TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Chemie der Biopolymere

Mapping the Human Bitopic Membrane Proteome for Self-Interacting Transmembrane Helices

Jan Kirrbach

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. D. Frischmann
Prüfer der Dissertation:		
	1.	Univ.-Prof. Dr. D. Langosch
	2.	Prof. Dr. I. T. Arkin, The Hebrew University of Jerusalem, Jerusalem, Israel

Die Dissertation wurde am 04.10.2012 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 10.01.2013 angenommen.

Abstract

Integral membrane proteins comprise 25-30% of any proteome and take part in countless cellular processes. Most membrane proteins form non-covalent homo- and hetero-oligomers. This is favored by the constraints of the lipid bilayer, but is also driven by sequence-specific interactions of α-helical transmembrane domains (TMDs) which involve amino acid motifs that form well-packed interfaces. The non-covalent assembly of TMDs is reflected by the pattern of residue conservation during evolution. TMDs are more conserved than soluble domains in bitopic membrane proteins and interior-facing residues of polytopic protein TMDs are more conserved than lipid-exposed amino acids and co-evolve more often. Taken together, it appears as if the evolved primary structures of TMDs mirror their ability to oligomerize. This raises the question to which extent bitopic membrane proteins can be grouped into families based on their potential transmembrane oligomerization domains.

This work systematically assesses the self-interaction of the human bitopic TMDs clustered by sequence homology. It is focused on revealing indicators for specific and efficient TMD self-interaction and addresses questions about the significance of TMD-based clusters, the distribution of self-interaction in the human bitopic membrane proteome, and the evolution of TMDs.

First, the level of sequence homology at which TMDs can be clustered had to be identified. By comparing all-against-all pairwise alignments of natural TMDs from human bitopic proteins and their randomized counterparts, a similarity threshold of $\geq 55\%$ was identified. The clustering based on this threshold groups the human bitopic proteins almost as efficiently as clustering based on the similarities of the complete sequences. The majority of the 33 largest clusters are functionally rather homogeneous as their members are annotated with similar function. However, pairwise TMD alignments also suggest relationships between TMDs that belong to proteins being apparently unrelated in function. The TMDs of those proteins were enriched in GxxxG motifs and might reflect convergent evolution of TMDs towards structures with similar properties.

The self-interaction of a representative TMD from each large cluster was investigated with the ToxR assay and revealed a broad distribution of affinities. A significant fraction of the representative TMDs exhibits high relative affinity within the range of Glycophorin A. Such high-affinity TMDs tend to exhibit orientation-dependence, which indicates preferential helix-helix interfaces, and mutation-sensitivity, which signifies well-packed interfaces. Taken together, the co-occurrence of high relative affinity, orientation-dependence and mutation-sensitivity indicates specific and efficient self-interaction.

Zusammenfassung

Integrale Membranproteine umfassen 25-30% jedes Proteoms und sind in zahlreiche Prozesse einer Zelle involviert. Ein Großteil aller Membranproteine bildet nichtkovalente Homo- oder Heterooligomere. Dies wird von Beschränkungen der Lipiddoppelschicht begünstigt, aber von der sequenzspezifischen Wechselwirkung α-helikaler Transmembrandomänen (TMDn), die eng gepackte Aminosäuremotive einbezieht, angetrieben. Die nichtkovalente Anordnung von TMDn ist in den Positionen der Aminosäurereste im Lauf der Evolution konserviert worden. TMDn von bitopischen Proteinen sind stärker konserviert als deren lösliche Domänen, wie auch proteinangrenzende Reste von polytopischen Proteinen stärker konserviert sind und öfter ko-evolvieren als deren lipidexponierte Reste. Zusammengefasst scheint es, als ob die evolvierten Primärstrukturen von TMDn deren Fähigkeit zur Oligomerisation widerspiegeln. Dies wirft die Frage auf, in welchem Umfang bitopische Membranproteine anhand ihrer potentiellen transmembranständigen Oligomerisierungsdomänen in Familien gruppiert werden können.

Diese Arbeit untersucht systematisch die Selbstinteraktion von menschlichen bitopischen TMDn, wobei diese anhand ihrer Sequenzhomologie gruppiert werden. Des Weiteren werden Indikatoren für spezifische und effiziente TMD-Selbstinteraktion aufgedeckt, sowie Fragen nach der Bedeutung von TMD-basierten Clustern, der Häufigkeit von Selbstinteraktion im menschlichen bitopischen Membranproteom und der Evolution von TMDn adressiert.

Als erstes wurde der Schwellenwert der Sequenzhomologie, bis zu dem TMDn gruppiert werden können, identifiziert. Durch den Vergleich paarweiser Alignments von jeder gegen jede menschliche TMD und deren randomisierten Gegenstücke, wurde eine Ähnlichkeitsgrenze von $\geq 55\%$ identifiziert. Die Clusterbildung basierend auf dieser Grenze gruppierte die TMDn von menschlichen bitopischen Proteinen beinahe so effizient wie eine vergleichbare Gruppierung basierend auf Ähnlichkeiten von kompletten Proteinsequenzen. Die Mehrheit der 33 größten Cluster sind funktionell eher homogen, da deren Mitgliedern ähnliche Funktionen zugeschrieben werden. Dennoch deuten einige paarweise TMD-Alignments auf eine Verwandtschaft zwischen TMDn, die offensichtlich funktionell nicht verwandt sind. Die TMDn von solchen Proteinen sind mit GxxxG-Motiven angereichert und könnten eine konvergente Evolution von TMDn zu Strukturen mit ähnlichen Eigenschaften widerspiegeln.

Die Selbstinteraktion der repräsentativen TMD von jedem der größeren Cluster wurde mithilfe des ToxR-Systems untersucht. Ein wesentlicher Teil der repräsentativen TMDn

weisen hohe relative Affinität im Bereich von Glycophorin A auf. Solche hochaffine TMDn tendieren zu Orientierungsabhängigkeit, die auf bevorzugte Helix-Helix-Kontaktflächen hindeutet, und zu Mutationssensitivität, die auf eng gepackte Kontaktflächen hinweist. Im Allgemeinen zeigt das gemeinsame Auftreten von hoher Affinität, Orientierungsabhängigkeit und Mutationssensitivität spezifische und effiziente Selbstinteraktion an.

Acknowledgments

During the past years of work for this thesis I was guided, inspired, encouraged, and supported by many people. It is my great pleasure to have the opportunity to thank all of them:

Prof. Dr. Dieter Langosch :: for being the best supervisor I can imagine; for the support from the very beginning of my work until its finish by offering me scientific freedom and inspiration at the same time; for highly encouraging trust in my work and person as well as the various opportunities to present my work and exchange scientific knowledge; for the many enjoyable scientific discussions and the easy chats in which we unleashed our phantasy and often had to laugh about some wild ideas. I am happy and proud to call him my "Doktorvater".

Prof. Isaiah (Shy) Arkin :: for supporting the collaboration for the project over all the time; for inviting me to Jerusalem and being a wonderful host for my terrific stay in Israel; additionally, for agreeing to review this thesis despite the distance.

Dr. Philipp Pagel :: for your scientific and technical guidance through bioinformatical and statistical challenges of this thesis as well as the frequent meetings and their inspiring discussions; for agreeing to review the scientific publication of this work.

