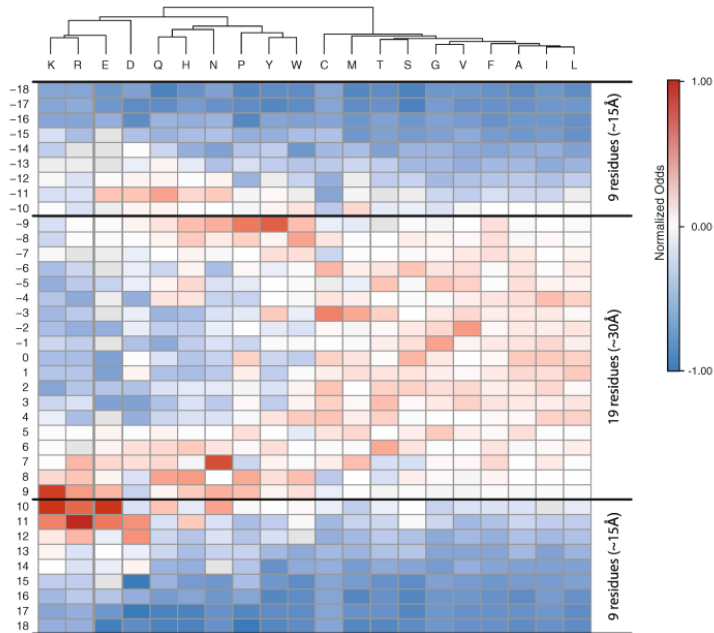


Fig. 4 Amino acid type and position distribution in TM helices. Each amino acid type and its positioning in the TM helix is represented by its position-normalized Odds (that is, for each column the Odds are normalized to an average of zero and a standard deviation of unity). The amino acids are clustered on the basis of their positional

normalized Odds within the helices. Positively labeled positions indicate the cytoplasmic side of the membrane and its flanking region whereas negatively labeled positions are indicative of extra-cytoplasmic regions

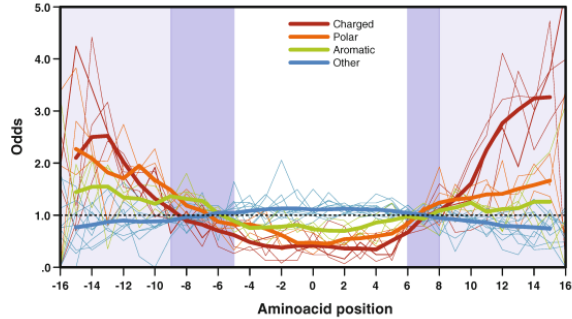
be accommodated close to the polypeptide backbone of the TM helices. This might be important for intramolecular helix-helix packing, for homo-oligomerization, or for recognition of other membrane proteins, among other factors. Indeed, it has been observed that Gly has the highest overall packing value in membrane proteins (Eilers et al. 2002). Ser or Thr within TM helices participate in hydrogen-bonding networks by hydrogen bonding of the side chain oxygen atom to the acceptor side chain or peptide bond groups. These effects, intimate packing (Gly) and hydrogen bonding (Ser and Thr), can be relevant at any position along the TM region, which could explain the absence of position prevalence of these amino acids in TM helices. Met or Cys are also frequent at different locations within the hydrophobic core, but relative prevalence can be observed in a region that would correspond to the initial portion of the polar headgroups of the phospholipids, consistent with the slightly amphipathic nature of these amino acids and in agreement with its distribution in the lipid bilayer recently obtained from molecular dynamics simulation (MacCallum et al. 2008).

Whereas Phe has a flat distribution in TM helices, behaving as a hydrophobic amino acid, distribution of Trp, Tyr, and Pro is biased—they are most likely to be found at the ends of the bilayer (i.e. at the interface between the hydrophobic core of the bilayer and the bulk water). At this location, aromatic amino acids may serve as anchors for the TM helices in the membrane. In fact, Trp and Tyr positioned 7–9 amino acids away from the center of a TM segment result in reduction of the free energy (Hessa et al. 2007), which correlates well with our statistical distribution from three-dimensional structures (Fig. 4). The biophysical reason for the observed distribution of Trp and Tyr could rely on the relatively amphipathic nature of their side chains, which can form hydrogen bonds and also have hydrophobic character. Actually, this prevalent location has previously been observed not only for α -helical but also β -barrel membrane proteins (Ulmschneider and Sansom 2001). A similar distribution is observed for Pro, although increased prevalence is detectable toward the center of the bilayer, which might be associated with the fundamental and subtle function of Pro in the dynamics, structure, and function of many membrane proteins of inducing the formation of molecular hinges (Cordes et al. 2002). Indeed, thirteen TM helices with known structure have Pro at the 0 position, which in all cases results in a kink in the helix. Nevertheless, it should be noted that the interfacial prevalence of these three amino acids is somehow more pronounced at the non-cytoplasmic interface. This was also observed for the aromatic amino acids (Trp and Tyr) in a membrane protein prediction analysis using sequence information from 107 genomes (Nilsson et al. 2005).

The distribution pattern for Asn, His, and Gln, corresponds to an interfacial preference close to the end of the TM regions, which is consistent with the amphipathic nature of these molecules. This pattern was previously reported for His (Ulmschneider and Sansom 2001), and is in good agreement with our results. Interestingly, in more recent studies using computer simulations, it has been noted that small molecule analogs of Asn (MacCallum et al. 2008) and Asn, His, and Gln (Johansson and Lindahl 2007) result in an energy minimum for partition into model lipid bilayers.

Because the energy cost of inserting an ionizable group in the hydrophobic environment of the membrane is very high (White and Wimley 1999), charged amino acids should generally be excluded from the hydrophobic core of the TM helices. Interestingly, nearly all membrane proteins with six or more predicted TM helices contain at least one ionizable amino acid (Arkin and Brunger 1998). However, charged amino acids consistently cluster at the TM flanking regions (Fig. 4). For example, increased distribution of acidic amino acids (Asp and Glu) occurs on both the cytoplasmic and extra-cytoplasmic sides of the membrane, although with some prevalence for the cytoplasmic region. Distribution of positively charged amino acids (Arg and Lys) is even more strongly asymmetric between opposite sides of the membrane, in good agreement with the positive-inside rule (von Heijne 1992). Moreover, it has been demonstrated experimentally that basic amino acids act as stronger topological signals than acidic amino acids (Nilsson and von Heijne 1990; Sauri et al. 2009), which is reflected by their different statistical occurrence on either end of the TM segments. Nevertheless, when considered globally, charged amino acids cluster predominantly near the cytoplasmic end of the TM segments (Fig. 5, orange line). This effect has already been noted in a previous structure-based analysis that included the fewer structures available at the time (Ulmschneider et al. 2005). In contrast, although polar amino acids (Gln, His, and Asn) mimic the distribution pattern of charged amino acids, avoiding the more hydrophobic region of the bilayer, they tend to occur in the extra-cytoplasmic region (Fig. 5). Trp, Tyr, and Pro are more abundant approximately eight or nine amino acid positions from the center of the membrane, that is, within the interface region, but with some bias toward the extra-cytoplasmic interface. The other natural amino acids are more abundant at the center of the bilayer, within seven amino acid positions on both sides of the membrane normal, but are also very frequently found beyond this boundary, as noted by their overall proximity to the Odd value of 1 for positions >10 on both sides of the center of the membrane (Fig. 5). Interestingly, the amino acid distribution patterns in both interface regions are slightly different. There is a sharper transition from mainly

Fig. 5 Most likely positions of amino acid groups in a membrane. *Thin lines* represent the positional Odds for each amino acid individually, whereas *thick lines* represent the average positional Odds for each group of amino acids obtained from Fig. 4. Amino acid types are grouped as in the dendrogram in Fig. 4, i.e. charged amino acids (*red* KRED), polar amino acids (*orange* QHN), aromatic amino acids plus Pro (*green* PYW), and the other amino acids (*blue* CMTSGVFAIL)



hydrophobic to charged, polar, and aromatic amino acids on the cytoplasmic side of the membrane (positions 6–8) than on the extra-cytoplasmic side (positions –5 to –9). The different lipid composition between the two lipid leaflets in biological membranes and the strong electrochemical potential over the prokaryotic inner cell membranes can exert an important effect, which may be reflected by this difference. For instance, asymmetry in the distribution of amino acids within TM segments from plasma membrane proteins has recently been reported (Sharpe et al. 2010), and has been attributed to asymmetry in the state of lipid order in the membrane. Such asymmetry is likely to be because of enrichment of lipids, for example sterols and sphingolipids, in the extra-cytoplasmic leaflet, where more gradual amino acid distribution can be expected.

Finally, we analyzed and plotted the odds ratio for each amino acid in three regions in a membrane, that is, taking the hydrophobic TM region as the central 19 positions (~30 Å) and nine amino acid positions (~15 Å) on both sides as the extra-cytoplasmic (from –10 to –18 amino acids) and cytoplasmic (from 10 to 18) flanking regions (Fig. 6). Hydrophobic amino acids (blue colored) predominated in the hydrophobic center. However, this trend is not observed for the more prevalent amino acids in TM segments (for example Leu, Fig. 2), which are also frequently found in the flanking regions. A minor increase is observed for Trp, Tyr, and Pro (green) in the extra-cytoplasmic flanking region. The absence of larger differences for the distribution of these amino acids is probably because of their precise location at the interface between the hydrophobic core and the flanking hydrophilic environment. Polar (orange) amino acids (Gln, His, and Asn) predominate in both flanking regions, because their presence within the membrane core is energetically unfavorable. These amino acids do not ionize at physiological pH

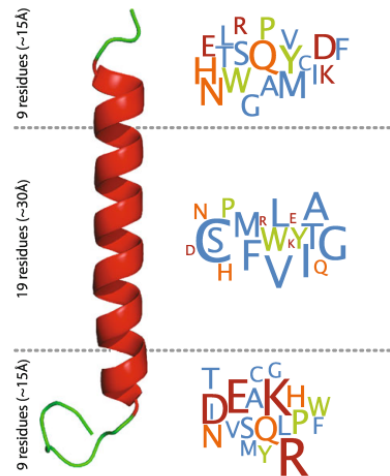


Fig. 6 Amino acid location prevalence in a membrane. Letter size is proportional to the odds (relative prevalence) of finding a given amino acid in the three regions in a membrane (i.e., from top to bottom outer, membrane, and inner regions). Amino acids are colored as in Fig. 5

and can donate and accept hydrogen bonds simultaneously. This effect is manifested as greater occurrence of Gln, His, and Asn in the rich hydrogen-bond network environment of the interface. Charged amino acids (red) were under-represented in the hydrophobic core and tended to occur in the cytoplasmic flanking region, with acidic amino acids more prevalent in the extra-cytoplasmic flanking region. Furthermore, basic amino acids are strong topological

determinants that heavily populate the cytoplasmic flanking region. The effect of positively charged amino acids located near the cytoplasmic end of hydrophobic segments has been estimated to contribute approximately -0.5 kcal/mol to the apparent free energy of membrane insertion (Lerch-Bader et al. 2008). This energy contribution can be extremely relevant for precise anchoring of hydrophobic regions to biological membranes.

Concluding remarks

We have compared the length and amino acid composition of helices in TM and water-soluble proteins. Overall, significant differences are observed for these proteins; these may be attributed to the biophysical differences between the two environments in which they fold.

- First, TM helices adapt their length to the dimensions and constraints of biological membranes, whereas water-soluble helices are statistically shorter because they do not have to satisfy the demanding restrictions imposed by the complexity of the membrane environment.
- Second, the observed differences indicate that in the lipid bilayer, an environment which forces secondary structure formation, amino acid side chain hydrophobicity prevails over helicity. Accordingly, aliphatic amino acids with reduced tendency to form a helix (Val, Ile, Gly, and Phe) are abundant in TM helices, whereas polar amino acids (Glu, Lys, and Arg) with high tendency to form a helix are consistently less frequent in TM helices.
- Third, half of the natural amino acids are equally distributed along TM helices whereas aromatic, polar, and charged amino acids plus Pro are biased toward the ends of the TM helices.
- Fourth, as previously observed, the distribution of charged amino acids was asymmetric, occurring more frequently on the cytoplasmic side of the membrane, which causes net charge unevenness on both sides of the membrane. In addition to this asymmetry, Trp, Tyr, and Pro were found to be more frequent at the extra-cytoplasmic interface of the membrane and the polar amino acids (Gln, His, and Asn) at the extra-cytoplasmic flanking region of the TM helices.
- Fifth, transitions between the different types of amino acid at the ends of the hydrophobic core occur in a more defined region on the cytoplasmic side than at the extra-cytoplasmic face, probably reflecting the different lipid composition of both leaflets of biological membranes.

The conclusions on TM helix architecture described here should prove useful for constructing models of

membrane proteins with desired properties, which could help filling in some of the many gaps in our knowledge in this field.

Acknowledgments This work was supported by grants BFU2009-08401 (to I.M.) and BFU2010-19310 (to M.A.M.-R.) from the Spanish Ministry of Science and Innovation (MICINN, ERDF supported by the European Union), and by PROMETEO/2010/005 and ACOMP/2012/226 (to I.M.) and ACOMP/2011/048 (to M.A.M.-R.) from the Generalitat Valenciana. C.B.-D. was recipient of a predoctoral FPI fellowship from the MICINN.

References

- Arkin IT, Brunger AT (1998) Statistical analysis of predicted transmembrane alpha-helices. *Biochim Biophys Acta* 1429:113–128
- Berman HM, Berman HM, Westbrook J, Westbrook J, Feng Z, Feng Z, Gilliland G, Gilliland G, Bhat TN, Bhat TN, Weissig H, Weissig H, Shindyalov IN, Shindyalov IN, Bourne PE, Bourne PE (2000) The protein data bank. *Nucleic Acids Res* 28:235–242
- Blaber M, Zhang XJ, Matthews BW (1993) Structural basis of amino acid alpha helix propensity. *Science* 260:1637–1640
- Bowie JU (1997) Helix packing in membrane proteins. *J Mol Biol* 272:780–789
- Bywater RP, Thomas D, Vriend G (2001) A sequence and structural study of transmembrane helices. *J Comput Aided Mol Des* 15:533–552
- Cordes FS, Bright JN, Sansom MSP (2002) Proline-induced distortions of transmembrane helices. *J Mol Biol* 323:951–960
- Eilers M, Patel AB, Liu W, Smith SO (2002) Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys J* 82:2720–2736
- Engel DE, DeGrado WF (2004) Amino acid propensities are position-dependent throughout the length of alpha-helices. *J Mol Biol* 337:1195–1205
- Hessa T, Kim H, Bihlmaier K, Lundin C, Boeckl J, Andersson H, Nilsson I, White SH, von Heijne G (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433:377–381
- Hessa T, Meindl-Beinker NM, Bernsel A, Kim H, Sato Y, Lerch-Bader M, Nilsson I, White SH, von Heijne G (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* 450:1026–1030
- Holt A, Killian JA (2009) Orientation and dynamics of transmembrane peptides: the power of simple models. *Eur Biophys J* 39:609–621
- Huang Y, Huang Y, Niu B, Niu B, Gao Y, Gao Y, Fu L, Fu L, Li W, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26:680–682. Available at: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/link.fcgi?dbfrom=pubmed&id=20053844&retmode=ref&andcmd=prlinks>
- Illergård K, Kauko A, Elofsson A (2011) Why are polar residues within the membrane core evolutionary conserved? *Proteins* 79:79–91
- Jayasinghe S, Hristova K, White SH (2001a) Energetics, stability, and prediction of transmembrane helices. *J Mol Biol* 312:927–934
- Jayasinghe S, Jayasinghe S, Hristova K, Hristova K, White SH, White SH (2001b) MProtop: a database of membrane protein topology. *Protein Sci* 10:455–458
- Johansson ACV, Lindahl E (2007) Position-resolved free energy of solvation for amino acids in lipid membranes from molecular dynamics simulations. *Proteins* 70:1332–1344

- Lereh-Bader M, Lundin C, Kim H, Nilsson I, von Heijne G (2008) Contribution of positively charged flanking residues to the insertion of transmembrane helices into the endoplasmic reticulum. *Proc Natl Acad Sci USA* 105:4127–4132
- Li SC, Deber CM (1994) A measure of helical propensity for amino acids in membrane environments. *Nat Struct Biol* 1:558
- Lomize MA, Pogozheva ID, Joo H, Mossberg HI, Lomize AL (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res* 40:370–376
- MacCallum JL, Bennett WFD, Telemann DP (2008) Distribution of amino acids in a lipid bilayer from computer simulations. *Biophys J* 94:3393–3404
- Martínez-Gil L, Saurí A, Martí-Renom MA, Mingarro I (2011) Membrane protein integration into the endoplasmic reticulum. *FEBS J* 278:3846–3858
- Nilsson I, von Heijne G (1990) Fine tuning the topology of a polytopic membrane protein: role of positively and negatively charged amino acids. *Cell* 62:1135–1141
- Nilsson I, Johnson AE, von Heijne G (2003) How hydrophobic is alanine? *J Biol Chem* 278:29389–29393
- Nilsson J, Persson B, von Heijne G (2005) Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins* 60:606–616
- Orzáez M, Salgado J, Giménez-Giner A, Pérez-Payá F, Mingarro I (2004) Influence of proline residues in transmembrane helix packing. *J Mol Biol* 335:631–640
- Pal L, Chakrabarti P, Basu G (2003) Sequence and structure patterns in proteins from an analysis of the shortest helices: implications for helix nucleation. *J Mol Biol* 326:273–291
- Saurí A, Tamboreno S, Martínez-Gil L, Johnson AE, Mingarro I (2009) Viral membrane protein topology is dictated by multiple determinants in its sequence. *J Mol Biol* 387:113–128
- Scenes A, Gerstein M, Engelman DM (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol* 296:921–936
- Sharpe LJ, Stevens TJ, Munro S (2010) A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell* 142:158–169
- Ullmschneider MB, Sansom MS (2001) Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta* 1512:1–14
- Ullmschneider MB, Sansom MSP, Di Nola A (2005) Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* 59:252–265
- von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* 225:487–494
- Wallin E, von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* 7:1029–1038
- White SH (2004) The progress of membrane protein structure determination. *Protein Sci* 13:1948–1949
- White SH (2009) Biophysical dissection of membrane proteins. *Nature* 459:344–346
- White SH, Wimley WC (1999) Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct* 28:319–365
- Williams RW, Chang A, Juretić D, Loughran S (1987) Secondary structure predictions and medium range interactions. *Biochim Biophys Acta* 916:200–204

ANNEX IV

Baeza-Delgado, C., von Heijne, G., Marti-Renom, M. A. and Mingarro, I. (2016) ‘**Biological insertion of computationally designed short transmembrane segments.**’, *Scientific reports*. Nature Publishing Group, 6(March), p. 23397.