Christian Ried :: for being the best colleague one can wish for; for your kind personality as a scientist and a friend which made every day so much more enjoyable; for your irreplaceable help and input for each and every part of this work; furthermore, for the review and correction of every important document I had to create during my PhD.

Eliane Küttler :: for your advices which improved my life in a way I never had imagined; for carefully proof-reading this thesis and many other writings.

Oxana Pester :: for being my office room mate for three years and helping me in keeping motivated in times when everything seemed to collapse.

Miriam Krugliak :: for your tremendous work in the ToxR lab in Jerusalem and a more than pleasant collaboration.

Barbara Rauscher :: for being the angel of the ToxR lab and your great work in preparing nearly every lab material.

Dr. Jana Herrmann :: for passing on your wide knowledge about ToxR experiments and its scientific background as well as some inspiration for the creation of this document.

Drs. Angelika Fuchs and Sindy Neumann :: for your support in bioinformatical questions and data management as well as for all the fun time we had together.

Sebastian Kube, Felix Behr, and Manuel Mohr :: for all your input into my project and the knowledge I gained from your questions during your thesis time.

My past and present colleagues (Markus, Walter, Martin, Karo, Aline, Marcella, Susanne, Ellen, Elke, Doreen, Christoph, Steven, Ute, Sevnur, Oli, Edwin, and Rashmi) as well as all students :: for the wonderful atmosphere you created and the cooperation and inspiring discussions with you; for the many fun and scientific moments we had together.

Last but not least I want to thank my parents Konrad and Christine and my sister Mandy as well as all my friends (Wedge, Nina, Jackie, Tristan, Sqeedy, Simon, Jana, Steph, Alex, Leon, Achim, Flo, Marina, Nadia, Patrick, Gesine, and many, many more...) for their constant support and all the great time we had during my doctoral studies and even long before.

Contents

List of Figures

List of Tables

1 Introduction

All organisms comprise of the same basic structural and functional unit of life: The cell. For over 300 years, biologists all over the world have been researching the function and architecture of this general module and its components. They have revealed common features and fundamental molecular processes in all analyzed species. One of the major functions of the cell is the separation of various contents from the surrounding environment. This is achieved by the cell membrane, a phospholipid bilayer, which envelops the cytoplasm. Membranes also create compartments within the cell. By dividing the cell in defined spaces with distinct physicochemical conditions, different biochemical processes can simultaneously take place within a single cell. This compartmentalization enabled the development of complex organisms inhabiting our planet nowadays.

The following introduction aims at summarizing present knowledge about membranes and membrane-spanning proteins. The first chapter briefly outlines the current view of membranes as they influence the structural and energetic properties of integrated proteins. In the following, the relevance, the structural characteristics, and the biogenesis of membrane proteins are described. Further, the current knowledge regarding the interaction of membrane spanning α-helices is outlined as well as methodology for the analysis of such interactions. Finally, a basic understanding of the relation of sequence, structure and function of proteins is presented.

1.1 Membranes

Biological membranes are lipid bilayers composed of various phospholipids. The viscosity of eukaryotic membranes is dependent on chain length and saturation of the lipids as well as the cholesterol content. The thickness of membranes averages to 60 Å. Acyl chains of phospholipids constitute the 30 Å thick hydrophobic core region of the membrane,

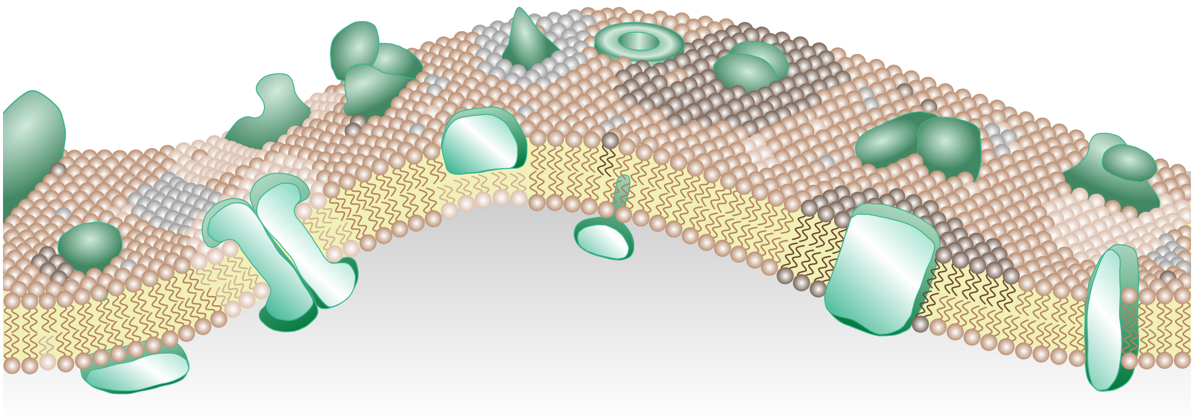

Figure 1.1: Schematic illustration of a lipid bilayer. Two layers of various phospholipids (red) constitute the two-dimensional viscous fluid membrane in which proteins (green) are embedded or at which they are anchored or attached. Different lipid species (different shades of brown and gray) can be unequally distributed over the leaflets of a bilayer but also laterally in plane. The enrichment of specific lipids causes membrane heterogeneity. Those microdomains may stabilize membrane proteins or they may be stabilized by proteins.

whereas their polar head groups add a 15 Å thick boundary to either membrane surface. The polarity within a membrane drastically decreases from water environment via the water/membrane boundary to the core region. Therefore, the membrane represents a complex heterogeneous environment [1].

The Fluid-Mosaic-Model of Singer and Nicolson describes membranes as two-dimensional viscous fluids containing freely diffusing membrane proteins [2]. However, different lipid species are not alone distributed among the leaflets of a bilayer but also organized laterally in the plane (figure 1.1). Such membrane microdomains of variable lipid composition are referred to as rafts [3].

1.2 Membrane proteins

Membranes are barriers which molecules generally cannot pass without assistance. However, cells have to exchange molecules between compartments, the cytoplasm, and their environment in order to sustain life. Nutrients need to be absorbed, produced waste has to be disposed, and reagents as well as products of reactions have to be carried from one compartment to another. Therefore, proteins providing aid to transport across membranes are embedded into the complex and highly dynamical lipid structure of biological membranes. Around 30% of the genes in eukaryotic species are encoding integral membrane proteins [4, 5, 6]. The following sections comprise the significance, the structural characteristics, the biogenesis, and the evolution of membrane proteins.

1.2.1 Biological and medical importance

Membrane proteins take part in countless cellular processes, i.e transport of substances, signal transduction, cell-cell communication and recognition, and energy production. Since only hydrophobic molecules are able to permeate through lipid membranes, the crossing of hydrophilic substances like ions is achieved and controlled by selective channels and pumps [7, 8]. For example, neuronal cells use ion channels to transmit electrical impulses. Membrane associated proteins also mediate the signal transduction across the membrane. For instance receptors of growth factors influence the development of cells by regulating the proliferation and differentiation [9, 10]. Furthermore, cells use membrane proteins like cadherins [11] and integrins [12] to contact adjacent cells or the extracellular matrix. This enables cells to connect to their surrounding tissue and adapt their morphology and motion. Membranes are also involved in energy production. Either cellular respiration and photosynthesis use an electric potential across the membrane to generate chemical energy in form of ATP which is consumed in many different cellular processes [13, 14]. Concluding from the importance of membrane proteins, it is not surprising that they constitute more than 60% of all targets of medical research [15].