Selected as **Paper of the Month** by the Spanish Society of Biophysics (SBE, *Sociedad Española de Biofísica*), March 2016.

All the results of this publication are contained within this Thesis. Carlos Baeza has performed the experiments and contributed for their design and to the writing of the manuscript.

Part of the experiments of this publication were realized in the Department of Biochemistry and Biophysics of the Stockholm University, under supervision of Prof. Gunnar von Heijne.





SCIENTIFIC REPORTS

OPEN

Biological insertion of computationally designed short transmembrane segments

Carlos Baeza-Delgado¹, Gunnar von Heijne², Marc A. Marti-Renom^{3,4,5} & Ismael Mingarro¹

Received: 20 January 2016
Accepted: 07 March 2016
Published: 18 March 2016

The great majority of helical membrane proteins are inserted co-translationally into the ER membrane through a continuous ribosome-translocon channel. The efficiency of membrane insertion depends on transmembrane (TM) helix amino acid composition, the helix length and the position of the amino acids within the helix. In this work, we conducted a computational analysis of the composition and location of amino acids in transmembrane helices found in membrane proteins of known structure to obtain an extensive set of designed polypeptide segments with naturally occurring amino acid distributions. Then, using an *in vitro* translation system in the presence of biological membranes, we experimentally validated our predictions by analyzing its membrane integration capacity. Coupled with known strategies to control membrane protein topology, these findings may pave the way to *de novo* membrane protein design.

Transmembrane (TM) helices are the building blocks of the vast majority of integral membrane proteins. To perform their biological functions, integral membrane proteins must first insert their TM segments into the membrane in a helical conformation, and then acquire a defined three-dimensional structure by assembling their TM helices. In eukaryotic cells, the insertion and assembly of membrane proteins is mediated by the concerted action of a translating ribosome and the endoplasmic reticulum (ER) translocon, which allows lateral integration of TM helices into the ER membrane¹. Although it is well accepted that thermodynamically favorable partitioning of TM helices from the translocon into the more hydrophobic (lipidic) environment is important at the insertion stage^{2,3}, the limits for the insertion of TM helices with naturally occurring amino acid distributions has not been systematically explored.

The minimum hydrophobic length necessary to form a TM helix has been investigated using model membrane-inserted hydrophobic peptides⁴. These results show that for alternating Leu and Ala peptides (which have a hydrophobicity typical of natural TM helices), 13 consecutive residues is the minimum necessary to adopt a predominantly TM disposition in synthetic bilayers with a biologically relevant thickness. More recently, TM disposition of poly-Leu sequences were analyzed using synthetic peptides and oriented phospholipid bilayers, *in vitro* insertion into microsomal membranes and molecular dynamics simulations⁵. Sequences with either blocks of or dispersed hydrophobic residues have also been analyzed using similar methods⁶. The picture that emerges from these studies is that lipid bilayers adapt to TM helices as short as 10–12 leucines. However, TM helices in native membrane proteins vary significantly in length, being 24.0 (±5.6) amino acids long on average⁷.

How exactly a particularly short TM helix responds to the surrounding lipid bilayer will depend not only on the nature of the lipids with which it is in contact but also on its amino acid composition and the distribution of amino acids along the helix. While the studies cited above are all based on model hydrophobic sequences composed of only a few different kinds of amino acids, TM helices of amino acid composition that more closely matches natural TM helices need to be studied to lay a foundation for more advanced TM helix designs.

Here we used a database including 792 TM helices from membrane proteins with known three-dimensional structure and topology to generate a collection of random sequences with naturally occurring TM compositions of different lengths. The membrane insertion efficiency of these sequences were predicted computationally and

¹Departament de Bioquímica i Biologia Molecular, ERI BioTecMed, Universitat de València. E-46100 Burjassot, Spain. ²Dept. of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University, 10691 Stockholm, Sweden. ³CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain. ⁴Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain. ⁵Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain. Correspondence and requests for materials should be addressed to I.M. (email: Ismael.Mingarro@uv.es)

systematically examined experimentally using a microsomal *in vitro* expression system. The results reveal a correlation between the predictions and the experimental data, but also highlight the importance of local residue positioning, particularly with respect to the presence of proline residues and putative salt bridges within the TM segment.

Results

Predicted insertion capacity for designed sequences. The membrane insertion efficiency of the computationally designed sequences (see Methods and Supplementary Information Fig. S1 for details of the design procedure) was predicted using the experimentally based ΔG Prediction Server (<http://dgpred.cbr.su.se>). In an initial screen, the insertion efficiency of polyleucine segments of different lengths (l , from 9 to 25 residues) was calculated. Stretches of 11 or more leucine residues were predicted to be fully inserted, while 9 consecutive leucines was not enough to be inserted and 10 leucines resulted in a probability of insertion (p_i) of 0.59 (Fig. 1A, first row). As expected, these results were in excellent agreement with the previous experimental data³ that was used to construct the ΔG Predictor. These extremely short sequences would likely provoke the adaptation of the surrounding bilayer to reduce the putative hydrophobic mismatch by changes in the lipid order parameters in the peptide neighborhood, according to molecular dynamics simulations⁵.

We thus began by testing computationally designed segments with more variable amino acid composition, initially formed by leucine and alanine residues, which are the two most prevalent residues in TM helices. Sets of 1,000 sequences from 9 to 25 residues long were computed with the leucine-alanine relative ratio (58.2%Leu-41.8%Ala) found in TM helices in membrane proteins of known structures⁷. At least 13 residues were needed to obtain a significant level of predicted insertion (Fig. 1A second row), which was in good agreement with what it was found for alternating leucine and alanine peptides in lipid vesicles⁴. Subsequently, we included one by one the rest of the 20 natural amino acids following the order of prevalence found in native membrane proteins⁷. For instance, sets of 1,000 sequences formed by leucines, alanines and valines were generated using 42.6%Leu, 30.6%Ala and 26.8%Val residues, for each sequence length (SI Table 1). As expected, sequence sets including less hydrophobic residues needed longer segments to achieve insertion p_i values greater than 0.5 (Fig. 1A). Interestingly, when the less abundant lysine, glutamine, cysteine and aspartate residues were included at their naturally observed frequencies, the algorithm predicted that these computationally designed sequences were not expected to be efficiently inserted into biological membranes through the ER translocon. However, some natural sequences with these precise compositions are successfully inserted *in vivo*. Therefore, in line with previous findings³, we hypothesized that not only the amino acid composition is relevant for TM insertion, but that the actual position of the amino acid residues with respect to the center of the TM segment could also have an important influence on the insertion process.

To address this question, we changed our computational design algorithm to reflect the amino acid frequency values stratified at different position within a TM segment, taking as a reference the TM center⁷. Again, sets of 1,000 sequences were designed, using this position-dependent distribution of residues in natural TM helices, and their p_i values were predicted (Fig. 1B). The predicted insertion efficiency for sets with higher proportion of hydrophilic residues increased significantly when residue position constrains were included in the computational design. Hence, series of position-defined sequences including arginine, histidine, glutamate, lysine, glutamine, cysteine and aspartate were predicted to insert more efficiently at any given segment length than the previous sets based only on the global TM helix residue composition. The differences between the two data sets were most evident for TM sequences that include the less prevalent, more hydrophilic residues (Fig. 1C).

Correlation between predicted and experimentally determined insertion efficiencies. To be able to directly compare the predicted insertion efficiencies to insertion into the mammalian ER, some of the computationally designed sequences were analysed using a well-established *in vitro* assay for quantifying the efficiency of membrane integration of designed TM sequences into dog pancreas rough microsomes (RMs)⁸. The host protein (Lep) consists of two TM segments (H1 and H2) connected by a cytoplasmic loop (P1) and a large C-terminal domain (P2), and inserts into ER-derived RMs with both termini located in the lumen (Fig. 2A). The designed sequence ("TM-tested") was engineered into the luminal P2 domain and flanked by two acceptor sites (G1 and G2) for N-linked glycosylation. The engineered glycosylation sites can be used as membrane insertion reporters because G1 will always be glycosylated due to its native luminal localization, but G2 will be glycosylated only upon translocation of the analyzed region through the microsomal membrane. A singly glycosylated construct in which TM-tested is inserted into the membrane has a molecular mass ~ 2.5 kDa higher than the molecular mass of Lep expressed in the absence of microsomes; the molecular mass shifts by ~ 5 kDa upon double glycosylation (i.e. membrane translocation of the TM-tested).

We measured membrane insertion efficiencies of systematically designed sequences GGPG-X₁₁₋₂₃-GPGG, in which the flanking tetrapeptides are included to insulate the central computed stretches of different lengths (11 to 23 residues long) from the surrounding sequence in the Lep model protein. The insertion efficiency was calculated on the basis of the fractions of singly (f_{1g}) and doubly (f_{2g}) glycosylated forms by using $p_i = f_{1g}/(f_{1g} + f_{2g})$ determined from quantitative analyses of SDS-PAGE gels. Examples of SDS-PAGE gels showing the translation products of non-position defined leucine and alanine (LA) stretches 11, 13, 15 and 17 residues long and position-defined LAVIGFTSMYWPNR stretches of 17, 19, 21 and 23 residues long are shown in Fig. 2B.

Figure 3 shows plots of the probability of membrane insertion as a function of TM length for series of sequences that included the more prevalent amino acid residues in TM helices. Hydrophobic Leu and Ala account for more than one quarter (27.7%) of all amino acids in TM helices, and together with Val, Ile, Gly and Phe constitute the bulk of the amino acids embedded into the hydrophobic core of the membrane, accounting for almost two thirds (64.4%) of all amino acids in this membrane region⁷. The grey area corresponds to the 500 sequences between percentiles 0.25 and 0.75 of the 1,000 predicted p_i values (SI Fig. S2). The figure also shows

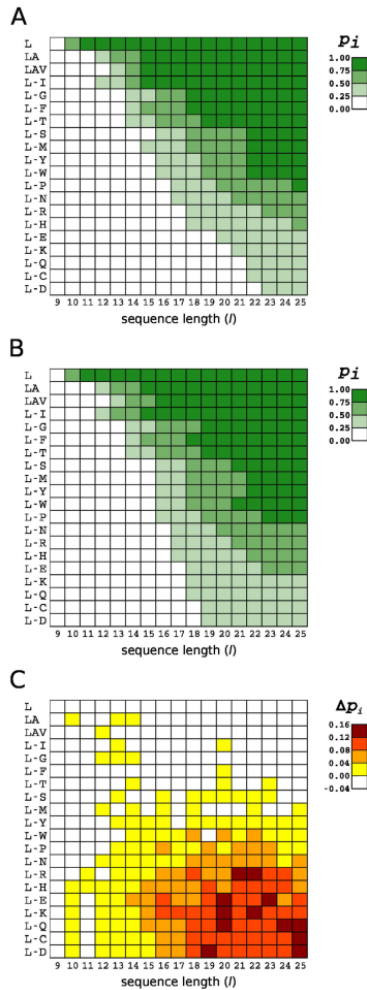


Figure 1. Predicted membrane insertion efficiencies for computationally designed sequences. (A) Probability of insertion (p_i) for a series of computationally designed TM segments of different lengths. First row corresponds to the predictions of p_i values for polyleucine stretches of different lengths (l). Each row in descending order represents the inclusion of a specific amino acid to the TM segment composition used in the previous row. The row order is derived from the prevalence for each amino acid type in TM helix composition in a previous structure-based statistical analysis⁷. (B) Similar to (A) but including in the computational design information of the position-dependent distribution for each amino acid type in TM helices⁷. (C) Differences between the position-defined (B) p_i values and those obtained for the computed sequences using only amino acid composition constraints (A).

ID	Sequence	$\Delta G_{\text{app}}^{\text{pred}}$	$\Delta G_{\text{app}}^{\text{exp}}$	p_i^{pred}	p_i^{exp}
L-K23	FFVHLE FIWLSVVMLPFAIEAYN	+0.96	-0.36	0.17	0.65
L-K23 H3V/V4H	FFVHLE FIWLSVVMLPFAIEAYN	+0.71	+0.14	0.23	0.44
L-R23	LMFRVATSAPLIILYGFMLTTT	-0.14	+0.62	0.56	0.26
L-R23 P10L/L14P	LMFRVATS ALLI IPYGFMLTTT	+0.13	+0.70	0.45	0.24
L-R23 A6P/P10A	LMFRV PTSA LIILYGFMLTTT	-0.30	+0.11	0.62	0.45

Table 1. Thermodynamic cost of L-K23- and L-R23-derived TM segments integration. The predicted and experimental (ΔG_{app}) energetic cost in kcal/mol of the computationally designed TM segments. Negative values are indicative of TM disposition, while positive values indicate non-TM disposition. Charged residues studied are highlighted in color and mutated residues are shown in bold.

the probability of insertion for the experimentally measured sequences (orange line) as well as their particular predicted values (blue line) (see SI Table S2 for details on the TM segments analyzed).

As expected, the inclusion of less prevalent amino acid residues in the designed TM sequences increased the variability of the predictions (compare the grey areas in Figs 3 and 4A) for all sets of sequence lengths. Moreover, the differences between the predictions and the experimental measurements were larger for some sequences (Fig. 4A). Next, we introduced constrains in our computational designs derived from the position-defined distributions found in native TM helices⁷. The introduction of the amino acid position constrains in our computational algorithm increased both the predicted as well as the experimentally measured p_i values for the TM segments containing less abundant (polar) residues (Fig. 4B). Nevertheless, we noticed that in some cases the differences between the predicted and experimental values were large. In these cases, we re-ran the ΔG Predictor algorithm but this time the algorithm was allowed to identify subsequences (i.e., with lower ΔG estimated values). The new predictions, in all cases, approached the experimental values (Fig. 4A,B, dashed lines), reinforcing that biological membranes can adapt to accommodate sequences harboring deviations from canonical hydrophobic regions.

Considering the complexity of the biological system, the two sets of p_i values are well correlated (Fig. 5): the linear fit has a slope of 0.80 with an r value of 0.93. Interestingly, the only outliers (highlighted as empty dots) are the longer sequences designed without position-dependent constrains, which include polar/charged residues. A closer look to the first of these sequences (L-K23, empty orange dot) revealed the presence of a histidine and glutamate residues (Table 1). This construct inserted experimentally much more efficiently than predicted by the ΔG algorithm. Given the 3.6-residue periodicity of an ideal α -helix, an intrahelical charge pair would be expected for this ($i, i+3$) His-Glu pair. To test this hypothesis, we swapped the histidine residue with its neighboring valine residue (L-K23 H3V/V4H, Table 1) generating an ($i, i+2$) periodicity for the His-Glu pair and likely precluding intrahelical pairing by orienting the two side-chains toward opposite faces of the helix. Noticeably, this mutant resulted in a slightly increased predicted p_i value but consistently diminished experimental insertion efficiency. The combined effect of these data improved the correlation between the experimental and prediction values for the mutant sequence (Fig. 5, arrow pointed dot), which has the same amino acid composition as the L-K23 sequence. These results support the idea that intra-helical salt-bridge formation between residues located on the same face of a TM helix ($i, i+3$) may reduce the free energy of membrane partitioning^{9,10}, whereas the presence of His and Glu on opposite faces of the helix ($i, i+2$) is unfavorable and lowers the ER translocon membrane insertion efficiency.

Next, we analyzed the L-R23 (Fig. 5, empty blue dot) sequence. In this case the predicted value suggests a higher propensity to insert than the measured experimental value. Inspection of L-R23 sequence (Table 1) highlighted the presence of a proline residue in a central position within the hydrophobic region (-2 relative to the center, negatively labeled positions indicate extra-cytoplasmic face). In general, the presence of proline residues in an α -helix generates a constrained Φ rotamer at the position of the proline, the loss of a hydrogen bond donor and the appearance of steric clashes between the proline cyclic side chain and the peptide backbone. In the case of TM helices, all these effects may eventually increase the polarity of the carbonyl groups of the TM helix at the positions three and four residues N-terminal of the proline location¹¹, reducing insertion efficiency¹².

Based on the L-R23 construct, we made two different mutants with the proline residue placed at different positions. In particular, we compared the insertion efficiency of positioning the proline roughly one helical turn (4 residues) towards the N or C terminus by swapping mutagenesis. When the proline residue was moved towards the C terminus ($+2$ position, mutant P10L/L14P) the experimental p_i value remained very similar to the one obtained for the original sequence (see Table 1), whereas the construct in which the proline was moved four residues towards the N terminus (-6 position, mutant A6P/P10A) was inserted more efficiently into the ER (Table 1). The different effect observed for these two mutants can be explained by the different location of the proline residue in relation to the midpoint of the TM segment. Hence, in the case of P10L/L14P mutant ($+2$ position) the distortions produced by the proline occur around the center of the membrane plane where the system is probably more sensitive to distortions. On the contrary, in the case of A6P/P10A mutant (-6 position) the presence of the proline closer to the interface would locate the unsatisfied carbonyl group in a less hydrophobic environment^{13,14}, probably reducing the free energy of membrane partitioning.

Discussion

In this study, we have systematically explored how the amino acid composition and positioning affect the efficiency of membrane insertion capacity of computationally designed TM segments, using both prediction and experimental measurements. We used a microsomal *in vitro* expression system to examine the translocon

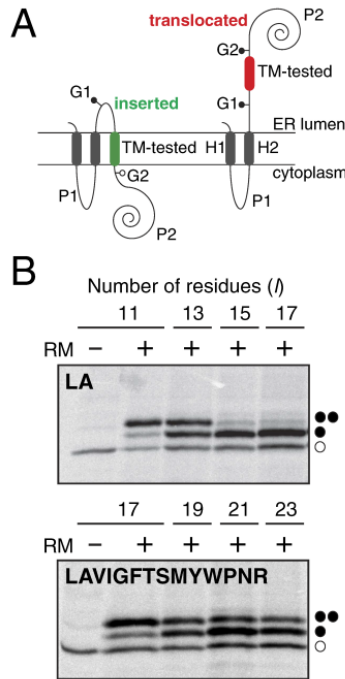


Figure 2. Integration of computationally designed TM segments into microsomal membranes. (A) Schematic of the engineered leader peptidase (Lep) model protein. Lep, consisting of 2 TM segments (H1 and H2) and a large luminal domain (P2), inserts into rough microsomes in an $N_{\text{lumen}}-C_{\text{lumen}}$ orientation. Computationally designed TM sequences were engineered into the P2 domain with flanking glycosylation sites (G1 and G2). For sequences that integrate into the membrane (green), only the G1 site is glycosylated (left), whereas both G1 and G2 are modified for sequences (red) that do not integrate into the membrane (right). (B) *In vitro* translation in the presence (+) or absence (-) of rough microsomes (RM) of computationally designed TM sequences of different length (l) composed of Leu and Ala residues (top), and of Leu, Ala, Val, Ile, Gly, Phe, Thr, Ser, Met, Tyr, Trp, Pro, Asn and Arg residues with position-defined constraints (bottom). Non-glycosylated protein bands are indicated by an empty dot; singly and doubly glycosylated proteins are indicated by one or two black dots, respectively.

insertion efficiency of chosen examples of the designed sequences. To generate the sequences we took advantage of the calculated distributions of amino acids from our previous structure-based statistical analysis of TM helices⁷.

Prior work showed that polyleucine sequences of 9–10 residues were efficiently inserted by the ER translocon into microsomal membranes^{5,15} and by the *E. coli* translocon¹⁶. However, TM segments in natural membrane proteins are not made exclusively of leucines. To expand our knowledge towards natural membrane proteins, we have designed and analyzed large sets of sequences with amino acid compositions that become more and more like the natural ones. Using the ΔG Prediction Server we found that 12–14 consecutive hydrophobic residues is about the minimum required for insertion into biological membranes through the ER translocon for highly hydrophobic sequences composed by leucine, alanine, valine and isoleucine (L-I series), which account for almost half (47.8%) of amino acid residue composition in TM helices. Sequences containing less prevalent (more hydrophilic) amino acid residues in TM segments have to be longer to efficiently insert into the membrane. Not surprisingly, there is a correlation between amino acid abundance in TM helices and hydrophobicity (SI Fig. 3), which explains the need for an increased sequence length compensating the lower hydrophobicity. However, this effect can be partly balanced by taking into account amino acid position-dependent contributions. Thus, when this last parameter

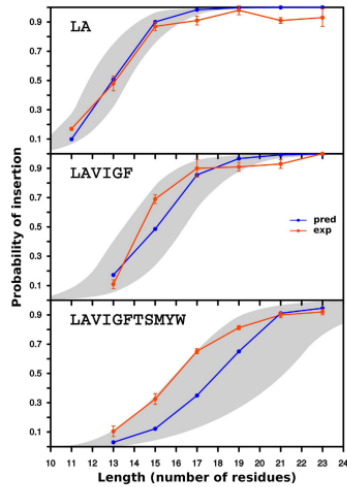


Figure 3. Predicted and experimental p_i values of sequences including the most prevalent amino acid residues in TM helices. Upper panel: Computationally designed leucine and alanine sequences of different lengths (l). The predicted values for each given sequence are shown in blue and the measured values obtained for three independent experiments in orange. The gray area represents the predictions of the p_i values for the 500 sequences between percentiles 0.25 and 0.75 of the total population (1,000 computed sequences, see Fig. S2). Central panel: Similar to upper but the computationally designed TM sequences contained leucine, alanine, valine, isoleucine, glycine and phenylalanine. Bottom panel: Similar to upper but including, in addition to the hydrophobic, the more prevalent polar and aromatic residues in TM segments. See Supplementary Information Table 2 for details on the TM-tested sequences used.

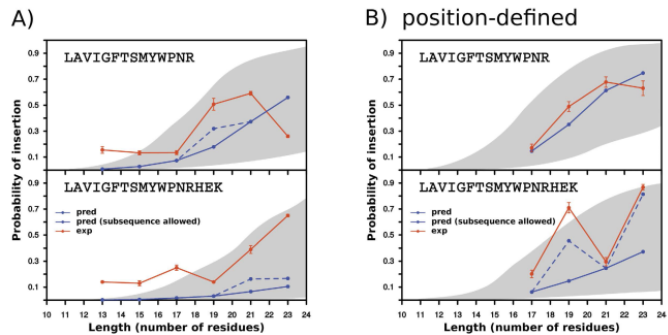


Figure 4. Predicted and experimental p_i values of computationally designed TM sequences including polar and charged amino acid residues. Experimental and predicted values are shown as in Fig. 3. (A) Computationally designed TM sequences were constrained only by amino acid composition constraints. (B) Computationally designed TM sequences were constrained both by amino acid composition and distribution along the helix. See Supplementary Information Table 2 for details on the TM-tested sequences used.

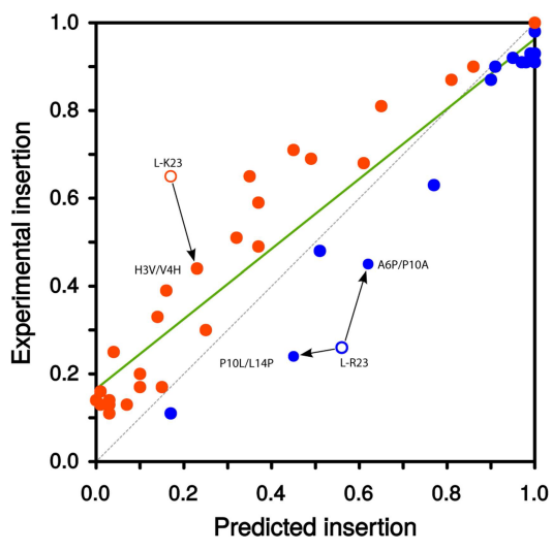


Figure 5. Correlation between experimentally measured and predicted p_i values. For each sequence analyzed, predicted values higher than the experimental ones (i.e., below the grey dashed line) are shown in blue, whereas experimental values higher than the predicted ones (i.e., above the grey dashed line) are shown in orange. The correlation between the experimental and predicted insertion probabilities is indicated by a green line. Outliers are shown as empty circles and the results of their mutated sequences (that is, P10L/L14P, A10P/P10A and H3V/V4H) are indicated by arrows.

was included in our computational designs, the predictions for the insertion efficiency of sequences harboring polar and charged residues increased significantly (Fig. 1C).

The results of our experimental assay using microsomal membranes (Figs 3 and 4) mirrored the ΔG algorithm predictions. Our data reinforces the accepted idea that there is an unfavorable free energy associated with locating hydrophilic residues in the hydrophobic core of the membrane. However, this effect can be reduced by allocating non-hydrophobic residues close to the polar headgroup region of the lipids at the membrane interface¹⁷, as well as by engaging polar residues in salt-bridge pair formation¹⁸. This is in line with our analysis of specific sequences with experimental/predicted p_i values that deviate from a linear correlation (Fig. 5). One of these outlier sequences (L-K23) was inserted by the translocon surprisingly efficiently compared with its ΔG Predictor value (Table 1). We suggest that the stability of L-K23 TM helix within the lipid membrane is derived from intra-helical electrostatic and/or salt-bridge interaction between the histidine and the glutamic acid side chains positioned at $(i, i + 3)$ periodicity. In TM helices, intra-helical charge pairs within the same helix have been reported for appropriately spaced $(i, i + 3)$ and $(i, i + 4)$, oppositely charged residues¹⁹. We support this hypothesis by locating the His-Glu pair at $(i, i + 2)$ periodicity, which is non-compatible with intra-helical electrostatic and/or salt-bridge interaction. This mutation strongly reduced the experimental insertion efficiency. As expected, ionizable histidine and glutamic acid residues are present in TM helices at a low frequency level (1.7% and 1.6%, respectively). Nevertheless, among the 792 TM helices included in our database, 84 helices (10.6%) contained both amino acid residues in their sequence, and 15 of these helices present the His-Glu pair at $(i, i + 3)$ periodicity. Approximately, only one fourth of these His-Glu pairs (4) are partly exposed to the lipid face, whereas the rest are buried in the protein interior, emphasizing the necessity to shield the polarity of this interaction from the hydrophobic environment of the membrane core. Charged and polar residues can face a high energetic barrier when inserting into a biological membrane. Nevertheless, positive and negative charges within the same^{6,10} and different^{19,20} TM helices can interact with each other, thereby drastically reducing this barrier. In fact, these interactions can occur in the proximity of the translocon^{21,22}, where a strict coupling of correct tertiary structure formation and membrane insertion can be achieved²³.

The other outlier sequence (L-R23) displayed a completely different behavior, since in this case the experimentally measured p_i value was lower than the predicted value (Fig. 5). This sequence contains a proline residue

at a nearly central position in the TM segment. Proline is rarely found in the middle of helices from soluble proteins because it results in distortion of the canonical helical geometry and loss of at least one backbone hydrogen bond^{24,25}. However, proline residues are relatively common in TM helices^{7,11}. This suggests that proline residues may be of particular structural and/or functional significance in membrane proteins, even though they invariably produce deviations from canonical helical structure²⁶. To learn how the proline present in the L-R23 mutant reduces membrane insertion, we analyzed the insertion of a mutant with the proline residue positioned near the middle of the helix (P10L/L14P), and a mutant with the proline residue located near the N-terminus of the helix (A6P/P10A), both with the same amino acid composition. The A6P/P10A mutant inserted more efficiently than the original sequence, while the P10L/L14P sequence resulted in only minor changes in terms of membrane insertion. These results indicate that proline residues are not easily accommodated in the center of the helix. Structural studies have shown that proline substitutions at the end of a TM helix can be accommodated by movement of a small part of the helix, while proline substitutions in the middle can require more complex and difficult to accommodate structural changes²⁶. Statistical analyses of TM helices show a similar pattern for proline residue distribution (SI Fig. S4). Altogether, these results indicate that the proline in the original sequence can diminish membrane insertion efficiency depending on the position along the TM segment.

In summary, our analysis of the membrane insertion of computationally designed TM sequences resulted in a good correlation between the values predicted by the ΔG Predictor and experimentally measured values. Nevertheless, our data indicate that some extra attention has to be paid to accommodate intra-helical salt-bridge formation and proline residues when designing short TM helices as building blocks of membrane proteins, a major challenge when engineering new membrane proteins to perform biomimetic functions.

Methods

Computational sequence design. All statistics of amino acid composition at the TM level as well as the positional level was derived from our previously published datasets⁷. Briefly, our dataset included a total of 170 non-redundant structures described in the MPro database²⁷ containing 792 TM helices of length from 17 and 38 residues, which resulted in a total of 19,356 amino acids. The dataset was further parsed to compute the probability of a given amino acid to be included in a TM helix. Using our dataset, we generated a series of designed sequences of a given length by populating them with an increased number of amino acid types. First, we generated sequences with only Leu amino acid type (the most common in TMs) of lengths ranging from 9 to 25 residues. Next, we included the second most common amino acid type in TMs (that is, Ala) with its relative probability compared to Leu. Again, we generated 1,000 sequences of each length where the sequences were obtained by shuffling a string of L and A letters with the proportions defined in Supplementary Information Table S1. The same procedure was repeated each time including a new amino acid type until we had all 20. Next, we generated a set of sequences with both, amino acid propensity as well as positional effect by taking TMs residues and annotating their position with respected the central part of the TM helix. In such case, we computed the probability of an amino acid type in each of the positions of a TM starting from the central one (position 0) and increasing the position number as we approached the cytoplasmic side of the TM or a negative number as we approached the extracellular side of the TM. The designed sequences were built taking into account the position by selecting amino acids for each pool across the TM helices. Similarly to the no position-defined dataset, 1,000 sequences of each different length were generated for each amino acid type compositions.

Prediction of the ΔG values and probability of insertion. All sets of 1,000 sequences of desired length, amino acid composition and non-positional or positional effect were used to generate a series of predicted insertion capacity scores, which were predicted using the experimentally based ΔG Prediction Server (<http://dgpred.cbr.su.se/>)²³. The generated sequences and their ΔG values can be obtained in the provided Excel file (SI Table S3).

Enzymes and chemicals. All enzymes as well as plasmid pGEM1, the TNT SP6 Quick Coupled System and rabbit reticulocyte lysate were from Promega (Madison, WI). ER rough microsomes from dog pancreas were from tRNA Probes (College Station, TX). [³⁵S]Met were from Perkin Elmer. The restriction enzymes were purchased from Roche Molecular Biochemicals. The DNA purification kits were from Thermo (Ulm, Germany). All the oligonucleotides were purchased from Sigma-Aldrich (Switzerland).

DNA manipulation. For experimental analysis of the computationally designed sequences, oligonucleotides encoding the computed hydrophobic (TM) regions were introduced into the P2 domain of *E. coli* leader peptidase (Lep). Tested sequences were constructed using two double-stranded oligonucleotides with 5' phosphorylated overlapping overhangs at the ends. Pairs of complementary oligonucleotides were first annealed at 85°C for 10 min followed by slow cooling to 30°C, after which the two or three annealed double-stranded oligonucleotides were mixed, incubated at 65°C for 5 min, cooled slowly to room temperature and ligated into the vector. Mutations at the designed TM segments were obtained by site-directed mutagenesis using the QuikChange kit (Stratagene, La Jolla, California). All TM segment inserts and mutants were confirmed by sequencing of plasmid DNA.

In vitro transcription and translation. Constructs in pGEM1 were transcribed and translated in the TNT SP6 Quick Coupled System (Promega). 75 ng DNA template, 0.5 μ L ³⁵S-Met (5 μ Ci) and 0.25 μ L microsomes (tRNA Probes) were added at the start of the reaction, and samples were incubated for 90 min at 30°C. Translation products were diluted in 50 μ L of loading buffer and analyzed by SDS-PAGE. The gels were quantified using a Fuji FLA-3000 phosphorimager and Image Reader 8.1j software. The membrane-insertion probability of a given TM sequence was calculated as the quotient between the intensity of the singly glycosylated band divided by the summed intensities of the singly glycosylated and doubly glycosylated bands.