1.2.2 Structural characteristics

Membrane proteins can be bound by peripheral or integral membrane attachment (figure 1.1, page 2). Peripheral membrane proteins are associated with the membrane surface either to acyl chains using hydrophobic interaction or to lipid head groups via electrostatic interaction. In some cases, such proteins are also anchored covalently to a fatty acid or lipid. In contrast, integral membrane proteins traverse both lipid layers of a membrane. They are either bitopic or polytopic and thus possess one or more transmembrane domains (TMDs), respectively. Bitopic proteins account for a substantial fraction of membrane proteins that increases from $\sim$15% in bacteria to $>$ 40% in humans [16]. The membrane-spanning parts of integral membrane proteins are limited to two building principles: the α-helix-bundle and the β-sheet-barrel. Unfolded membrane proteins are unable to permeate lipid bilayers due to their free polar peptide bonds that are unfavorable in the apolar environment of membranes. The folding into α-helical and β-sheet structures saturates the hydrogen bond potential of the peptide backbone and enables the integration into membranes [1, 17].

Proteins with β-barrel structure are constructed cylindrically from anti-parallel β-strands. Such proteins constitute water-filled pores in outer membranes of gram-negative bacteria, chloroplasts and mitochondria. In contrast, integral membrane proteins with α-helical TMDs are found within all cellular membranes. They are also more abundant, more versatile in their structure and function, and include almost all membrane proteins of medical importance.

A canonical α-helix, which perpendicularly traverses the 30 Å thick hydrophobic core region of a membrane, comprises around 20 amino acids. Each turn of a right-handed helix includes 3.6 residues. The axial shift between two succeeding residues is 1.5 Å and they are twisted by around 100 °. In eukaryotes the length of a transmembrane helix is 20 to 30 amino acids [18, 19, 20] that are mostly hydrophobic [21].

1.2.3 Biogenesis

The folding of membrane proteins and their integration into the lipid bilayer are linked processes [1]. Polypeptide chains with adequate length and hydrophobicity are able to insert spontaneously into a lipid membrane as they form hydrogen bonds and adopt an α-helical structure [22]. The dissociation of those non-covalent bonds is energetically unfavorable and thus the insertion of TMDs is stable. However, most membrane proteins are not released directly into the cytoplasm since their hydrophobic characteristics would lead to incorrect folding and aggregation.

Most α-helical TMDs are inserted co-translationally via a complex protein localization machinery. In bacteria, the folding and insertion of membrane proteins is mainly mediated by the SecYEG translocon [23, 24, 25]. Additionally, there is a SecA/SecB [26] or YidC dependent process [27]. In eukaryotes, the Sec61 translocon inserts membrane proteins into the membrane of the endoplasmic reticulum similar to the SecYEG in prokaryotes. During the synthesis at the ribosome, membrane proteins are recognized at their signal peptide or their first hydrophobic segment by the signal recognition particle (SRP). After binding of the SRP at its membrane bound receptor the nascent polypeptide chain is transferred together with the ribosome to the translocon complex located at the membrane of the endoplasmic reticulum. The translocation process is resumed after the dissociation of the SRP and the insertion into the membrane takes place. Hydrophobic segments, e.g. signal sequences and TMDs, laterally leave the water-filled channel of the translocon and enter the membrane. The process of membrane protein translocation and biogenesis is further described in the literature [23, 28, 29, 30].

The topology of α-helical membrane proteins is determined by interaction of nascent polypeptide chains with the translocon complex. Positively charged amino acid residues in flanking regions of the TMDs influence their orientation. The "positive-inside rule" implies that positively charged arginine and lysine residues are located more frequently at the cytoplasmic side of TMDs than at the periplasmic or extracellular regions [31, 32]. This principle was first observed in bacterial inner membrane proteins but also applies to eukaryotic membrane proteins [4, 33]. However, the topology of membrane proteins has little effect on the TMDs to partitioning out of the translocon and thus their integration efficiency [34].

Most probably only the interaction of the specific protein segment with lipids is decisive for its integration into the membrane [35, 36]. Therefore, the amino acid sequence of a TMD determines the partitioning out of the translocon into the membrane bilayer. Protein segments get distributed between the lipid phase of the membrane and aqueous phase in the translocon channel dependent on their length and hydrophobicity. Thereby, energetic effects are important, i.e. hydrophobic effects, the expenditure of energy for the removal of the hydrate coating of the peptide, interactions of polar amino acids with lipid head groups, the loss of entropy caused by the decreased degrees of freedom of the protein's side chains, and the TMD's influence to the lipids. Translocon mediated integration of membrane proteins is based on direct protein-lipid interaction and thus dependent on amino acid composition but also on the sequence which determines the position of specific residues within the membrane [36, 37]. However, as long the gain of free energy for the insertion of hydrophobic residues into the hydrophobic core of the membrane exceeds the cost for the insertion of polar and ionizable residues, even those energetic unfavorable residues can be inserted into the membrane [23, 35]. The knowledge about hydrophobicity of TMDs and the positive-inside rule led to development of prediction tools (e.g. TMHMM [38], Phobius [39], and SignalP [40]) for the identification of membrane proteins and their topologies.

1.2.4 Evolution and oligomerization

Evolution has found means of diversification that lead to the known variety of functions mediated by integral membrane proteins. Various evolutionary mechanisms like gene duplication followed by structural and functional diversification of individual copies formed families of paralogs [41, 42]. The evolution of soluble proteins appears to differ significantly from that of integral membrane proteins. While the diversity of soluble proteins is further increased by gene fusion, fission, and swapping of domains, domain recombination is not common for integral membrane proteins [43]. It has therefore been argued that complex membrane protein functions in higher organisms may be supported by formation of non-covalent homo- and heterooligomeric complexes, rather than by recombination of TMDs. Indeed, most membrane proteins form oligomers [44, 45, 46, 47]. This oligomerization often ivolves the assembly of TMDs [48, 49, 50, 51].

The non-covalent assembly of TMDs is reflected by the pattern of residue conservation during evolution. TMDs are more conserved than soluble domains in bitopic membrane proteins [52]. Further, one-sided conservation of TMDs from bitopic proteins [53] is consistent with the observation that interior-facing residues of polytopic protein TMDs are more conserved than lipid-exposed amino acids [54, 55, 56] and co-evolve more often [57]. Taken together, it appears as if the evolution of protein TMDs mirrors their ability to oligomerize.

1.3 Interaction of transmembrane helices

The correct folding and association of subunits of proteins is crucial for their functionality. Hence, the interaction of the membrane embedded segments of a membrane protein is often essential for the assembly and function of the whole protein. Thereby, persistent and temporary interaction regulate the activity of a membrane protein. Transmembrane helix interactions are prevalent in cell membranes and are not only involved in assembly and folding of membrane proteins but also in signaling and subcellular localization [58, 59, 60].