References

- Martinez-Gil, L., Sauri, A., Marti-Renom, M. A. & Mingarro, I. Membrane protein integration into the ER. *FEBS J* **278**, 3846–3858, doi: 10.1111/j.1742-4658.2011.08185.x (2011).
- Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433**, 377–381 (2005).
- Hessa, T. *et al.* Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **450**, 1026–1030 (2007).
- Krishnakumar, S. S. & London, E. Effect of sequence hydrophobicity and bilayer width upon the minimum length required for the formation of transmembrane helices in membranes. *J Mol Biol* **374**, 671–687, doi: 10.1016/j.jmb.2007.09.037 (2007).
- Jaud, S. *et al.* Insertion of short transmembrane helices by the Sec61 translocon. *Proc Natl Acad Sci USA* **106**, 11588–11593, doi: 0900638106.10.1073/pnas.0900638106 (2009).
- Stone, T. A., Schiller, N., von Heijne, G. & Deber, C. M. Hydrophobic blocks facilitate lipid compatibility and translocon recognition of transmembrane protein sequences. *Biochemistry* **54**, 1465–1473, doi: 10.1021/bi5014886 (2015).
- Baeza-Delgado, C., Marti-Renom, M. A. & Mingarro, I. Structure-based statistical analysis of transmembrane helices. *Eur Biophys J* **42**, 199–207, doi: 10.1007/s00249-012-0813-9 (2013).
- Saaf, A., Wallin, E. & von Heijne, G. Stop-transfer function of pseudo-random amino acid segments during translocation across prokaryotic and eukaryotic membranes. *Eur J Biochem* **251**, 821–829 (1998).
- Bano-Polo, M. *et al.* Polar/ionizable residues in transmembrane segments: effects on helix-helix packing. *PLoS One* **7**, e44263, doi: 10.1371/journal.pone.0044263 (2012).
- Chin, C. N. & von Heijne, G. Charge pair interactions in a model transmembrane helix in the ER membrane. *J Mol Biol* **303**, 1–5 (2000).
- Cordes, F. S., Bright, J. N. & Sansom, M. S. Proline-induced distortions of transmembrane helices. *J Mol Biol* **323**, 951–960 (2002).
- Orzaez, M., Salgado, J., Gimenez-Giner, A., Perez-Paya, E. & Mingarro, I. Influence of proline residues in transmembrane helix packing. *J Mol Biol* **335**, 631–640, doi: 10.1016/j.jmb.2003.10.062 (2004).
- White, S. H. & Wimley, W. C. Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct* **28**, 319–365 (1999).
- MacCallum, J. L., Bennett, W. F. & Tieleman, D. P. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophys J* **94**, 3393–3404 (2008).
- Kuroiwa, T., Salaguchi, M., Mihara, K. & Omura, T. Systematic analysis of stop-transfer sequence for microsomal membrane. *J Biol Chem* **266**, 9251–9255 (1991).
- Chen, H. F. & Kendall, D. A. Artificial transmembrane segments - Requirements for stop transfer and polypeptide orientation. *J Biol Chem* **270**, 14115–14122 (1995).
- White, S. H. & von Heijne, G. Do protein-lipid interactions determine the recognition of transmembrane helices at the ER translocon? *Biochem Soc Trans* **33**, 1012–1015 (2005).
- Jayasinghe, S., Hristova, K. & White, S. H. Energetics, stability, and prediction of transmembrane helices. *J Mol Biol* **312**, 927–934, doi: 10.1006/jmb.2001.5008 (2001).
- Johnson, E. T. & Parson, W. W. Electrostatic interactions in an integral membrane protein. *Biochemistry* **41**, 6483–6494 (2002).
- Bano-Polo, M. *et al.* Charge pair interactions in transmembrane helices and turn propensity of the connecting sequence promote helical hairpin insertion. *J Mol Biol* **425**, 830–840, doi: 10.1016/j.jmb.2012.12.001 (2013).
- Sadlish, H., Pitzonzo, D., Johnson, A. E. & Skach, W. R. Sequential triage of transmembrane segments by Sec61alpha during biogenesis of a native multispanning membrane protein. *Nat Struct Mol Biol* **12**, 870–878 (2005).
- Sauri, A., McCormick, P. J., Johnson, A. E. & Mingarro, I. Sec61alpha and TRAM are sequentially adjacent to a nascent viral membrane protein during its ER integration. *J Mol Biol* **366**, 366–374 (2007).
- Cymer, F., von Heijne, G. & White, S. H. Mechanisms of integral membrane protein insertion and folding. *J Mol Biol* **427**, 999–1022, doi: 10.1016/j.jmb.2014.09.014 (2015).
- Barlow, D. J. & Thornton, J. M. Helix geometry in proteins. *J Mol Biol* **201**, 601–619 (1988).
- von Heijne, G. Proline kinks in transmembrane α -helices. *J Mol Biol* **218**, 499–503 (1991).
- Yohannan, S. *et al.* Proline substitutions are not easily accommodated in a membrane protein. *J Mol Biol* **341**, 1–6, doi: 10.1016/j.jmb.2004.06.025 (2004).
- Jayasinghe, S., Hristova, K. & White, S. H. MPtopo: A database of membrane protein topology. *Protein Sci* **10**, 455–458, doi: 10.1110/ps.43501 (2001).

Acknowledgements

This work was supported by grants BFU2012-39482 to I.M. and BFU2010-19310/BMC and BFU2013-47736-P to M.A.M.-R. from the Spanish Ministry of Economy and Competitiveness (MINECO, co-financed by European Regional Development Fund), PROMETEOII/2014/061 from the Generalitat Valenciana to I.M., and by grants from the Swedish Foundation for Strategic Research, the European Research Council (ERC-2008-AdG 232648), the Swedish Cancer Foundation, the Swedish Research Council, and the Knut and Alice Wallenberg Foundation to GvH. C.B.-D. was recipient of a predoctoral FPI fellowship and an FPI short-staying grant from the MINECO to visit the laboratory of G.v.H.

Author Contributions


C.B.-D., M.A.M.-R. and I.M. conceived the experiments. C.B.-D. and M.A.M.-R. carried out the computational sequence design. C.B.-D. performed the translocon experiments. C.B.-D., M.A.M.-R., G.v.H. and I.M. analysed the data. I.M. wrote the paper, all authors discussed the results and edited the manuscript.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Baeza-Delgado, C. *et al.* Biological insertion of computationally designed short transmembrane segments. *Sci. Rep.* **6**, 23397; doi: 10.1038/srep23397 (2016).

 This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

ANNEX V

Bañó-Polo, M., Baeza-Delgado, C., Orzáez, M., Marti-Renom, M. A., Abad, C. and Mingarro, I. (2012) '**Polar/Ionizable Residues in Transmembrane Segments: Effects on Helix-Helix Packing**', *PLoS ONE*, 7(9), pp. 1–8.

These results are part of the background of this Thesis. Carlos Baeza contributed by performing the computational analysis and figures 1 and 2.

Polar/Ionizable Residues in Transmembrane Segments: Effects on Helix-Helix Packing

Manuel Baño-Polo¹, Carlos Baeza-Delgado¹, Mar Orzáez², Marc A. Martí-Renom^{3,4}, Concepción Abad¹, Ismael Mingarro^{1*}

1 Departament de Bioquímica i Biologia Molecular, Universitat de València, Burjassot, Spain, **2** Centro de Investigación Príncipe Felipe, Valencia, Spain, **3** Genome Biology Group, Structural Genomics Team, Centre Nacional d'Anàlisi Genòmic, Barcelona, Spain, **4** Structural Genomics Group, Center for Genomic Regulation, Barcelona, Spain

Abstract

The vast majority of membrane proteins are anchored to biological membranes through hydrophobic α -helices. Sequence analysis of high-resolution membrane protein structures show that ionizable amino acid residues are present in transmembrane (TM) helices, often with a functional and/or structural role. Here, using as scaffold the hydrophobic TM domain of the model membrane protein glycoporphin A (GpA), we address the consequences of replacing specific residues by ionizable amino acids on TM helix insertion and packing, both in detergent micelles and in biological membranes. Our findings demonstrate that ionizable residues are stably inserted in hydrophobic environments, and tolerated in the dimerization process when oriented toward the lipid face, emphasizing the complexity of protein-lipid interactions in biological membranes.

Citation: Baño-Polo M, Baeza-Delgado C, Orzáez M, Martí-Renom MA, Abad C, et al. (2012) Polar/Ionizable Residues in Transmembrane Segments: Effects on Helix-Helix Packing. PLoS ONE 7(9): e44263. doi:10.1371/journal.pone.0044263

Editor: Peter Butko, Nagoya University, Japan

Received: April 25, 2012; **Accepted:** July 31, 2012; **Published:** September 12, 2012

Copyright: © 2012 Baño-Polo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants BFU2009-08401 (to IM) and BFU2010-19310 (to MAM-R) from the Spanish Ministry of Science and Innovation (MICINN, ERDF supported by the European Union), as well as by PROMETEO/2010/005 (to IM) from the Generalitat Valenciana. MB-P and CB-D were recipients of FPU and FPI predoctoral fellowships from the MICINN, respectively. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Ismael.Mingarro@uv.es

Introduction

The vast majority of membrane proteins are anchored to biological membranes through hydrophobic α -helices. These transmembrane (TM) α -helices, rather than serving solely as featureless hydrophobic stretches required for anchorage of proteins in membranes, have structural and/or functional roles well beyond this canonical capacity. In fact, the folding and assembly of membrane proteins rely in part on interacting TM helices, which was conceptualized as a two-stage process [1]. In the first stage, TM helices are inserted into the membrane by the translocon. The driving force for this process derives primarily from the transfer of hydrophobic side chains from the aqueous channel of the translocon to the apolar region of the bilayer [2]. In the second stage, the protein attains its native tertiary structure through the packing of its TM helices. In the apolar environment of the membrane core, van der Waals packing, hydrogen bonding and ionic interactions are the dominant contributors to TM helix packing.

Sequence analysis of high-resolution membrane protein structures show that ionizable amino acid residues are present in TM helices, although at a low frequency level [3]. Insertion of these residues through the translocon has been proved to be feasible thanks to the overall hydrophobicity of the TM segment [4] and depending on their position along the hydrophobic region [5]. In many cases, ionizable residues are involved in TM helix packing [6,7,8]. Likely, hydrogen bonding [6,7] or salt-bridge [9] formation with other membrane-spanning hydrophilic residues drives these interactions, while at the same time, reduces the

unfavorable energetics of inserting polar or ionizable residues into the hydrophobic membrane core.

Homo-oligomeric membrane proteins provide attractive systems for the study of TM helix packing because of their symmetry and relative simplicity. These model systems can serve as an excellent starting point to understand the structural dynamics and folding pathways of larger membrane proteins. One of the best-suited models of membrane protein that oligomerizes (more specifically, dimerizes) through non-covalent interactions of its TM α -helix is undoubtedly glycoporphin A (GpA) [10,11]. The wide use of this protein as a model membrane protein is partially based on its intrinsic simplicity, since the free energy decrease associated with TM helix-helix interactions is enough to confer detergent resistant dimerization to the protein. Thus, those factors that could affect or modify the dimerization process can be analyzed using sodium dodecyl sulfate (SDS)-PAGE. The GpA homodimer, defines a dimerization interface that has been extensively studied by diverse techniques such as saturation mutagenesis [12] and alanine-insertion scanning [13] in SDS micelles, solution NMR in dodecyl phosphocholine micelles [14] and solid-state NMR in lipid membranes [15]. The output of these studies describes a dimerization motif in the TM segment composed of seven residues, L⁷²IxxGVxxGVxxI⁸⁷, which is responsible for the dimerization process. More recently, using proline-scanning mutagenesis it was demonstrated that Leu75 is not so clearly involved in the packing process [16], focusing the interaction on the central G⁷⁹VxxGVxxI⁸⁷ motif, which includes the widely proved framework for TM helix association, GxxxG [17,18].

Nevertheless, the sequence context highly determines the thermodynamic stability of GxxxG-mediated TM helix-helix interactions (recently reviewed [19]).

In the present study, we have analyzed the distribution of ionizable (Asp, Glu, Lys and Arg) amino acid residues in TM segments from high-resolution membrane protein structures, which have to energetically accommodate into the highly hydrophobic core of biological membranes by interacting favorably with its local environment. Then, we address the consequences of replacing specific residues by ionizable amino acids along the hydrophobic region of the GpA TM domain on the dimerization of this model membrane protein, both in detergent micelles and in biological membranes. Our findings demonstrate that ionizable residues are stably inserted in hydrophobic environments, and tolerated in the dimerization process when oriented toward the lipid face, emphasizing the complexity of protein-lipids interactions in biological membranes.

Results and Discussion

Ionizable amino acid residues in TM α -helices

TM helices of lengths between 17 and 38 residues were selected from the MPDPO database [20], which included helical segments that do completely span the hydrophobic core of the membrane. TM helices shorter than 17 residues as well as larger than 38 residues were excluded since they may not cross entirely the membrane or may contain segments parallel to the membrane [3], respectively.

As expected, ionizable residues (Asp, Glu, Lys, and Arg) are present at a low frequency level. All together, these residues constitute only 6.6% of the residues within TM helices. Despite their lower presence, strongly polar residues are evolutionary conserved in TM proteins, which can be partially explained by their tendency to be buried in the protein interior and also in many cases due to their direct involvement in the function of the protein [21,22]. Among the 792 TM helices included in our database, 366 helices (46.2%) contained at least one ionizable residue within the hydrophobic region (that is, the central 19 amino acid residues). A summary of the statistics is presented in Figure 1. Furthermore, 96 TM helices contained at least one acidic plus one basic residue in their sequence, and 20 of these helices present oppositely charged residues with the appropriate periodicity ($i, i+4$) to form intrahelical charge pairs. To gain more detailed insight into the structural role of these ionizable residues within the membrane core, we analyzed the environment of all these 20 helices. Approximately half of the ionizable residues (51%) found in these helices are buried in the protein interior, but the rest are partly exposed to the lipid face. Some of these lipid facing ionizable residues are located in pairs at the appropriate distance to form a salt-bridge, as in the sarcoplasmic/endoplasmic reticulum calcium ATPase 1 protein (Fig. 2).

Effects on SDS-resistant TM helix packing

Next, we investigated the effect of strongly polar residues in TM helix packing using the GpA TM segment as a model (scaffold) segment. Initial polar mutations (I87D, T87K, I91D, and I91K) made on residues located at the helix-helix interface (Fig. 3A) abolished dimerization (Fig. 3B). Furthermore, it has been reported that T87S (which retains the side chain γ oxygen) permits dimer formation both in SDS micelles [23] and in *E. coli* membranes [24], whereas a bulkier hydroxylated side chain (I87Y) is strongly disruptive (Fig. 3B). However, point mutations corresponding to replacements of nonpolar residues located at the lipid-facing interface (Fig. 3A) by ionizable residues gave rise to a

Charged Residues in Transmembrane Segments

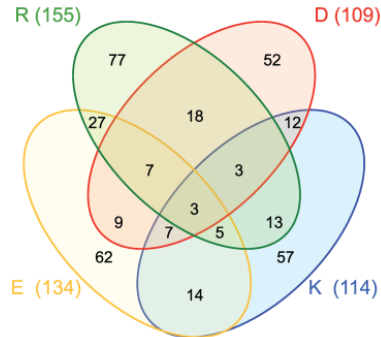


Figure 1. Venn diagram of TM segments (the central 19 residues) containing charged residues: Asp (D, red), Lys (K, blue), Glu (E, yellow) and Arg (R, green). The value in parenthesis is the total TM helices that contain at least one of such residues. The values inside the ellipses indicate the number of TM helices in each combination of these four amino acids. For example, there are 57 TM helices with only Lys as a charged residue, 12 helices with only Lys and Asp, 7 helices with Lys, Asp and Glu, and 3 helices with all four ionizable residues. doi:10.1371/journal.pone.0044263.g001

more tolerated response (Fig. 3B). When I85 was substituted by ionizable side-chain residues, either negatively charged (I85D) or positively charged (I85K and I85R), the dimerization level was similar to native GpA TM sequence as shown under SDS-PAGE analysis (Fig. 3C, compare lanes 2, 3 and 4 to lane 1). It is commonly assumed that single ionizable residues should exist in their uncharged form within membrane-spanning helices [25]. In fact, the pK_a values observed for Asp residues in hydrophobic helices were somewhat elevated (5–8.5) relative to those for Asp residues in solution [26]. Furthermore, the replacement of Leu89 by basic residues (L89K and L89R) had almost no effect, while its substitution by an acidic residue (L89D) abolished dimerization (Fig. 3B and 3C). The opposing consequences observed for Leu89 mutants can be explained taking into account the nature of the SDS-micelles used in these experimental conditions. These results suggest that L89D mutation alters the interaction of the protein with the negatively charged detergent micelle, possibly resulting in a structure that differs from a "transmembrane" α -helix due to helix distortions and interaction with the polar micelle surface. This effect was not observed when the Asp residue was located in a more central position (I85D), where its carboxylate should be located away from the negatively charged sulfate groups of the SDS molecules. In this regard, the capacity of SDS to respond to such nuance of sequence in terms of SDS solvation of TM segments within protein-SDS detergent complexes has been proved to be highly sequence (position) dependent [27]. Nevertheless, the comparable electrophoretic migration observed for I85D and L89D (Fig. 3C) suggests that the monomers associate with SDS quite similarly. To identify the helix interface responsible of dimer formation in the Leu89 mutants, we designed double mutants that contained a non-polar highly disruptive mutation (G83L). Gly83 has been proved to be extremely sensitive, since all mutations tested disrupted the dimer completely [12]. As

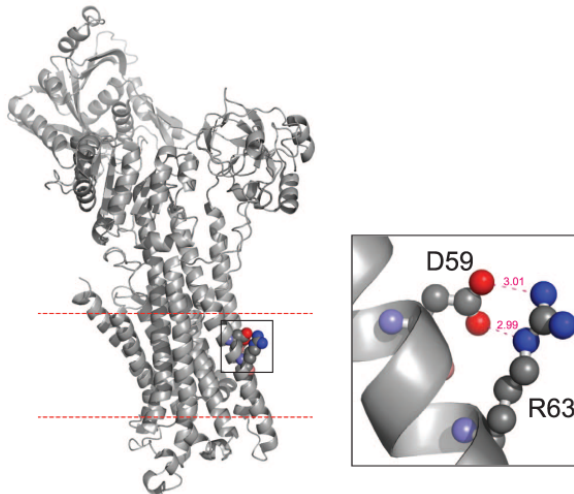


Figure 2. Structure of the calcium ATPase 1. Left panel: cartoon representation of sarcoplasmic/endoplasmic reticulum calcium ATPase 1 (PDB ID: 1SU4) with cytosolic domain in the up side and transmembrane aspartic 59 and arginine 63 residues in spheres representation (C atom gray, O atom red and N atom blue). Membrane boundaries (dashed red lines) were obtained from the PPM Server [44]. Right panel: zoom view centered on the salt bridge between Asp59 and Arg63, dashed pink lines indicate O to N atom distances. doi:10.1371/journal.pone.0044263.g002

shown in Fig. 3B, G83L mutant did not form any detectable dimer, and both double mutant proteins (G83L/L89D and G83L/L89D) containing this mutation did not dimerize, suggesting that the lysine residue introduced was not participating in the dimerization process, instead, the native dimerization motif is responsible of helix-helix interaction.

Given the 3.6-residue periodicity of an ideal α -helix, intrahelical charge pairs would be expected for $(i, i+4)$ Lys-Asp pairs. To further assess if intrahelical charge pair formation can be tolerated in dimerizing TM sequences, we performed a double mutation combining two strongly dimerizing sequences (I85D/L89K), which only reduced dimerization by about 50% compared to the wild-type sequence (Fig. 3B). Similarly, I85K/L89K mutant retained the same level of dimerization, likely favored by a beneficial SDS solvation effect on the lysine residues. On the contrary, when oppositely charged residues were located at the TM-interacting interface (I87D/I91K) dimerization was abrogated (Fig. 3B). Furthermore, when charge pairs include L89D mutation although facing the lipids, as for I85K/L89D, we found no evidence for dimer formation (Fig. 3B). These results suggest that charge pairs are tolerated only when located at the non-interacting interface, but solely at specific positions.

Recent mutational analysis of strongly self-interacting TM segments demonstrated that basic and acidic residues located at the helix-interacting interface participate in homotypic interactions [25]. In this case, basic and acidic residues spaced $(i, i+1)$ and $(i, i+2)$ contribute to the interaction of model TM segments. To

test this idea in the GpA sequence, we designed two mutants with appropriately spaced basic and acidic residues (L89D/I91K and L90D/I91K), and no dimeric forms were observed in any of these proteins.

In light of our experiments in SDS micelles, it can be concluded that nonpolar to ionizable substitutions away from the dimer interface (lipid facing) in combination with N-terminal native GpA dimerization motif (including GxxxG sequence) does not perturb the dimerization process, while similar mutations positioned at the helix-interacting interface strongly compromise dimer formation.

Effects on insertion and packing into biological membranes

To test the molecular effect of the ionizable residues in biological membranes we used a glycosylation mapping technique to measure changes in the insertion capacity of the GpA TM domain after introduction of ionizable residues at the more tolerant positions in terms of TM packing. The glycosylation mapping technique has been used previously to investigate the membrane insertion level of hydrophobic regions and to systematically examine the effects of individual residues on their position in the membrane [16,28,29]. The method is based on the observation that the endoplasmic reticulum (ER) enzyme oligosaccharide transferase (OST) can only transfer a sugar moiety to Asn-X-Thr/Ser acceptor sites when they are oriented toward the lumen of the ER membrane. To assess the effect of the presence of ionizable residues on the GpA TM segment insertion into

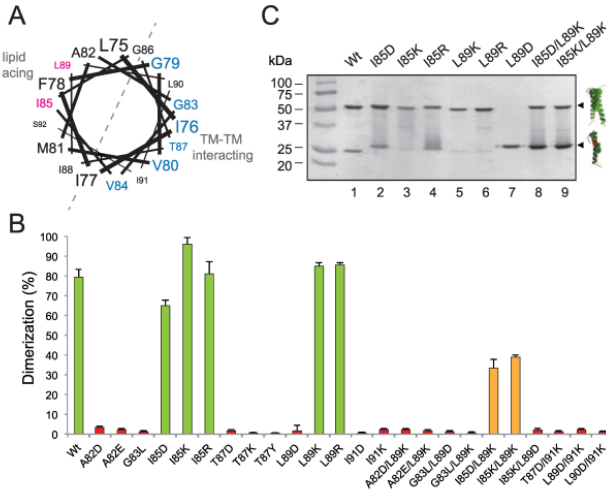


Figure 3. Dimerization in SDS micelles. (A) Helical wheel projection of GpA TM sequence. The residues associated with dimer formation as defined by Orzaez et al. [16] are shown in blue. Non-interacting residues susceptible of ionizable residue substitution are shown in magenta. (B) Green colored bars denoted dimerization levels similar to wild-type sequence. Bars for intermediate dimerization levels (~40%) are colored orange. Red colored bars denote non-dimerizing sequences (dimerization <3%). (C) SDS-PAGE analysis of GpA mutants. Chimeric proteins were purified in the presence of SDS and analyzed by PAGE. Positions of the monomer and dimer of the chimeras are marked on the right as single and double helices, respectively. doi:10.1371/journal.pone.0044263.g003

biological membranes, we located this hydrophobic sequence (Fig. 4A) in place of the second TM fragment of the well-characterized *Escherichia coli* inner membrane protein leader peptidase (Lep). Although of bacterial origin, Lep integrates efficiently into dog pancreas microsomes with the same topology as in *E. coli* [30] (*i.e.*, with both the N- and C-termini exposed to the luminal side of the ER membrane) and the presence of its first TM segment together with the cytoplasmic P1 domain (Figure 4B) is sufficient for proper targeting of chimeric proteins to the eukaryotic membrane [30,31]. An engineered glycosylation site placed at the C-terminal P2 domain is glycosylated efficiently upon correct insertion into the microsomal membrane (Fig. 4B), serving as a reporter to distinguish between a luminal (glycosylated) and a cytoplasmic (unglycosylated) location. Glycosylation of the molecule results in an increase in molecular mass of about 2.5 kDa relative to the observed molecular mass of Lep expressed in the absence of microsomes. The efficiency of glycosylation of Lep under standard conditions is 80–90% [31,32]. The strength of the Lep system is that it provides a comparative scale for the energetic cost of inserting a broad range of model and actual TM sequences into biological membranes, closely mimicking the *in vivo* situation.

The wild-type sequence of GpA TM segment efficiently inserts into the ER-derived microsomal membranes, while I85D mutation severely diminished membrane insertion capacity (Fig. 4C). On the contrary, L89K mutation allowed efficient insertion (Fig. 4C,

lane 6). The different effect observed for these two mutants can be explained by differences in amino acid side chain size and the position of the residue in relation to the midpoint of the TM sequence (Fig. 4A). Hence, in the case of L89K, the longer side chain of this cationic amino acid and its proximity to the membrane interface compared to I85D may allow the hydrophilic moiety of the lysine residue to snorkel, that is, to approach its ϵ -amino group toward the interfacial and aqueous region, close to the negatively charged phospholipid head groups. Next, a construct with an Asp-Lys pair at the same positions (double mutant I85D/L89K) was glycosylated somewhat more efficiently than the I85D construct (Fig. 4C, lanes 4 and 8), supporting the idea that an intrahelical salt-bridge or hydrogen bond interactions between Lys and Asp side chains located on the same face of a TM helix can facilitate its insertion into biological membranes by reducing the free energy of membrane partitioning, as previously suggested in a similar system [9]. Furthermore, the predicted insertion frequencies from the biological hydrophobicity scale [2,5] for these mutants using the ΔG Prediction Server v1.0 (<http://dgpred.cbr.su.se/>) are shown in Table 1. In this algorithm, the predicted insertion frequency comes from the apparent free-energy difference (ΔG_{app}) from insertion into ER membranes. Since very low and very high insertion efficiencies cannot be accurately measured, ΔG_{app} values outside the interval ± 1.5 kcal/mol are only qualitative. The positive value of ΔG_{app} predicted

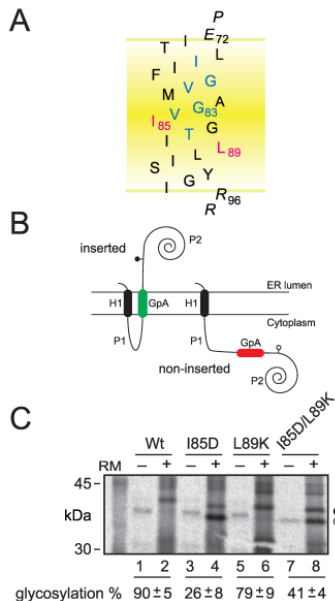


Figure 4. Insertion of GpA-derived segments into microsomal membranes. (A) Model of the GpA TM wild-type sequence. GpA residues involved in dimer formation are blue, the hydrophobic residues replaced to ionizable residues are magenta, and flanking residues are shown in italic. (B) Membrane topology of Lep chimeras. The second TM segment of Lep was replaced by the GpA TM amino acid sequence (gray). The glycosylation acceptor site located in the beginning of the P2 domain will be modified only if GpA-derived TM sequence inserts into the membrane. (C) *In vitro* translation in the presence of rough microsomal membranes (RM). Proper insertion of the GpA-derived TM sequences results in an increase in molecular mass of about 2.5 kDa relative to the observed molecular mass of the proteins expressed in the absence of microsomes. Bands of nonglycosylated and glycosylated proteins are indicated by white and black dots, respectively. Average \pm s.d. of glycosylation results from four independent experiments are shown at the bottom. doi:10.1371/journal.pone.0044263.g004

that the tested-sequence is not TM. The high negative value for the GpA wild-type sequence agrees with our experimentally measured glycosylation data showing the highest insertion efficiency. A closer analysis of the output data highlighted I85D mutation as precluding TM disposition. Hence, replacing Ile85 with aspartic acid reduced ΔG_{app} by almost 2 kcal/mol (Table 1), which correlates with our lowest glycosylation efficiency. However, replacing Leu89 with lysine has a lower energy cost (ΔG_{app} close to 0) that is reflected by a higher insertion level (Fig. 4C). Finally, the double mutant I85D/L89K results in the highest predicted penalty for TM disposition, whereas experimentally we find no evidence

that GpA TM segment is significantly compromised by the presence of two polar/ionizable residues. Such phenomena points towards an intra-helical interaction between the ionizable residues and should be taken into account to improve TM prediction algorithms.

Finally, the effect of ionizable residues in TM packing in bacterial cytoplasmic membranes was assessed using the ToxCAT assay [33]. This assay uses a chimeric construct composed of the ToxR N-terminal transcriptional activation domain [34] fused to the GpA TM segment and a C-terminal maltose binding protein (MBP) domain (Fig. 5A). TM-mediated dimerization of the chimera in the *E. coli* inner membrane results in transcriptional activation of a reporter gene encoding chloramphenicol acetyltransferase (CAT), with the level of CAT protein expression indicating the strength/intensity of TM helix-helix interactions. After transformation of these ToxCAT constructs into *E. coli* NT326 cells, we tested the ability of the wild-type and mutant fusion proteins carrying ionizable residues to complement the *malE* phenotype of the NT326 strain by growing each construct on plates containing maltose as the sole carbon source. Cells containing a construct that lack a TM segment do not grow (*pecKAN*), but the wild-type and all point mutants support growth on maltose (Fig. 5B), indicating that the MBP domains of these chimeric proteins are properly targeted to the periplasm of the NT326 cells. Consequently, the expected topology (Fig. 5A) is being achieved by these proteins, in agreement with GpA wild-type and point mutants in ToxR [35] and ToxCAT [33] assays. Dimerization of wild-type and mutant sequences carrying ionizable residues was assessed along with a GpA point mutant (G83I) that disrupts homodimerization as negative control. The I85D mutant was found to dimerize in this system to about 35% of the level shown by wild-type GpA (Fig. 5C). Interestingly, L89D mutant, which precludes dimer formation in the presence of SDS micelles (Fig. 3C), appears to retain some dimerization capacity ($21 \pm 4\%$, normalized dimerization), which highlights the influence of the specific lipid environment during the assembly of TM segments [36]. Nevertheless, differences in TM segment length and flanking residues sequences (see Fig. S1) may alter the dimerization process in the two systems, which are difficult to rationalize. Mutation of Leu89 to lysine (L89K) had a smaller effect on TM dimerization, and double mutant I85D/L89K still retained some dimerization capacity (Fig. 5C). In agreement with these data, recent molecular dynamics simulations suggested that a lysine residue outside the contact interface could exert a significant influence on TM helix association affinity of the bacteriophage M13 major coat protein because the extent of their burial in the membrane could be different in monomers and dimers [37]. Together, our data indicate that the presence of ionizable residues does not preclude membrane insertion and allows dimer formation in bacterial cells.

Conclusions

Ionizable amino acid residues are functionally and/or structurally important residues in membrane proteins. Therefore, although the insertion of such residues into the membrane hydrophobic core may be energetically unfavourable, there is often a functional and/or structural necessity to accommodate them. In the light of our experiments it can be concluded that nonpolar to ionizable point substitutions at specific positions away from the dimer interface ('lipid facing') in combination with a N-terminal GxxxG motif does not preclude neither the dimerization process nor TM helix insertion, while point mutations of nonpolar (or polar nonionizable) to ionizable residues in the 'helix facing',

Table 1. Thermodynamic cost of GpA-derived TM segments integration.

GpA-derived region	AG _{pred}	Glycosylation % (measured)	Sequence
Wt	-1.646	90±5	ITLLIFGV MAGVIGT ITLLISYGI
I85D	+ 0.413	26±8	ITLLIFGV MAGVIGT D GTLLISYGI
L89K	+ 0.112	79±9	ITLLIFGV MAGVIGT K LISYGI
I85D/L89K	+ 2.561	41±4	ITLLIFGV MAGVIGT DGT KLISYGI

The predicted (AG_{pred}) energetic cost in kcal/mol of inserting versions of the GpA TM spanning region estimated using the biological hydrophobicity scale [2,5] are provided solely for the basis of comparison. Negative AG_{pred} values are indicative of TM disposition, while positive values indicate non-TM disposition. Mutated residues at positions 85 and 89 are shown in bold. doi:10.1371/journal.pone.0044263.t001

e.g. I91D/E, I91K/R, or T87D strongly compromise dimer formation. These notions need to be considered if we are to develop a predictive understanding of TM helix interactions in membrane proteins.

Materials and Methods

Helix data set

All α -helical membrane proteins deposited in the MPTOPO database (last updated on January 19th, 2010) [20], and thus with known membrane insertion topology, were selected. The initial set was further filtered by: (i) removing any entry of unknown structure as based on the MPTOPO entry classification (i.e., keeping only entries described as "3D_helix" and "1D_helix"); and (ii) removing redundant pairs at 80% sequence identity by applying the *cd-hit* program [38]. The final data set of TM helices contained 170 non-redundant structures, 837 TM helices, and 20,079 amino acids. Furthermore, to properly analyze the amino acid propensities in single membrane spanning TM helices, we discarded any helix shorter than 17 amino acids or larger than 38 amino acids. The resulting TM data subset contained 792 TM helices, and 19,356 amino acids.

Plasmid constructs

Construction of the plasmids encoding the His-tagged chimeric proteins (SN/GpA) have been described [13,39]. Mutations at the TM fragment of GpA were obtained by site-directed mutagenesis using the QuikChange site directed mutagenesis kit (Stratagene, La Jolla, California). Introduction of the TM segment from GpA into the *Lep* sequence was described elsewhere [16]. The ToxCAT vector pccKAN, and the derivatives carrying the TM domain of GpA (pccGpA) and a disruptive GpA mutant (pccGpA-G83D) fused to the ToxR transcription activator and to maltose-binding protein (MBP) were described previously [33]. All mutants were confirmed by DNA sequencing.