Individual helix dimers of bitopic TMDs serve as an important model system for studying lateral transmembrane helix-helix interactions [61, 62, 63, 64, 65, 66, 67]. Transmembrane α-helices might interact to form helix dimers and subsequent interactions eventually form the final higher ordered oligomeric structures. Therefore, individual helix-helix interaction determine folding of polytopic membrane proteins [48] along with constraints by the covalent loops linking them.

In general, the interaction of helix dimers is categorized into homotypic interaction of two TMDs of the same type and in heterotypic interaction of two different types of TMDs. Defined interactions, which involve packing interactions, hydrogen bonding, aromatic interactions and salt bridges, can determine sequence specific packing of transmembrane helices in bitopic as well as in polytopic transmembrane proteins [48, 49, 68].

The following sections deal with the physical chemistry and sequence specific interaction motifs of interacting transmembrane helices.

1.3.1 Physical chemistry of helix-helix interaction

The association of molecules goes along with a reduction of their degrees of freedom and thus with a loss of entropy. This influences the interaction equilibrium in favor of monomers. In the case of helix-helix interaction, the loss of entropy for the peptide backbone is low since the TMDs are already folded before association. The loss of side chain rotamer entropy at contact surfaces of interacting helices also counteracts TMD di- and oligomerization. However, TMDs associate despite the cost of entropy if favorable enthalpic contributions dominate [69].

Due to the strict confinement of lipids and transmembrane helices within the bilayer, helices and lipids will interact all the time with a maximum number of neighbor molecules. This interaction can be lipid-lipid, lipid-helix, or helix-helix. A transmembrane helix monomer interacts better with lipids than with other helices. Similarly, helix self-assembly will result if the sum of helix-helix and lipid-lipid interactions is favored over helix-lipid interactions [51]. Therefore, unfavorable helix-lipid interactions could be as important as favorable helix-helix interactions in determining the propensity of transmembrane helices in membranes to self-associate. In the following, the principles of lateral interactions between helices and lipids in membranes are described.

1.3.1.1 Helix-helix packing

Membrane proteins have been shown to be on average packed tighter than soluble proteins [70]. The major source of this observation are closely packed small residues [71] resulting in a better fit between helix surfaces which gives rise to more favorable van der Waals interactions [51]. From the perspective of helix-helix interactions, this effect has long been described as “knobs-into-holes” [64] and “ridges-into-grooves” [61]. In these cases, the interacting TMDs are tightly packed and sterically complementary to each other. The close assembly of such TMDs leads to the accumulation of van der Waals interactions as a driving force for helix-helix interaction despite they are the weakest form of attractive forces. Van der Waals interactions are formed between permanent or induced dipoles. Their strength is dependent on the polarity, the polarizability, and the distance between involved molecules. Since the strength decreases with the sixth power of the distance, van der Waals interactions are short ranged and weak.

1.3.1.2 Polar, ionic, and aromatic interactions

The introduction of a polar side chain into the hydrophobic membrane creates a unfavorable energy cost if it is exposed to the lipid hydrocarbon [1]. Therefore, salt bridges and hydrogen bonds between polar groups within the hydrophobic environment can reduce the energy cost [72] and drive TMDs to dimerize. Additionally, aromatic π-π interactions may occur between aromatics in interacting TMDs and further stabilize helix dimers [73, 74, 75]. However, the helix-bilayer system can respond to unfavorable membrane-embedded side chains in other ways than dimerization, i.e. shifting its position vertically in the bilayer.

Hydrogen bonds are established between two properly arranged electronegative atoms which compete for a hydrogen atom. The hydrogen atom is covalently bound to the hydrogen donor and is partially positively charged caused by the higher electonegativity of the donor atom. In this way, the hydrogen atom can simultaneously interact with the second electronegative atom which is referred to as hydrogen bond acceptor. For instance, hydrogen bonds can be formed between polar and/or ionizable amino acid side chains, or amide and carbonyl groups of the polypeptide backbone. Also aromatic rings of aromatic side chains can act as hydrogen bond acceptor. The strength of hydrogen bonds exceeds the attractive forces of van der Waals interactions and they have a direction. Thus, hydrogen bonds increase the specificity and stability of helix-helix interactions.

The strongest non-covalent interaction promoting TMD-TMD interactions are salt bridges. They occur between ions such as in charged amino acid residues and have a large effect. Whether the side chains of D, E, H, R and K are charged is presumably dependent on their close vicinity. Since the strength of electrostatic interactions (i.e. van der Waals, polar, ionic, and aromatic interactions) is dependent on the dielectricity of the environment, such interactions are stronger in apolar membranes than in aqueous milieu.

1.3.1.3 Influences of the lipid bilayer

Beside the lipids, membranes have several other properties to influence or control the structure of TMD dimers and oligomers [48]. The length of the lipid acyl chains can vary significantly and thereby determine the thickness of a membrane. In case of differing transmembrane helix lengths and thickness of the membrane's hydrophobic core, "hydrophobic mismatches" can occur. A structural adaption of membrane proteins to the membrane thickness can be achieved by an adjustment of the tilt angle of the TMD within the membrane or a rearrangement of amino acid side chains at the ends of the helix [76, 77, 78]. Another possibility to overcome the hydrophobic mismatch is the lateral association of TMDs to reduce the contact surface with the lipid bilayer or the aqueous milieu [79, 80]. The order of acyl chains also influences fluidity of the membrane and thus the stability of helix dimers [81]. Cholesterol not only slightly increases the thickness of membranes but also results in the ordering of lipid acyl chains. This decreases the lipid chain entropy and leads to unfavorable helix-lipid interactions which then can drive the self-association of TMDs [77, 81, 82]. Furthermore, the nature of the lipid head group can influence the structure and function of surrounded proteins significantly. Charged head groups might attract protons and thereby lowering the local pH. The reversible protonation of individual amino acid side chains stabilizes or destabilizes the oligomeric structure of TMDs [48]. The amino acids R, K, W and Y interact directly with lipids. They anchor the TMD helices within a membrane and thus control the structure of TMD oligomers [83, 84]. Different lipid composition within each leaflet of a bilayer membrane add another level of complexity. Their lipid composition differs in eukaryotic membranes [85] as well as in some prokaryotes [86]. This asymmetry might influence the formation and stabilization of distinct transmembrane protein structures. Besides the differences between the two monolayers of a membrane, a lateral bilayer asymmetry, which creates local lipid domains with defined properties or results in local concentration of TMDs, might also control helix-helix interaction [48]. The prokaryotic model organism *E. coli* has a simpler lipid composition than eukaryotic cells. The lack of cholesterol and sphingolipids reduces the complexity of the lipid phase and influence the stability of integral membrane proteins [58].

1.3.2 Dimerization motifs

On the one hand, TMD association is favored by the constraints of the lipid bilayer which concentrate and pre-orient the proteins [87]. On the other hand, the oligomerization of many membrane proteins is driven by sequence-specific interactions of their α-helical TMDs which involve amino acid motifs that form well-packed interfaces [45, 60, 88]. The following sections describe the general principles of dimerization motifs and the most common amino acid patterns with examples of known TMD dimer structures.