Protein expression and purification

Overexpression and purification of His-tagged SN/GpA constructs from transformed *Escherichia coli* BL21 (DE3) cells was performed as described [40]. *In vitro* transcription/translation of *Lep*-derived constructs was done in the presence of reticulocyte lysate and [³⁵S]-labeled amino acids as described [16].

Charged Residues in Transmembrane Segments

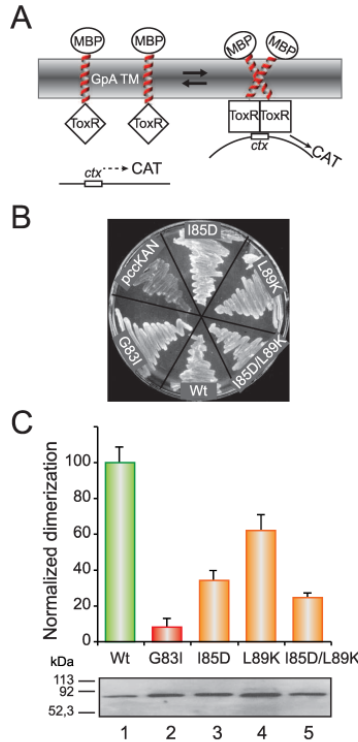


Figure 5. Dimerization in *E. coli* membranes. (A) Schematic representation of the ToxCAT assay. ToxR domains (squares) can activate transcription of the reporter gene (CAT) if brought together by the GpA-derived TM domains (right). The maltose binding protein domain (ellipses) helps direct the insertion of the construct into the membrane, complements the *malE* mutation in the host cells, and serves as an epitope for quantifying the expression level of fusion protein. (B) Complementation assays for wild-type and selected mutant ToxR(GpA)MBP fusion constructs. NT326 cells (*malE* deficient) carrying various constructs were streaked on a plate with maltose as the sole carbon source and grown for three days at 37 °C. All ToxR(GpA)MBP chimeras permit growth of NT326 cells on maltose, while control transformants (pccKAN) do not. (C) Normalized dimerization of the indicated TM domain variants as measured by CAT-ELISA relative to the wild-type GpA TM domain. Bars for intermediate dimerization and non-dimerizing levels are colored orange and red, respectively. Average \pm s.d. of results from four independent experiments are shown. Levels of expression of selected ToxR(GpA)MBP constructs as analyzed by immunoblotting are shown at the bottom. doi:10.1371/journal.pone.0044263.g005

SDS-PAGE analysis

Purified SN/GpA proteins were loaded onto SDS 12% polyacrylamide mini-gels. The loading buffer contained 2% (w/v) SDS, and samples were boiled for five minutes prior to electrophoresis. Gels were stained with Coomassie blue, and the percentages of monomer and dimer were estimated with a ImageQuantTM LAS 4000mini Biomolecular Imager (GE Healthcare). Gels with radioactive Lep-derived samples were dried at 80°C and scanned using a Fuji FLA-3000 phosphor-imager using the ImageGauge software.

ToxCAT methods

Plasmids encoding ToxR(GpA)MBP chimeras were transformed into *Escherichia coli* NT326 cells (kindly provided by D. M. Engelman) and plated onto Luria Bertani (LB) plates (with 50 µg/ml ampicillin, 25 µg/ml streptomycin); colonies were inoculated into LB medium (with 50 µg/ml ampicillin, 25 µg/ml streptomycin), and glycerol stocks were made at $A_{600} \approx 0.2$ and stored at -80°C. LB cultures (with 50 µg/ml ampicillin, 25 µg/ml streptomycin) were inoculated from frozen glycerol stocks and grown at 37°C until approximately $A_{120} \approx 0.6$, when culture densities were equalized by dilution into fresh culture tubes, and $6.0 A_{120}$ units of cells were harvested by centrifugation and washed with 0.4 ml of sonication buffer (25 mM Tris-HCl, 2 mM EDTA, pH 8.0) [41]. Cells were then resuspended in 0.6 ml of sonication buffer and lysed by probe sonication. After removing an aliquot (20 µl) for Western blot analysis, the remaining lysate was clarified by centrifugation at 13,000×g, and the supernatant was stored on ice until the spectrophotometric assay was performed. All constructs conferred the ability to grow on maltose plates to the *malE* strain NT326, which indicates that proper membrane insertion of the ToxR(GpA)MBP fusion protein has occurred [33]. For maltose complementation assays, *E. coli* NT326 cells expressing ToxR(GpA)MBP constructs were streaked on M9 minimal media plates containing 0.4% maltose as the only carbon source, and incubated for 3 days at 37°C. All constructs showed similar expression levels of ToxR(GpA)MBP fusion protein as determined by Western blot using an anti-MBP antibody. The

self-association ability of the TM domain triggers expression of a chloramphenicol transferase (*cat*) gene reporter and production of CAT protein can be quantified by a CAT-ELISA kit (Roche Diagnostics) [42]. CAT measurements and construct expression measurements were performed in at least triplicate and were normalized for the relative expression level of each construct using Western blotting [43]. All constructs showed similar expression levels of ToxR(GpA)MBP fusion proteins as determined by Western blot using an anti-MBP antibody. For Western blots samples were mixed with equal volumes of 2× SDS-PAGE sample buffer heated to 95°C for 10 min, separated on 10% (w/v) polyacrylamide mini-gels, blotted onto nitrocellulose membranes, and blocked in skim milk. ToxR(GpA)MBP chimera were detected with biotinylated anti-MBP primary antibody (NEB) and visualized with streptavidin-horseradish peroxidase conjugate and ECL reagent (GE Healthcare). Bands were quantified with an ImageQuantTM LAS 4000mini Biomolecular Imager (GE Healthcare).

Supporting Information

Figure S1 TM segments and flanking residues sequences. The primary sequences of the GpA TM regions used in both the SDS-PAGE and ToxCAT analyses are shown. Hydrophobic residues are boxed in yellow and flanking residues are highlighted (italic). (EPS)

Acknowledgments

We thank D.M. Engelman (Yale University) for ToxCAT vectors and *E. coli* NT326 cells, and M.D. Oliver for preliminary results.

Author Contributions

Conceived and designed the experiments: MO IM. Performed the experiments: MB-P. Analyzed the data: MB-P CB-D MAM-R CA IM. Contributed reagents/materials/analysis tools: MB-P CB-D. Wrote the paper: MAM-R IM.

References

- Popot JL, Engelman DM (1990) Membrane protein folding and oligomerization - The 2-stage model. *Biochemistry* 29: 4031-4037.
- Hessa T, Kim H, Bihlmaier K, Lindin C, Boekel J, et al. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433: 377-381.
- Baeza-Delgado C, Marti-Renom MA, Mingarro I (accepted) Structure-based statistical analysis of transmembrane segments. *Eur Biophys J*. DOI 10.1007/s00249-012-0813-9
- Martinez-Gil L, Perez-Gil J, Mingarro I (2003) The surfactant peptide KLF4 sequence is inserted with a transmembrane orientation into the endoplasmic reticulum membrane. *Biophys J* 95: L36-38.
- Hessa T, Meindl-Beinker NM, Bernsel A, Kim H, Sato Y, et al. (2007) Molecular code for transmembrane-helix recognition by the SecE translocon. *Nature* 450: 1026-1030.
- Zhou FX, Merianos HJ, Brunger AT, Engelman DM (2001) Polar residues drive association of polyeuclycine transmembrane helices. *Proc Natl Acad Sci U S A* 98: 2250-2255.
- Gratkowski H, Lear JD, DeGrado WF (2001) Polar side chains drive the association of model transmembrane peptides. *Proc Natl Acad Sci U S A* 98: 880-885.
- Hermansson M, von Heijne G (2003) Inter-helical Hydrogen Bond Formation During Membrane Protein Integration into the ER Membrane. *Journal of Molecular Biology* 334: 803-809.
- Chin CX, von Heijne G (2000) Charge pair interactions in a model transmembrane helix in the ER membrane. *J Mol Biol* 303: 1-5.
- DeGrado WF, Gratkowski H, Lear JD (2003) How do helix-helix interactions help determine the folds of membrane proteins? Perspectives from the study of homo-oligomeric helical bundles. *Protein Sci* 12: 647-665.
- Mackenzie KR (2006) Folding and stability of alpha-helical integral membrane proteins. *Chem Rev* 106: 1931-1977.
- Lenmon MA, Flanagan JM, Treutlein HR, Zhang J, Engelman DM (1992) Sequence specificity in the dimerization of transmembrane α -helices. *Biochemistry* 31: 12719-12725.
- Mingarro I, Whitley P, Lenmon MA, von Heijne G (1996) Ala-insertion scanning mutagenesis of the glycoporphin A transmembrane helix. A rapid way to map helix-helix interactions in integral membrane proteins. *Protein Sci* 5: 1339-1341.
- MacKenzie KR, Prestegard JH, Engelman DM (1997) A transmembrane helix dimer: Structure and implications. *Science* 276: 131-133.
- Smith SO, Song D, Shekar S, Groesbeck M, Ziliox M, et al. (2001) Structure of the transmembrane dimer interface of glycoporphin A in membrane bilayers. *Biochemistry* 40: 6553-6558.
- Orzaez M, Salgado J, Gimenez-Giner A, Perez-Paya E, Mingarro I (2004) Influence of proline residues in transmembrane helix packing. *J Mol Biol* 335: 631-640.
- Russ WP, Engelman DM (2000) The GxxxG motif: a framework for transmembrane helix-helix association. *J Mol Biol* 296: 911-919.
- Senes A, Engel DE, DeGrado WF (2004) Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr Opin Struct Biol* 14: 465-479.
- Cymer F, Veerappan A, Schneider D (2012) Transmembrane helix-helix interactions are modulated by the sequence context and by lipid bilayer properties. *Biochimica et Biophysica Acta* 1818: 963-973.
- Jayasinghe S, Hristova K, White SH (2001) MPOpop: A database of membrane protein topology. *Protein Sci* 10: 455-458.
- Illergard K, Kauko A, Eilsson A (2011) Why are polar residues within the membrane core evolutionary conserved? *Proteins* 79: 79-91.
- Wong WC, Maurer-Stroh S, Eisenhaber F (2011) Not all transmembrane helices are born equal: Towards the extension of the sequence homology concept to membrane proteins. *Biology direct* 6: 57.

23. Orzaez M, Lukovic D, Abad C, Perez-Paya E, Mingarro I (2005) Influence of hydrophobic matching on association of model transmembrane fragments containing a minimised glycophorin A dimerisation motif. *FEBS Lett* 579: 1633–1638.
24. Duong MT, Jaszewski TM, Fleming KG, MacKenzie KR (2007) Changes in apparent free energy of helix-helix dimerization in a biological membrane due to point mutations. *Journal of molecular biology* 371: 422–434.
25. Herrmann JR, Fuchs A, Panitz JC, Eckert T, Unterreimter S, et al. (2010) Ionic interactions promote transmembrane helix-helix association depending on sequence context. *Journal of molecular biology* 396: 452–461.
26. Caputo GA, London E (2004) Position and ionization state of Asp in the core of membrane-inserted alpha helices control both the equilibrium between transmembrane and nontransmembrane helix topology and transmembrane helix positioning. *Biochemistry* 43: 8794–8806.
27. Tulumello DV, Deber CM (2009) SDS micelles as a membrane-mimetic environment for transmembrane segments. *Biochemistry* 48: 12096–12103.
28. Monne M, Nilsson I, Johansson M, Elmhed N, von Heijne G (1998) Positively and negatively charged residues have different effects on the position in the membrane of a model transmembrane helix. *J Mol Biol* 284: 1177–1183.
29. Garcia-Saez AJ, Mingarro I, Perez-Paya E, Salgado J (2004) Membrane-insertion fragments of Bel-1, Bax, and Bid. *Biochemistry* 43: 10930–10943.
30. Gavélin G, Sakaguchi M, Andersson H, von Heijne G (1997) Topological rules for membrane protein assembly in eukaryotic cells. *J Biol Chem* 272: 6119–6127.
31. Vilár M, Sauri A, Monne M, Marcos JF, von Heijne G, et al. (2002) Insertion and topology of a plant viral movement protein in the endoplasmic reticulum membrane. *J Biol Chem* 277: 23447–23452.
32. Johansson M, Nilsson I, von Heijne G (1993) Positively charged amino acids placed next to a signal sequence block protein translocation more efficiently in *Escherichia coli* than in mammalian microsomes. *Mol Gen Genet* 239: 251–256.
33. Russ WP, Engelman DM (1999) TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc Natl Acad Sci USA* 96: 863–868.
34. Kolmar H, Hennecke F, Gotze K, Janzer B, Vogt B, et al. (1995) Membrane insertion of the bacterial signal transduction protein ToxR and requirements of transcription activation studied by modular replacement of different protein substructures. *EMBO J* 14: 3895–3904.
35. Langosch D, Brosig B, Kolmar H, Fritz HJ (1996) Dimerization of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. *J Mol Biol* 263: 525–530.
36. Martinez-Gil L, Sauri A, Martt-Renom MA, Mingarro I (2011) Membrane protein integration into the ER. *FEBS J* 278: 3846–3853.
37. Zhang J, Lazaridis T (2009) Transmembrane helix association affinity can be modulated by flanking and noninterfacial residues. *Biophysical journal* 96: 4418–4427.
38. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26: 680–682.
39. Lemmon MA, Flanagan JM, Hunt JF, Adair BD, Bormann BJ, et al. (1992) Glycophorin A dimerization is driven by specific interactions between transmembrane 9-helices. *J Biol Chem* 267: 7683–7689.
40. Orzaez M, Perez-Paya E, Mingarro I (2000) Influence of the C-terminus of the glycophorin A transmembrane fragment on the dimerization process. *Protein Sci* 9: 1246–1253.
41. Sulistijo ES, Jaszewski TM, MacKenzie KR (2003) Sequence-specific dimerization of the transmembrane domain of the “BH3-only” protein BNIP3 in membranes and detergent. *The Journal of biological chemistry* 278: 51950–51956.
42. Vilár M, Charalampopoulos I, Kenchappa RS, Simi A, Karaca E, et al. (2009) Activation of the p75 neurotrophin receptor through conformational rearrangement of disulphide-linked receptor dimers. *Neuron* 62: 72–83.
43. Johnson RM, Rath A, Melynk RA, Deber CM (2006) Lipid solvation effects contribute to the affinity of Gly-xxx-Gly motif-mediated helix-helix interactions. *Biochemistry* 45: 8507–8515.
44. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Research* 40: D370–376.

ANNEX VI

Prado-Martinez J., Hernando-Herraez I., Lorente-Galdos B., Dabad M., Ramirez O., Baeza-Delgado C., Morcillo-Suarez C., Alkan C., Hormozdiari F., Raineri E., Estellé J., Fernandez-Callejo M., Valles M., Ritscher L., Schöneberg T., de la Calle-Mustienes E., Casillas S., Rubio-Acero R., Melé M., Engelken J., Caceres M., Gomez-Skarmeta J. L., Gut M., Bertranpetit J., Gut I. G., Abello T., Eichler E. E., Mingarro I., Lalueza-Fox C., Navarro A. and Marques-Bonet T. (2013) ‘**The genome sequencing of an albino Western lowland gorilla reveals inbreeding in the wild**’, *BMC genomics*, 14(1), p. 1.

These results are part of the background of this Thesis. Carlos Baeza contributed by performing the membrane integration experiments shown in figure 2.



The genome sequencing of an albino Western lowland gorilla reveals inbreeding in the wild

Prado-Martinez *et al.*



Prado-Martinez *et al.* *BMC Genomics* 2013, **14**:363
<http://www.biomedcentral.com/1471-2164/14/363>

RESEARCH ARTICLE

Open Access

The genome sequencing of an albino Western lowland gorilla reveals inbreeding in the wild

Javier Prado-Martinez¹, Irene Hernando-Herraez¹, Belen Lorente-Galdos¹, Marc Dabad¹, Oscar Ramirez¹, Carlos Baeza-Delgado², Carlos Morcillo-Suarez^{1,3}, Can Alkan^{4,5}, Fereydoun Hormozdiari⁴, Emanuele Raineri⁶, Jordi Estellé^{6,7}, Marcos Fernandez-Callejo¹, Mònica Valles¹, Lars Ritscher⁸, Torsten Schöneberg⁸, Elisa de la Calle-Mustienes⁹, Sònia Casillas¹⁰, Raquel Rubio-Acero¹⁰, Marta Melé^{1,11}, Johannes Engelken^{1,12}, Mario Caceres^{10,13}, Jose Luis Gomez-Skarmeta⁹, Marta Gut⁶, Jaume Bertranpetit¹, Ivo G Gut⁶, Teresa Abello¹⁴, Evan E Eichler^{4,15}, Ismael Mingarro², Carles Lalueza-Fox¹, Arcadi Navarro^{1,3,13,16} and Tomas Marques-Bonet^{1,13*}

Abstract

Background: The only known albino gorilla, named *Snowflake*, was a male wild born individual from Equatorial Guinea who lived at the Barcelona Zoo for almost 40 years. He was diagnosed with non-syndromic oculocutaneous albinism, i.e. white hair, light eyes, pink skin, photophobia and reduced visual acuity. Despite previous efforts to explain the genetic cause, this is still unknown. Here, we study the genetic cause of his albinism and making use of whole genome sequencing data we find a higher inbreeding coefficient compared to other gorillas.

Results: We successfully identified the causal genetic variant for *Snowflake*'s albinism, a non-synonymous single nucleotide variant located in a transmembrane region of *SLC45A2*. This transporter is known to be involved in oculocutaneous albinism type 4 (OCA4) in humans. We provide experimental evidence that shows that this amino acid replacement alters the membrane spanning capability of this transmembrane region. Finally, we provide a comprehensive study of genome-wide patterns of autozygosity revealing that *Snowflake*'s parents were related, being this the first report of inbreeding in a wild born Western lowland gorilla.

Conclusions: In this study we demonstrate how the use of whole genome sequencing can be extended to link genotype and phenotype in non-model organisms and it can be a powerful tool in conservation genetics (e.g., inbreeding and genetic diversity) with the expected decrease in sequencing cost.

Keywords: Gorilla, Albinism, Inbreeding, Genome, Conservation

Background

The only known albino gorilla named *Snowflake* (Figure 1) was a male wild-born Western lowland gorilla (*Gorilla gorilla gorilla*) from Equatorial Guinea. He was brought to the Barcelona Zoo in 1966 at young age [1], where he gained popularity worldwide. *Snowflake* presented the typical properties of albinism as seen in humans: white hair, pink skin, blue eyes, reduced visual acuity and photophobia. Given his lack of pigmentation and thus

reduced protection from UV light, the aged albino gorilla developed squamous-cell carcinoma that led to his euthanasia in 2003 [2].

Snowflake was diagnosed with non-syndromic albinism (Oculocutaneous Albinism, OCA). This is a group of Mendelian recessive disorders characterized by the generalized reduction of pigmentation in skin, hair, and eyes. Pigmentation is determined by melanin compounds, which are produced in melanocytes and are transported via melanosomes into keratinocytes of the epidermis and hair follicles. It has been widely studied in humans and four genes are found to be causative of this disorder: (i) OCA1A/B (MIM 203100,606952) are caused by mutations in the gene *TYR* (*Tyrosinase*) (ii) mutations in the *OCA2* gene (previously known as *P-gene*) can cause OCA2

* Correspondence: tomas.marques@upf.edu

¹Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Barcelona 08003, Spain

¹³Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain

Full list of author information is available at the end of the article



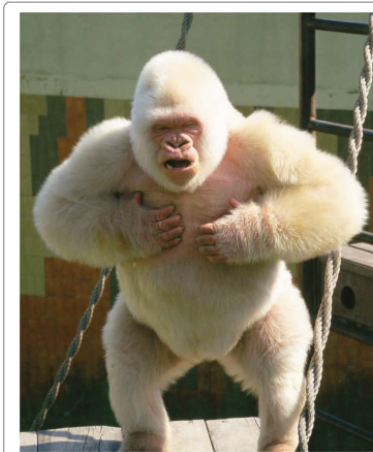


Figure 1 *Snowflake*, the only known albino gorilla. This Western lowland gorilla was wild-born in Equatorial Guinea and he presented the typical characteristics of oculocutaneous albinism.

phenotype (MIM 203200) (iii) mutations in *TYRP1* cause OCA3 (MIM 203290) and (iv) OCA4 (MIM 606574) is caused by mutations in *SLC45A2* (formerly known as *MATP* and *AIM1*) [3]. Tyrosinase and *TYRP1* are critical in the melanin synthesis pathway whereas P protein (OCA2) and *SLC45A2* are involved in melanocytes maintenance or formation.

A previous study tried to assess whether the causative mutation of *Snowflake's* albinism was located in the *TYR* gene but no causative mutation was found [4]. Here, we make use of whole genome sequencing to provide a better characterization of all known genes related to albinism to try to ascertain the genetic component causing this phenotype and to study genome wide patterns that can help the field of conservation genetics. Most of the knowledge about ecology, population dynamics, demography and social behavior about gorillas has been collected from mountain gorillas (*Gorilla beringei beringei*) and until recently this has not expanded to Western lowland gorillas [5,6]. This effort has been extremely helpful to improve our knowledge and conservation of this endangered species. With the development of conservation genetics we have gained insights into population genetics [7], demographic history [8] and group relationships through the usage of both microsatellites and mitochondrial markers. The main difficulty of these studies

is that non-invasive samples such as hair or feces cannot provide DNA of high quality.

Here, using high quality DNA and next-generation sequencing, we have studied for the first time the whole genome of a wild born Western lowland gorilla. It is important to stress that previous whole-genome sequencing projects of Western lowland gorillas, involved captive-born individuals, Kamilah [9] and Kwan [10], individuals that do not belong to a wild population as it has been recently studied with microsatellite markers [11]. Studying this unique albino gorilla, we find the first evidence of inbreeding in wild Western lowland gorillas.

Results

We sequenced the genome of *Snowflake* at 18.7× effective coverage using the Illumina GAIIX platform (114 bp paired-end reads). We aligned the reads to the reference human genome (NCBI build 37) using GEM [12], and used samtools [13] to identify single nucleotide variants (SNVs) (Methods). We found 73,307 homozygous non-synonymous *Snowflake's* mutations compared to the human reference genome. Out of those, 20 were found within candidate genes for albinism (OCA related genes), but a single mutation was private compared to two other sequenced gorillas (Additional file 1: Tables S1 and S2) [9,10]. This substitution is located in the last exon of the *SLC45A2* gene at the position hg19: chr5_33944794_C/G and it causes a substantial amino acid change, Glycine to Arginine, (pGly518Arg) in a predicted transmembrane region of the protein. We then resequenced this mutation using capillary sequencing and it was confirmed as homozygous in *Snowflake* and heterozygous in all five tested non-albino offspring, as expected in Mendelian recessive disorders. To rule out the possible participation of other candidate genes, we also looked for structural variants that may be disrupting other genes related to pigmentation. We applied computational methods based on paired-end and split read approaches to detect genomic deletions (Methods), followed by experimental validation using array-comparative genomic hybridization (aCGH). We identified 1,390 validated deletions totaling to 9.5 Mbps, a similar proportion of the genome compared to previous reports [5] (Additional file 1: Table S3). These deletions overlap completely with 36 RefSeq transcripts and partially (>10%) with 660 transcripts (Additional file 1: Table S4) but none of them has a direct association with albinism.

Several pieces of evidence support the hypothesis that the non-synonymous mutation found in *SLC45A2* might be responsible for *Snowflake's* albinism. First, this specific Glycine residue is conserved throughout all available vertebrate taxa (Additional file 2: Figure S1), suggesting a conserved role of this amino acid. Second, we predicted whether this amino acid change may affect the protein

structure and function based on sequence conservation and protein properties using SIFT [14] and PolyPhen-2 [15]. It is predicted as a "damaging" mutation by SIFT, and "probably damaging" by PolyPhen-2. Third, this gene was reported to be the genetic cause of albinism in several other species (e.g. mouse [16], medaka fish [17], horse [18] and chicken [19]). Last, previous reports showed that Glycine to Arginine mutations within other transmembrane regions of *SLC45A2* in humans result in severe albino phenotypes [20].

We followed up on this finding with an experimental study to determine how this amino acid substitution affects the transmembrane segment where this mutation is present. For this purpose we used a functional assay based on *Escherichia coli* inner membrane protein leader peptidase (Lep) that detects and permits accurate measurements of the apparent free energy (ΔG_{app}) of translocation-mediated integration of transmembrane helices into the endoplasmic reticulum (ER) membranes [21-23]. This procedure allows the quantification of the proper integration of the transmembrane region with the normal sequence and with the mutation. When we assayed the construct with the wild type sequence, we observed that 90% of the proteins were properly recognized for membrane insertion. However, translation of the mutant (G518R) found in *Snowflake* resulted in a significant reduction (~25%, p-value = 0.036 Mann-Whitney U test) in the membrane integration capability (Figure 2), suggesting that the replacement of a glycine by an arginine residue lowers the affinity of the transmembrane region and possibly alter the topology of the *SLC45A2* gene product.

Finally, the last piece of evidence supporting the role of the mutation in the phenotype is based on genome-wide patterns of heterozygosity in the genome of *Snowflake* (Additional file 2: Figure S2). We found that *SLC45A2* gene is located in a large run of homozygosity (40 Mb) orthologous to human chromosome 5 (Figure 3a), meaning that this allele was inside a block identical by descent, which is characteristic for Mendelian recessive disorders. The other three candidate genes are not found in any autozygous regions. Overall, we found 25 large runs of homozygosity (longer than 2 Mb), and a general reduction of heterozygosity compared to the other known genomes sequenced of the same species (Figure 3b). Some of the runs of homozygosity are particularly large, such as a continuous 68 Mb segment in chromosome 4 (Additional file 2: Figure S2). This reduction of variation might allow the emergence of certain phenotypes otherwise masked by dominance, and they could lead to inbreeding depression [24], as previously reported in chimpanzee and other primates [25].

These patterns of heterozygosity allow the estimation of the amount of autozygosity, i.e. long regions of the genome identical by descent as a result of inbreeding. The

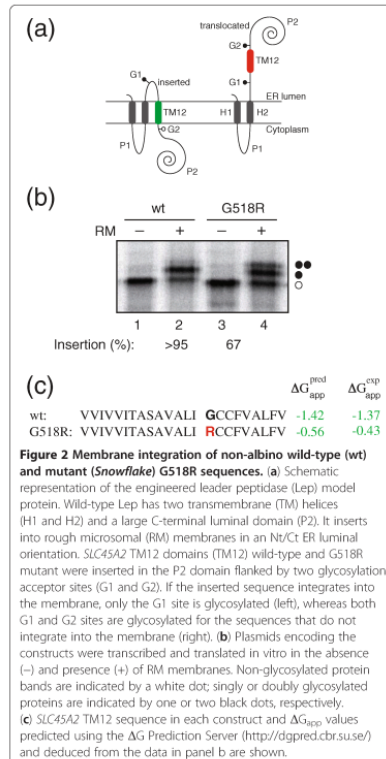
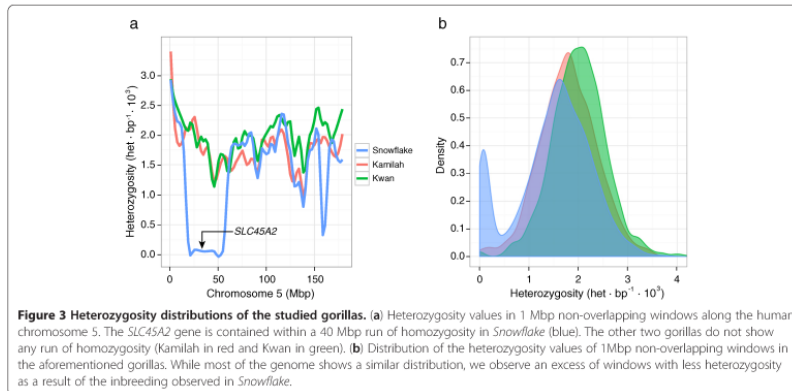


Figure 2 Membrane integration of non-albino wild-type (wt) and mutant (*Snowflake*) G518R sequences. (a) Schematic representation of the engineered leader peptidase (Lep) model protein. Wild-type Lep has two transmembrane (TM) helices (H1 and H2) and a large C-terminal luminal domain (P2). It inserts into rough microsomal (RM) membranes in an N/Ct ER luminal orientation. *SLC45A2* TM12 domains (TM12) wild-type and G518R mutant were inserted in the P2 domain flanked by two glycosylation acceptor sites (G1 and G2). If the inserted sequence integrates into the membrane, only the G1 site is glycosylated (left), whereas both G1 and G2 sites are glycosylated for the sequences that do not integrate into the membrane (right). (b) Plasmids encoding the constructs were transcribed and translated in vitro in the absence (-) and presence (+) of RM membranes. Non-glycosylated protein bands are indicated by a white dot; singly or doubly glycosylated proteins are indicated by one or two black dots, respectively. (c) *SLC45A2* TM12 sequence in each construct and ΔG_{app} values predicted using the ΔG Prediction Server (<http://dgpred.cbr.su.se/>) and deduced from the data in panel b are shown.

minimum threshold of these stretches has been estimated in humans ranging between 1-2.25 Mbp depending on the population [26]. We conservatively quantified the inbreeding coefficient in *Snowflake* based on autozygosity (F_{ROH}) of 0.118 (306 Mb) out of 2,587 Mb) compared to 0.002 and <0.001 estimated from the previously sequenced gorillas, Kamilah and Kwan respectively, using the same criteria (Methods). Assuming no previous inbreeding in any of the parents, 0.125 would correspond to parents being grandparent-grandchild, half-siblings, or uncle-niece/aunt-nephew. We performed a set of simulations that replicated the patterns of autozygosity in different pedigrees, accounting for the differential amount of recombinations derived from the number of meiosis



and the sex of the ancestors that is known to influence recombination rates [27]. These simulations reconstruct the different recombination patterns in different possible pedigrees under a random paternal transmission, and we estimated which fragments may appear as autozygous in the offspring. Finally, we compared the distribution of sizes and number of autozygous segments in *Snowflake* with the different simulated outcomes to calculate a likelihood for each case. The uncle/niece or aunt/nephew combination is the most probable scenario, although there was not a unique best statistical pedigree (Additional file 2: Figure S3C-F and Additional file 1: Table S5).

Discussion

We sequenced the whole genome of a phenotypically unique gorilla, identified and characterized the causative mutation for his albinism, and explored the origin of this trait. We found a private non-synonymous substitution in one of the candidate genes - the *SLC45A2* gene - associated with the OCA4 class of albinism. We provided several lines of evidence based on evolution, human disease and a functional assay supporting that this mutation in a transmembrane domain can modify the topology of the translated protein, therefore reinforcing its causative role in this rare case of albinism. Moreover, long runs of homozygosity in this wild born individual explain the emergence of this recessive trait through identity by descent, suggesting that inbreeding was an important factor towards the emergence of this phenotype.

We inferred that *Snowflake* was an offspring of closely related individuals supported by an inbreeding coefficient of 0.118. In general, inbreeding is avoided in the wild because the offspring within gorilla societies disperse to

other groups before maturity [28]. This is strictly true in patriarchal groups that are commonly composed of a silverback male and several females (97% of all the gorilla groups) [6] whereas in multimale groups, females can remain and have their first birth in the natal group. Multimale groups are usually composed of related males and therefore, newborn females are likely to be also related to them (commonly with relationships such as half brothers or half uncles). However, multimale groups have mainly been observed in mountain gorillas, while only two multimale groups have ever been reported in Western lowland gorillas, suggesting that they are extremely rare in these populations [6]. Therefore, it seems unlikely that a multimale group would explain the inbreeding found in this Western lowland gorilla.

Previous parentage studies in wild Western lowland gorillas have never found inbred mating, suggesting that is probably a rare behavior [29]. Despite this and considering that the observation of inbreeding in a single individual could be an extreme case, some social observations may point that inbreeding may still occur. First, gorillas seem to follow a patrilineal social structure, i.e. silverbacks are usually related to one or more nearby silverbacks [30]. Additionally, females transfer several times during their lifespan after the dispersal from their natal group [31], which may result in the arrival of a female to a new group where the silverback is related to her. Although father-daughter inbreeding is completely avoided, this hypothesis is feasible because other mating relationships, with half-brothers or even full brothers, are possible; suggesting that females do not detect consanguinity [6]. Other factors such as habitat loss, small population sizes and population fragmentation may influence the disposal of breeding

groups and therefore of unrelated silverbacks which may in turn favor inbreeding [32]. Other potential explanations are less likely; male takeovers are highly avoided and the death of a male silverback normally results in the disintegration of the group and female dispersal.

A previous study using microsatellite markers in captive gorilla populations showed that their genetic diversity is comparable to wild gorilla populations [11]. However, in our study, *Snowflake* shows different patterns of heterozygosity compared to the captive born gorillas. The gorilla studbooks show that *Kamilah* (Studbook ID: 661) is a first generation captive-born gorilla, while *Kwan* (Studbook ID: 1107) is a second generation captive-born gorilla. When we compared the heterozygosity genome-wide, we observed that *Kwan* is the gorilla with higher heterozygosity, despite we cannot rule out that this was a result of some false positives due to the lower sequencing coverage. *Kamilah* and *Snowflake* have lower heterozygosity, with the albino gorilla showing the lowest values compared to the other captive-born individuals (Figure 3b) even accounting for the regions of inbreeding. This suggests that breeding programs could result in an increase of genetic variation but a bigger sample size would be needed to systematically explore this effect.

In this particular study, we show that high throughput sequencing can be used not only to unravel the genetic mechanisms of fundamental phenotypes (including disease) in non-model organisms, but also to provide insights into conservation genetics through the detection of inbreeding of endangered species such as gorilla. However, in order to systematically explore relationships and breeding patterns from wild specimens using whole genome sequencing data, high quality DNA is required and in most field studies, non-invasive samples such as feces or hair are used, and the amount of DNA extracted from such samples precludes the application of this methodology to conservation studies. Still, it can be applied to analyze the genomes of wild born individuals in zoos where blood samples are usually taken during routine veterinary check-ups. However, sequencing technologies quickly and constantly improve, and recent developments that includes library construction with very little amounts of DNA [33] or single-cell sequencing [34,35] may allow the implementation of this kind of analyses into conservation genetics in the near future.