1.3.2.1 General principles

Cymer *et al.* compared the structures of 11 transmembrane helix dimers which have been solved mainly by NMR spectroscopy [48]. The analysis comprised the homodimers of glycophorin A, growth factor receptors ErbB2 and ErbB3, receptor tyrosine kinases EphA1 and EphA2, receptor kinase BNIP3, the T cell signaling module $\zeta\zeta$, and the signaling module DAP12. Also, the heterodimer structures of synaptobrevin 2 and syntaxin 1A, ErbB1 and ErbB2 receptor tyrosine kinases, and the integrins αIIb and β3 were included. The structures indicate that the majority of their helices interact based on common models [18, 61, 64, 89]. Most helix pairs exhibit a regular structure of their contact area which is determined by their crossing angle. In case of left-handed helix-helix pairs, the angle between their axes is positive and the connectivity between the interfacial residues of adjacent helices conforms to the knobs-into-holes type of side chain packing (figure 1.2 left). Equally to soluble coiled-coils, the amino acids are arranged in a repeated heptad motif ([a..de.g]$_n$). In contrast, right-handed helix-helix pairs exhibit a negative crossing angle and are characterized by a repeated tetrad motif ([ab..]$_n$) which results in a ridges-into-grooves type of interaction (figure 1.2 right).

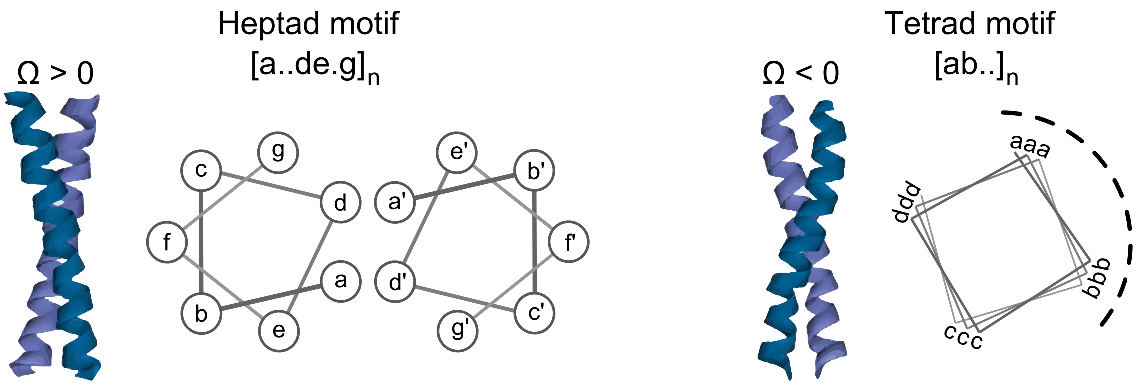

Figure 1.2: General interaction principles of membrane-spanning helix-helix pairs. The scheme on the left side shows a left-handed helix-helix pair with a positive crossing angle ($\Omega > 0$) between their axes. The helix wheel representation illustrates the arrangement of amino acids within heptad motifs [a..de.g]$_n$ including the interacting side chains at the a, d, e, and g positions. The image on the right side depicts the interacting side chains at the a and b position of a repeated tetrad motif. The helices of such a right-handed helix-helix pair cross with a negative angle ($\Omega < 0$). Figures are adapted from [88, 90, 91].

1.3.2.2 The GxxxG motif

In 1989 biochemical analyses have indicated that the transmembrane helix of human glycophorin A (GpA) forms a strong dimer (figure 1.3) *in vivo* and *in vitro* [92]. The sequence specific dimerization of GpA is accomplished by parallel, right-handed crossing of both transmembrane helices. After analyzing the interaction of this TMD in great detail the responsible amino acid motif was determined to LIxxGVxxGVxxT [62]. The core of this interface is the GxxxG motif. The two small side chains of the glycine residues might allow some structural flexibility of the helix and promote the close packing due to the formation of van der Waals interactions of neighboring amino acids as well as hydrogen bonding between C_α hydrogen atoms and carbonyl groups [93, 94]. Both glycines also create a void on one side of the surface of the transmembrane helix which can be filled with adjacent side chains of the interacting helix using the ridges-into-grooves principle [61].

The intense study of GpA contributed significantly to the understanding of structural and thermodynamic principles of TMD-TMD association [61, 62, 63, 95, 96]. The GxxxG dimerization motif is common in other membrane proteins [19, 20]. However,

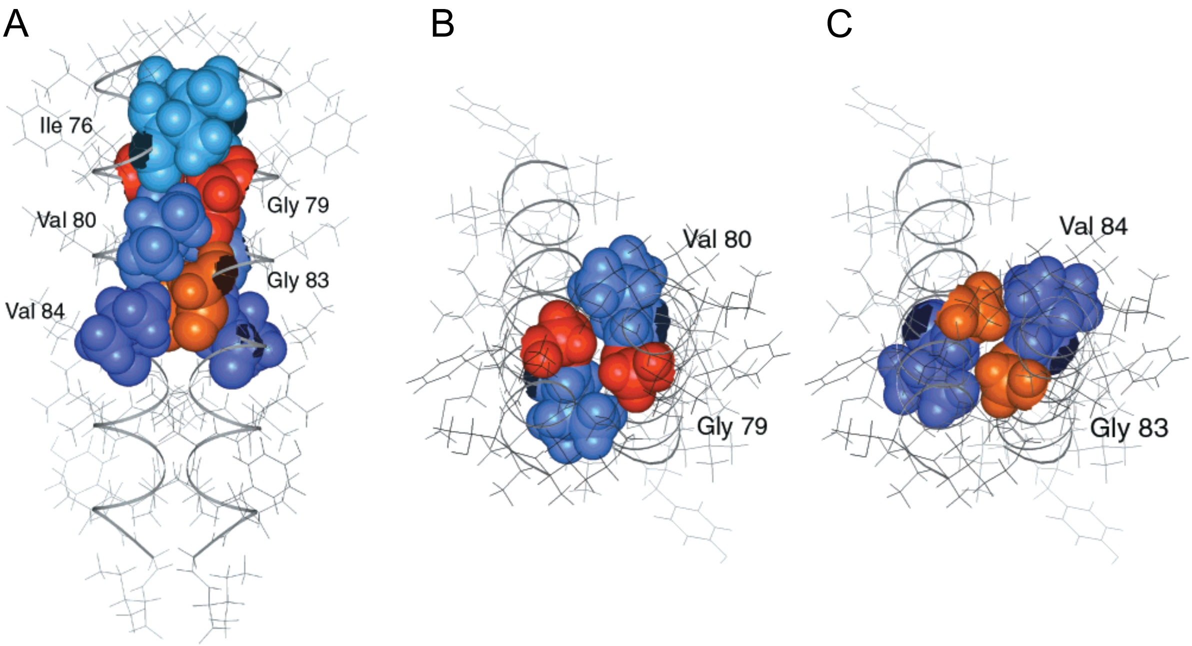

Figure 1.3: Structural model of the glycophorin A TMD dimer. (**A**) View along the dimer interface highlighting G_{79}, G_{83}, I_{76}, V_{80}, and V_{84}. The β-branched amino acids form a ridge that pack along the groove created by the Glycines. The view down the dimer axis shows the close packing of (**B**) G_{79}-G_{79} and (**C**) G_{83}-G_{83} (figures taken from [95]).

measurements of interaction energies have indicated that GxxxG-containing transmembrane helices may interact with remarkably diverse strength suggesting that sequence context is also important for the stability of TMD dimers [97, 98]. The small amino acids alanine and serine can replace one or both of the glycine residues [20, 89, 99]. In general, such amino acid patterns are referred to as GxxxG-like or Small-xxx-Small (SmxxxSm) dimerization motifs and can assist helix dimerization [100].