Conclusions

Here we make use of next-generation sequencing to study the complete genome of a wild born Western lowland gorilla genome. Using these data we have been able to identify the genetic cause of a rare phenotype --albinism-- in this non-model species and we provide several lines of evidence that reinforces this hypothesis, ranging from evolution to human disease. Moreover, we have been able to characterize that this individual was descendant of

close relatives by studying the patterns of autozygosity genome-wide, providing the first genetic evidence of inbreeding in this species. We discuss this finding from the perspective of gorilla societies and we link several pieces of information in order to provide plausible scenarios where this event could have happened. We envision that the analysis of whole genome data of endangered species will be a standard in future conservation and management studies and will make available relevant information that has been missed in previous studies.

Methods

Sequencing

We extracted DNA using phenol-chloroform from a frozen blood sample previously taken from *Snowflake* (*Gorilla gorilla gorilla*). We constructed Illumina libraries using the standard protocol with two different fragment sizes at 250 bp and 450 bp. We sequenced the genome at ~18x coverage with paired-end reads (114 nt). For comparison purposes, we analyzed the genomes of two other Western lowland Gorillas (*Kwan* [10]) and *Kamilah* [9]) (Additional file 1: Table S1). The research did not involve any experiment on human subjects or animals and for this reason no ethical approval was necessary, the blood used for the sequencing of *Snowflake* was extracted after the death of the gorilla.

Single nucleotide variants

We mapped all reads to the human reference genome (GRCh37) using GEM [12] allowing a divergence of 4% in order to capture all putative changes between human and gorilla keeping uniquely placed reads. We identified single nucleotide differences with *Samtools* [13] (v.0.1.9), and filtered out potential false positives by mapping quality and read depth (based on the different sequencing depths for the samples).

Copy number variation discovery and validation

We assessed genomic structural variants compared to the human genome using a combination of paired-end and split read methods to provide an initial catalog of potential deletions. In order to validate these regions using an independent approach, we further analyzed them with Array-comparative genomic hybridization. Finally, we reported the regions that revealed a variation and that were concordant using both methodologies (Additional file 2).

Inbreeding

To estimate the degree of heterozygosity in the genomic of *Snowflake*, we divided the genome into 1 Mbp non-overlapping windows and calculated the heterozygous positions per Kbp. To avoid divergent outliers in the estimates of each window, we removed all regions that overlapped more than 40% with duplications, and we corrected the

number of heterozygous positions by the remaining effective bases of the windows. To calculate the inbreeding coefficient, we conservatively considered regions with a loss of heterozygosity when at least two consecutive 1 Mbp windows showed a reduction of heterozygosity.

Inbreeding simulations

Computer simulations were run in order to infer the family history that may be responsible for the pattern of homozygous fragments found in the genome of *Snowflake*. We considered all the possible pedigrees: half-siblings, aunt/nephew, uncle/niece, grandfather/granddaughter and grandmother/grandson. A total of 10 models were defined to account for the different pedigree combinations that can generate the above parental origins (Additional file 2: Figure S3).

For each pedigree model, 10,000 simulated "Snowflakes" were created considering no relationship among founding members. We used different rates of recombination for males and females (8.9×10^{-9} crossovers/nucleotide for males, and 1.4×10^{-8} for females) following empirical data [27].

The simulations were performed using a Java program written ad hoc for this particular purpose. For every founder individual in the pedigree two sets of chromosomes are generated containing different alleles (zero inbreeding is assumed among all founders). In descendant individuals, chromosomes are generated by crossing over parental chromosomes and randomly passing one out of the two present in each parent to the offspring. Descendant individuals will have a mix of founder alleles in their chromosomes. Due to the inbred structure of the pedigrees, the simulated *Snowflake* is expected to present regions where both chromosomes have the same allele originated from a single founder individual.

Number and length distribution of homozygous fragments resulting from each model were compared with the actual values in the genome of *Snowflake* (Additional file 2: Figures S4 and S5). For each model, homozygous fragments obtained were classified according to their length into 5 Mbp bins and a multinomial distribution was defined using the resulting counts. The probability of these distributions of generating the actual *Snowflake* counts was used as a measure of likelihood for each model (Additional file 1: Table S5). To make the obtained data compatible with experimental data, we removed segments smaller than 2 Mbp and merged the segments separated by gaps smaller than 500 Kbp.

Mutant membrane integration

Wild type and *Snowflake* constructs in pGEM1 were transcribed and translated in the TNT[®] SP6 Quick Coupled System from Promega. DNA template (~75 ng), 1 μ l of [35S]Met/Cys (5 μ Ci), and 1 μ l of dog pancreas

RMs were added to 5 μ l of lysate at the start of the reaction, and samples were incubated for 90 min at 30°C. The translation reaction mixture was diluted in 5 volumes of phosphate buffer saline (pH 7.4). Subsequently, membranes were collected by layering the supernatant onto a 50 μ l sucrose cushion and centrifuged at 100,000 \times g for 20 min at 4°C in a Beckman tabletop ultracentrifuge with a TLA-45 rotor. Finally, pellets were analyzed by SDS-PAGE, and gels were visualized on a Fuji FLA3000 phosphorimager using the ImageGauge. (Additional file 2).

Additional files

Additional file 1: Contains the supplementary tables. **Table S1:** Summary of the samples used in this study. **Table S2:** Non-synonymous mutations found in OCA genes compared to human genes. **Table S3:** Summary of deletions found in *Snowflake* using different methodologies. **Table S4:** List of transcripts affected by deletions. **Table S5:** Likelihood values in the paternity simulations.

Additional file 2: Contains detailed explanation on some methods and supplementary figures.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JP-M and TM-B designed the study and drafted the manuscript. JP-M, BL-G, MD, CM-S, CA, FH, ER, JE, MF-C, SC, MM, RR and MC conducted bioinformatics analysis. IH-H, OR, CB-D, MV, LR, TS, EC-M, JE, JLG-S, MG, IGG and IM performed experiments. TA, JB, IM, EEE, CL-F and AN helped to write the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Jordi Camps, Luis Alberto Perez Jurado and Lorna Brockopp for technical help. The Spanish Government for grants BFU2010-14839 to JLG-S, Spanish Government and FEDER for grants BFU2009-13409-C02-02 and BFU2012-38236 to AN and JP-M, BFU2012-39482 to IM, and BFU2011-28549 to TM-B. The Andalusian Government for grants CSD2007-00008 and CVI-3488, supported by FEDER to JLG-S. The Barcelona Zoo (Ajuntament de Barcelona) for an award to JP-M. EEE is an investigator with the Howard Hughes Medical Institute. The European Community for an ERC Starting Grant (StG_20091118) to TM-B.

Author details

¹Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Barcelona 08003, Spain. ²Departament de Bioquímica i Biologia Molecular, Universitat de València, Burjassot E-46100, Spain. ³Instituto Nacional de Bioinformática, UPF, Barcelona, Spain. ⁴Department of Genome Sciences, University of Washington, 3720 15th AVE NE, Seattle, WA 98195, USA. ⁵Department of Computer Engineering, Bilkent University, Ankara, Turkey. ⁶Centro Nacional de Análisis Genómico, PCB, Barcelona 08028, Spain. ⁷Current address: INRA, UMR1313 GABI, Jouy-en-Josas, France. ⁸Institute of Biochemistry, University of Leipzig, Leipzig 04103, Germany. ⁹Centro Andaluz de Biología del Desarrollo, Consejo Superior de Investigaciones Científicas, Universidad Pablo de Olavide and Junta de Andalucía, Carretera de Utrera Km1, Sevilla 41013, Spain. ¹⁰Institut de Bioteconologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, 08193, Barcelona, Spain. ¹¹Current address: Centre for Genomic Regulation and UPF, Doctor Aiguader 88, Barcelona 08003, Catalonia, Spain. ¹²Department of Evolutionary Genetics, Max-Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany. ¹³Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain. ¹⁴Parc Zoològic de Barcelona, Barcelona 08003, Spain. ¹⁵Howard Hughes Medical Institute, 3720 15th AVE NE, Seattle, WA 98195, USA. ¹⁶Centre for Genomic Regulation and UPF, Doctor Aiguader 88, Barcelona 08003, Catalonia, Spain.

Received: 14 January 2013 Accepted: 23 May 2013
Published: 31 May 2013

References

- Sabatier PJ: An albino lowland gorilla from rio Muni, West Africa, and notes on its adaptation to captivity. *Folia Primatol* 1967, **7**:155-160.
- Márquez M, Serafini A, Fernández-Bellón H, Serat S, Fener-Admetlla A, Bertranpett J, Fener L, Pumarola M: Neuropathologic findings in an aged albino gorilla. *Ver Pathol* 2008, **45**:531-537. doi:10.1354/vp.45-4-531.
- Gronskov K, Ek J, Brondum-Nielsen C: Oculocutaneous albinism. *Orphanet J Rare Dis* 2007, **2**:43. doi:10.1186/1750-1172-2-43.
- Martínez-Arias R, Comas D, Andrés A, Abelló MT, Domingo-Roura X, Bertranpett J: The tyrosinase gene in gorillas and the albinism of "Snowflake". *Pigment Cell Res* 2000, **13**:467-470.
- Robbins MM, Bermejo M, Cipolletta C, Magliocca F, Parnell RJ, Stokes E: Social structure and life-history patterns in western gorillas (*Gorilla gorilla gorilla*). *Am J Primatol* 2004, **64**:145-159. doi:10.1002/ajp.20069.
- Harcourt AH, Stewart KJS: *Gorilla Society: conflict, compromise and cooperation between sexes*. Chicago: The University of Chicago Press; 2007.
- Vigliant L, Bradley BJ: Genetic variation in gorillas. *Am J Primatol* 2004, **64**:161-172. doi:10.1002/ajp.20070.
- Thalmann O, Fischer A, Lankester F, Paabo S, Vigliant L, Thalmann O, Fischer A, Lankester F, Paabo S, Vigliant L, Cleve CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, et al: Insights into hominid evolution from the gorilla genome sequence. *Nature* 2012, **483**:169-175. doi:10.1038/nature10842.
- Ventura M, Catacchio CR, Alkan C, Marques-Bonet T, Sajadian S, Graves TA, Hormozdiani F, Navarro A, Malig M, Baker C, Lee C, Turner EH, Chen L, Kidd JM, Archidiacono N, Shendure J, Wilson RK, Eichler EE: Gorilla genome structural variation reveals evolutionary parallels with chimpanzee. *Genome Res* 2011, **21**:1640-1649. doi:10.1101/g124461.111.
- Nsubuga AM, Holman J, Chemnick LG, Ryder OA: The cryptic genetic structure of the North American captive gorilla population. *Conserv Gen* 2009, **11**:161-172. doi:10.1007/s10592-009-0015-x.
- Marco-Sola S, Sammeth M, Guljog R, Ribeca P: The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* 2012. doi:10.1038/nmeth.2221.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The sequence alignment/Map format and SAMtools. *Bioinformatics* 2009, **25**:2078-2079. doi:10.1093/bioinformatics/btp352.
- Ng PC, Henikoff S: SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003, **31**:3812-3814. doi:10.1093/nar/gkg509.
- Azhubel IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: A method and server for predicting damaging missense mutations. *Nat Methods* 2010, **7**:248-249. doi:10.1038/nmeth0410-248.
- Newton JM, Cohen-Barak O, Hagiwara N, Gardner JM, Davison MT, King RA, Brilliant MH: Mutations in the human orthologue of the mouse underwhite gene (*uw*) underlie a new form of oculocutaneous albinism, OCA4. *Am J Hum Genet* 2001, **69**:981-988. doi:10.1086/324340.
- Fukamachi S, Shimada A, Shima A: Mutations in the gene encoding B, a novel transporter protein, reduce melanin content in medaka. *Nat Genet* 2001, **28**:381-385. doi:10.1038/ng584.
- Mariat D, Taouit S, Guérin G: A mutation in the *MATP* gene causes the cream coat colour in the horse. *Genet Sel Evol* 2003, **35**:119-133. doi:10.1186/1297-9686-35-1-119.
- Gunnarsson U, Hellström AR, Tixier-Boichard M, Minvielle F, Bed'hom B, Ito S, Jensen P, Rattink A, Vereijken A, Andersson L: Mutations in *SLC45A2* cause plumage color variation in chicken and Japanese quail. *Genetics* 2007, **175**:867-877. doi:10.1534/genetics.106.063107.
- Inagaki K, Suzuki T, Ito S, Suzuki N, Adachi K, Okuyama T, Nakata Y, Shimizu H, Matsuura H, Oono T, Iwamatsu H, Kono M, Tomita Y: Oculocutaneous albinism type 4: six novel mutations in the membrane-associated transporter protein gene and their phenotypes. *Pigment Cell Res* 2006, **19**:451-453. doi:10.1111/j.1600-0749.2006.00332.x.
- Hessa T, Kim H, Bihlmaier K, Lundin C, Boebel J, Andersson H, Nilsson L, White SH, Von Heijne G: Recognition of transmembrane helices by the endoplasmic reticulum translocator. *Nature* 2005, **433**:377-381.
- Martínez-Gil L, Sauri A, Vilar M, Pallás V, Mingarro I: Membrane insertion and topology of the p78 movement protein of Melon Necrotic Spot Virus (MNSV). *Virology* 2007, **367**:348-357. doi:10.1016/j.virol.2007.06.006.
- Martínez-Gil L, Pérez-Gil J, Mingarro I, Martínez-Gil L, Pérez-Gil J: The surfactant peptide KL 4 sequence is inserted with a transmembrane orientation into the endoplasmic reticulum membrane. *Biophys J* 2008, **95**:36-38. doi:10.1529/biophysj.108.138602.
- Charlesworth D, Willis JH: The genetics of inbreeding depression. *Nat Rev Genet* 2009, **10**:783-796. doi:10.1038/nrg2664.
- Charpentier MJE, Widdig A, Alberts SC: Inbreeding depression in non-human primates: a historical review of methods used and empirical data. *Am J Primatol* 2007, **69**:1370-1386. doi:10.1002/ajp.20445.
- Pemberton TJ, Altschul D, Feldman MW, Myers RM, Rosenberg NA, Li JZ: Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* 2012, **91**:275-292. doi:10.1016/j.ajhg.2012.06.014.
- Chowdhury R, Bois FRJ, Feingold E, Sherman SL, Vivian G: Genetic analysis of variation in human meiotic recombination. *PLoS Genet* 2009, **5**. doi:10.1371/journal.pgen.1000648.
- Harcourt AH, Stewart KJS, Fossey DJ: Male emigration and female transfer in wild mountain gorilla. *Nature* 1976, **263**:226-227.
- Douadi M, Gatti S, Leverro F, Duhamel G, Bermejo M, Vallet D, Menard N, Petit EJ: Sex-biased dispersal in western lowland gorillas (*Gorilla gorilla gorilla*). *Mol Ecol* 2007, **16**:2247-2259. doi:10.1111/j.1365-294X.2007.03286.x.
- Bradley BJ, Doran-Sheehy DM, Lukas D, Boesch C, Vigliant L: Dispersed male networks in western gorillas. *Curr Biol* 2004, **14**:510-513. doi:10.1016/j.cub.2004.02.062.
- Stokes EJ, Parnell RJ, Olejniczak C: Female dispersal and reproductive success in wild western lowland gorillas (*Gorilla gorilla gorilla*). *Behav Ecol Sociobiol* 2003, **54**:329-339. doi:10.1007/s00265-003-0630-3.
- Bergl RA, Vigliant L: Genetic analysis reveals population structure and recent migration within the highly fragmented range of the Cross River gorilla (*Gorilla gorilla diehli*). *Mol Ecol* 2007, **16**:501-516. doi:10.1111/j.1365-294X.2006.03159.x.
- Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J: Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* 2010, **11**:R119. doi:10.1186/gb-2010-11-12-r119.
- Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo B, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Kasnitz A, McCombie WR, Hicks J, Wigler M: Tumour evolution inferred by single-cell sequencing. *Nature* 2011, **472**:90-94. doi:10.1038/nature09807.
- Peters BA, Kermali BG, Sparks AB, Aflerov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J, Robasky K, Zaranek AW, Lee JH, Ball MP, Peterson JE, Perazich H, Yeung G, Liu J, Chen L, Kennerly MI, Pothuraju K, Korwicka K, Tsouplio-Sitnikova M, Pant KP, Ebert JC, Nilsen GB, Baccash J, Halpern AL, Church GM, Dmanac R: Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 2012, **487**:190-195. doi:10.1038/nature11236.

doi:10.1186/1471-2164-14-363

Cite this article as: Prado-Martinez et al.: The genome sequencing of an albino Western lowland gorilla reveals inbreeding in the wild. *BMC Genomics* 2013 **14**:363.