1.3.2.3 GxxxG-like motifs

SmxxxSm motifs have been shown to be involved in formation and stabilization of transmembrane helix-helix interactions in integrins or ErbB receptor kinases [60, 100, 101, 102]. In ErbB two GxxxG-like motifs are conserved and act in a switch-like fashion for the interaction between active and inactive states [103, 104]. The dimer is stabilized by the polar amino acid motif TxxxSxxxG. Hydrophobic residues L and V additionally stabilize the dimer by van der Waals packing interactions [105]. Some residues of the ErbB TMDs which are involved in the homo-dimerization are also participating in hetero-dimer formation [106]. In general, GxxxG-like motifs not necessarily support the formation of only a single TMD oligomer structure. Surrounding amino acids and the interacting helix determine the specificity and stability of a dynamic helix-helix-interaction. For example, the heterodimeric complex of αIIb and β3 integrin have been shown to be highly dynamic [107]. The interactions also involve conserved GxxxG-like motifs extended with strong van der Waals packing of aliphatic L, I, and V residues. However, other amino acids surrounding these critical residues also appeared to contribute to packing. Thus, stacking interactions of phenylalanine as well as a D-R electrostatic interaction appear to stabilize the integrin dimer structure.

1.3.2.4 Motifs with polar residues

Polar interactions are relevant in dimerization of transmembrane helices. They can extend and stabilize an existing dimer interface. For example, TMD dimers of the proapototic BNIP3 receptor kinase are stabilized by three small residues organized in a GxxxG-like glycine-zipper motif [108]. The interface involves also serine and histidine which form a hydrogen bond. This hydrogen bond might be a pH sensor triggering the structure and function of the helix dimer. In case of the human T-cell receptor DAP12, a pair of acidic residues is essential for the homo-dimerization. Both aspartate residues form interhelical hydrogen bonds [109] whereas the highly conserved GxxxG-like motif is located outside the contact surface and not involved in interhelical packing.

Polar residues do not introduce promiscuous interactions and rarely create a novel dimerization interface [51]. An example of such a rare sequence motif consisting of polar residues is the QxxS motif of the bacterial aspartate receptor TMD [110]. The TMD dimerization is directly dependent on the polar characteristic of the amino acid side chains since the interchange of both polar residues has no effect. In contrast the mutation to non-polar residues or the exchange to a GxxxG motif disrupts the dimer [110]. In the absence of GxxxG motifs, SxxSSxxT and SxxxSSxxT dimerization interfaces were found to be the most overrepresented in a pseudo-random genetic library which was selected for interacting helices in bacterial membranes [67]. That interaction was shown to be position specific and stabilized through interhelical hydrogen bonds.

1.3.2.5 Leucine zippers

The leucine zipper motif is an example for a simple repeated sequence motif comprised of hydrophobic residues. In a leucine zipper every first and fourth position in a seven residue heptad repeat is a leucine, isoleucine, or valine [111]. These hydrophobic residues form the contacts between the interacting helices. In contrast to the right-handed crossing angle of helices containing GxxxG and other SmxxxSm motifs, membrane-spanning leucine zippers are interacting with a left-handed crossing angle (figure 1.2 left, page 10). In studies of dimerizing helices where SmxxxSm motifs do not explain the obtained data, leucine zipper motifs have been considered [112, 113, 114, 115].

1.3.3 Analysis of TMD-TMD interaction

Some of the frequently used strategies to determine the strength, dynamics and specificities of TMD-TMD interactions by experimental (ToxR, TOXCAT, GALLEX, and FRET) or computational approaches (MD and Bioinformatics) are summarized below and reviewed in [50]. First, the ToxR system is introduced in detail.

1.3.3.1 The ToxR system

The ToxR system is a genetic assay for the detection of TMD-TMD interactions within the *E. coli* inner membrane. The name originates from the ToxR transcription activator from *Vibrio cholerae* that activates the expression of its virulence factors [116]. A periplasmic sensor domain is connected via a TMD to the cytoplasmic ToxR domain and thus can transport an extracellular signal across the membrane by dimerization of the protein and all of its domains. The dimer of the ToxR domain has an increased affinity to

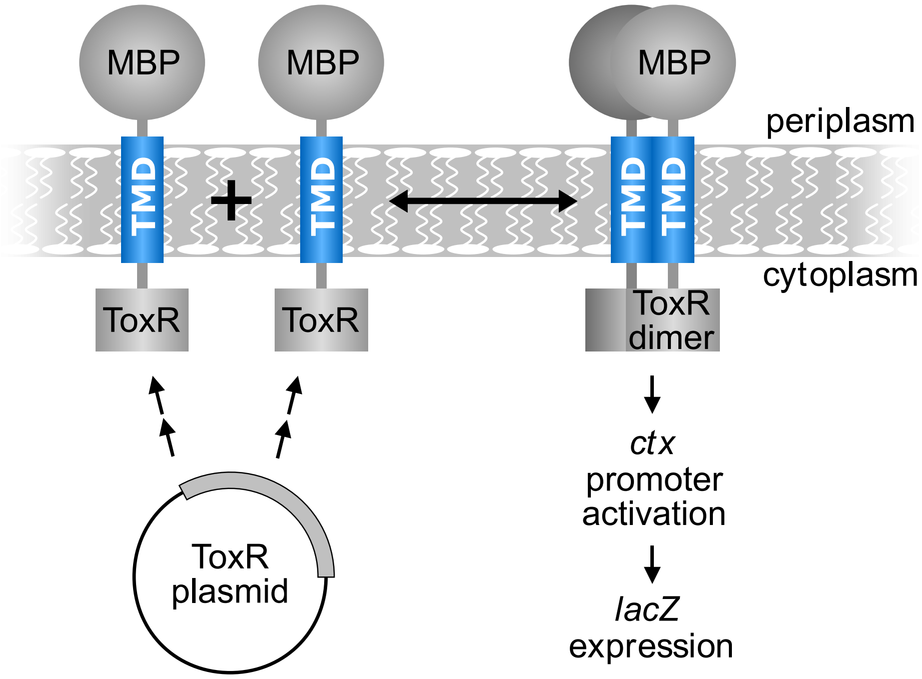

Figure 1.4: The ToxR reporter activator system. The self-interaction of TMDs leads to the di- or oligomerization of the cytoplasmic ToxR domains. The dimers activate the transcription of the β-Galactosidase reporter gene under the control of a *ctx* promoter. Periplasmic MalE domains allow for the detection of the chimeric protein with antibodies and for the analysis of correct membrane insertion (adapted from [121]).

the repetitive DNA binding site of the *ctx* promoter which then activates the expression of the cholera toxin [117, 118]. Because of its oligomerization dependent activity and modular structure that enables the exchange of domains [119, 120], the ToxR protein is suitable for investigation of TMD-TMD interaction. Therefore, a chimeric protein consisting of the ToxR transcription activator domain, a heterologous TMD, and the periplasmic maltose binding protein (MalE) was created [63, 120] (figure 1.4).

The close association of the TMD within the MalE-TMD-ToxR fusion protein leads to the di- or oligomerization of the protein and thus of the ToxR domain. Consequently, the ToxR dimer binds and activates the *ctx* promoter of the *E. coli* FHK12 indicator strain [119, 120]. Therefore, the *LacZ* reporter genes are expressed dependent on the TMD-TMD interaction (figure 1.4). The enzyme β-Galactosidase (β-Gal) catalyzes the hydrolysis of ortho-nitrophenyl-β-galactoside (ONPG) and the resulting amount of o-nitrophenol can be measured colorimetrically at an optic density of 405 nm within the cell lysate [122]. Dependent on the initial reaction rate, the quantity of β-Gal and thereby the relative interaction strength of a TMD can be determined. Furthermore, the spatial orientation of the TMD-TMD interface relative to the DNA-binding ToxR domain influences the efficiency of transcription activation. Each TMD has to be inserted at different phases into ToxR-TMD/MalE chimeric proteins. With the insertion of one amino acids at the N-terminus and the deletion of one amino acid at the C-terminus of the TMD, the potential interface is rotated by approximately 100 °. The *E. coli* FHK12 strain also possesses an F-plasmid encoded β-Gal fragment (ω-fragment) which can be expressed with the addition of isopropyl-β-D-1-thiogalactopyranosid (IPTG). The ω-fragment competes with complete β-Gal proteins for the assembly of functional enzyme complexes [112]. This competition can increase the variation in β-Gal reporter activity and simplifies the distinction between different strengths of TMD interaction.

In commonly used versions of the system (ToxRIV [113], ToxRV [123]), the expression of the chimeric ToxR proteins is regulated by the *pBAD* operator/promoter of the *araBAD* operon which can be repressed by the AraC protein. Arabinose can bind to the AraC protein and thus prevent its interaction with DNA. This mechanism is required to influence the concentration of the MalE-TMD-ToxR fusion protein which then affects the equilibrium between mono-, di-, and oligomer within the membrane in addition to the affinity of the TMDs. The expression level can be determined via SDS-PAGE [124] of the cell lysate and subsequent Western blot [125]. The detection of the 66 kDa chimeric protein is realized with anti-MalE antibodies. Only proteins with similar concentrations can be compared for their β-Gal reporter activity and thus for their TMD-TMD interaction. The MalE domain is also used to test for sufficient integration and correct topology (periplasmic MalE domain and cytoplasmic ToxR domain, figure 1.4, page 14) of the fusion protein within inner membrane of *E. coli.* As a part of the transporter system to take up maltose, MalE binds to the sugar and passes it to the membranous components [126]. In *E. coli* PD28 cells [127] the *MalE* gene is deleted. Therefore, PD28 cell have to be complemented with correct inserted MalE-TMD-ToxR protein to grow with maltose as sole carbon source. This is used in the PD28 assay for correct membrane insertion [127].

Using the ToxR system, self-interacting TMDs can be selected from large pools of transmembrane helices via combinatorial peptide libraries (1.3.3, page 13). Furthermore, their association strength can be estimated from measured β-Gal activities. To characterize helix-helix contact surfaces, each amino acid of the TMD can be mutated and its interaction can be compared to the wild-type TMD interaction measured with the ToxR system. The pattern of critical residues suggest the specific interaction motif of the TMD [74, 75, 128].

1.3.3.2 Other experimental techniques

Similar to the ToxR transcriptor activator system which uses the expression of β-Gal as reporter for TMD-TMD interaction, TOXCAT is based on the expression of chloramphenicol acetyltransferase [129]. Both systems have been widely used to asses mainly homo-dimerization but also hetero-dimerization. Another genetic assay for the identification of hetero-assembly is the GALLEX system [130]. In this system, β-Gal is constitutively expressed by *E. coli* cells until the interaction of TMDs is repressing the expression of the reporter. In contrast to enzymatic color reactions, some recently developed methods use the fluorescent characteristics of labels attached to potentially interacting TMDs. Using the analytical fluorescence resonance energy transfer (QI-FRET) technique, one can follow dynamic TMD-TMD interactions within the membrane milieu [131]. Thereby, two fluorescently tagged TMDs are scanned and visualized by confocal microscopy. Another technique termed stop flow fluorescence analysis is based on two distinct fluorescent labels present in a model TMD peptide. Both probes are sensitive to their environment and respond with a change in fluorescence [132].

1.3.3.3 Computational approaches

Many researchers combine or support their studies with computational analyses. Due to the lack of structural data, molecular dynamic (MD) simulations are widely used to investigate membrane protein interfaces [133]. For example, MD simulations are used to examine configuration changes within the tilt and position of TMDs relative to the membrane bilayer [134], to compare the dynamics of TMD self-assembly [135], to compare structural aspects of different dimerization motifs [136], to probe unknown interactions when there is lack of experimental data [137, 138], or even to design new protein-binding peptides [139]. Another frequently used computational method is the large-scale scan of TMD libraries for enriched amino acid patters. By calculating odds ratios of occurrence, potential interaction motifs can be preselected for experimental investigation. This approach was used to identify novel interacting motifs like SmxxxSm patterns in association with large aliphatic residues [20] or the combination of cationic interaction with GxxxG close packing [140]. Further examples include the discovery of the QxxS and WxxW dimerization motifs [110, 136].

1.4 Protein homology

A completely different application of bioinformatics is the comparison and alignment of amino acid sequences from different proteins. Protein sequence homology is often used to infer function or structure of an uncharacterized protein from a known one. If the similarity of two sequences is significantly non-random [141], two proteins may have a common origin and functional or structural relationship is likely. The homology of sequence is often used to cluster proteins or peptides sharing structure, function, or evolution.

1.4.1 Relation of protein sequence and structure

Based on the general concept that sequence similarity implies structural similarity [141], transmembrane domains with a certain sequence homology should adopt a similar structure. The Needleman-Wunsch algorithm [142] for detecting global alignments and the Smith-Waterman algorithm [143] for local alignments have been published many years ago. Although both algorithms are not fast enough to search large sequence databases, they quickly find the optimal alignment of pairwise compared sequences. In cases of full database search, algorithms like FASTA [144] and BLAST [145] emerged. All these algorithms use amino acid substitution matrices to score the alignment of sequences. In contrast to the identity of two domain sequences, the calculation of scores that depend on amino acid similarity is more sensitive, because not only identical but also similar amino acids positively contribute to the score.

Pairs of soluble proteins with a sequence identity higher than 35–40% are very likely to be structurally similar [146]. Structural similarity in pairs with a sequence identity of 20–35% (often referred to as “twilight zone” [147, 146]) is considerably less common. Above a cut-off of roughly 30% sequence identity, 90% of the pairs were found to be homologous. At the same time, the twilight zone is characterized by an explosion of false negatives, which means that many dissimilar sequences appear to be structural homologous. As a result, less than 10% of protein pairs with sequence identity below 25% have similar structures. For transmembrane proteins and especially for transmembrane helices such a cut-off is not yet revealed.

1.4.2 Classification of homologous proteins

Since evolution of membrane proteins may increase their ability to oligomerize (1.2.4, page 5), bitopic membrane proteins might be grouped into families based on their potential membrane-spanning oligomerization domains. Existing classification approaches operate at the level of sequence [148, 149, 150], structure [151], and/or function and have mostly been applied to polytopic membrane proteins. Almén *et al.* classified 6,718 predicted human membrane proteins including bitopic proteins based on full-length sequence similarity and grouped them into 234 functional families [6]. This greatly detailed annotation allowed the identification of new gene families and novel members of existing families without further structural analyses. The bitopic topology was found to be the least characterized structure type.

1.5 Motivation

As described in this introduction, the specific interaction of α-helical TMDs plays an important role for the folding, oligomerization, and function of membrane proteins. Thereby, the interaction is often mediated by specific amino acid residues dependent on their physical and chemical properties. The finding and characterization of helix-helix interaction motifs might improve the understanding of mechanisms that ensure specific TMD-TMD interactions and avoid non-specific ones. Former comprehensive studies included the search of combinatorial libraries for highly self-interacting helices and afterwards the measurement of their interaction with genetic or biophysical methods. Unfortunately, the size restriction of such libraries limits the coverage of the TMD sequence space which leads to a possible loss of interaction motifs. Since the analysis of all combinatorial TMD sequences is impossible at the moment, the search for interacting helices in natural proteins which are formed by evolution might be better suited to systematically investigate TMDs.

This work aims to systematically assess the self-interaction of a significant part of the human bitopic membrane proteome by clustering homologue TMDs and testing self-interaction of selected TMDs. This approach will provide a general impression about the interaction of bitopic membrane proteins and the occurrences of specific interaction motifs as well as their specificity. By detecting fundamental characteristics of the most common TMDs new motifs could be found and the knowledge about known interfaces might be broadened. The analysis of such a large amount of TMD sequences may also allow to draw some conclusions about the development and evolution of helix-helix interfaces.

The first step was to classify the transmembrane helices of the human bitopic membrane proteome on the basis of sequence similarity and derive a list of representative TMDs for self-interaction analyses. Therefore, a meaningful sequence similarity threshold for TMDs had to be identified first and then used for clustering human bitopic TMDs. The analysis for self-interaction of the representative TMD from each major cluster was anticipated to reveal the distribution of relative affinities. In a number of high-affinity TMDs, mutational analyses were planned to assess the sequence specificity of the self-interaction.

2

Material and Methods

This work combines bioinformatics and molecular biology. Therefore, this section is separated into two parts. The first part describes the data of human transmembrane protein sequences and bioinformatic methods for the classification of TMD sequences and the search for putative interaction motifs. The second part delineates the required laboratory materials and methods for the assessment of TMD self-interaction. An even more detailed description of laboratory methods can be found on the CD attached to the very last page of this thesis.

2.1 Sequence data and substitution matrices

For the comparison and grouping of TMDs, the human protein sequences and specific amino acid substitution matrices were required. This section describes the downloaded sequence data and applied substitution matrices.

2.1.1 The UniProtKB database

A database of human proteins was downloaded from the UniProtKB database [152] (release 57.9, Oct. 2009) containing 34,761 proteins.

2.1.2 Amino acid substitution matrices

For the comparison of protein sequences and calculation of their similarity from pairwise alignments, amino acid substitution matrices were required. Those matrices contain scoring values for each possible amino substitution and are required for calculating distances between protein sequences which then can be utilized e.g. for grouping similar sequences or drawing phylogenetic trees.

2.1.2.1 The PHAT matrix

The PHAT7573 amino acid substitution matrix [153] was used to compare transmembrane regions by calculating similarity scores. The matrix was created by Ng *et al.* by using predicted hydrophobic and transmembrane regions from the Blocks database [154]. The PHAT matrix was created to compensate for the different amino acid composition of transmembrane protein parts in contrast to soluble domains. There are more advanced matrices for the search of membranous domains in protein databases [155]. However, since solely the comparison of membranous regions among themselves was required the PHAT7573 amino acid substitution matrix was the most suitable.

2.1.2.2 The BLOSUM62 matrix

For the comparison of complete protein sequences the default amino acid substitution matrix of water (program module of the EMBOSS, 2.2.1.2, page 23) was used. Usually the choice of the required substitution matrix is dependent on the evolutionary distance and the type of alignment of the compared protein sequences. However, the same method of performing pairwise alignments between all sequences is crucial for comparability. In addition, the evolutionary distances between proteins may strongly vary within the human genome. Here, only the BLOSUM62 amino acid substitution matrix [156] was utilized, which performs well for local alignments.

2.2 Applied computer programs

The application of bioinformatic methods to biological sequence data requires widely used computer programs. The needed bioinformatic computer tools and programming software are introduced below.

2.2.1 Bioinformatic tools

This section comprises all programs used to analyze and manipulate biological sequence data.

2.2.1.1 Phobius

Phobius [39] (Version 1.01) is a combination of the bioinformatic tools TMHMM [38] and SignalP-HMM [40]. While TMHMM uses a Hidden Markov model for the prediction of transmembrane regions of a protein sequence, SignalP reveals signal peptides which are often falsely predicted as TMDs. Hence, the association of TMHMM with SignalP increases the prediction accuracy of TMDs and particularly the prediction of the correct number of membrane spanning regions.

2.2.1.2 EMBOSS

The European Molecular Biology Open Software Suite [157] (EMBOSS, Version 6.1.0) includes the module 'water' which was used to perform pairwise local alignments of TMD or protein sequences. The module calculates bit scores between two amino acid sequences. The scores again were used to calculate score/selfscore ratios (ssr, 2.3.1.3, page 25) which reflect the sequence similarity between protein sequences. The modules 'protdist', 'neighbor', and 'drawtree' were used to calculate phylogenetic trees starting from multiple alignments of several protein sequences generated with ClustalX2.

2.2.1.3 ClustalX2

ClustalX (version 2.0.1) [158] was used to generate multiple alignments of more than two protein sequences by using default parameters. Either the PHAT7573 or the BLOSUM62 amino acid substitution matrix was used for TMDs or for complete proteins, respectively. In cases of TMD alignments the gap opening and gap extend parameters were set to 100, thereby excluding the introduction of gaps.

2.2.1.4 CLANS

CLANS [159] was used to create the graphical expression of clustered TMDs in figure 3.3 on page 58. As input artificial p-values were used which only represent the type of affiliation, either TMD similarity or complete sequence similarity. The output coordinates from CLANS were used with R 2.10.0 to generate the final image.

2.2.1.5 WebLogo

WebLogo [160] (version 3.0) was used with standard parameters to visualize conserved residues within multiple alignments of TMD sequences.

2.2.2 Programming languages

Most computational methods were implemented in Java 1.6.0 programming language using the Fedora Eclipse 3.4.1 platform. Smaller scripts were written in Unix console script language. All scripts and programs ran on a Intel Pentium 4, 2 GHz, 2 GB RAM machine with a Fedora 9 operating system and a Linux 2.6.27.12 kernel. Biological sequences and score values were stored in a MySQL 5.0.88 database. Statistical tests were performed with R 2.10.0 for statistical computing.

2.3 Computational methods

This section describes the bioinformatic methods for the clustering of similar TMD sequences and the search for specific interaction pattern as well as some computational methods applied on the results of ToxR interaction assays. A rough summary of the most important steps for classification of TMDs of the human bitopic membrane proteome is given in figure 2.1.

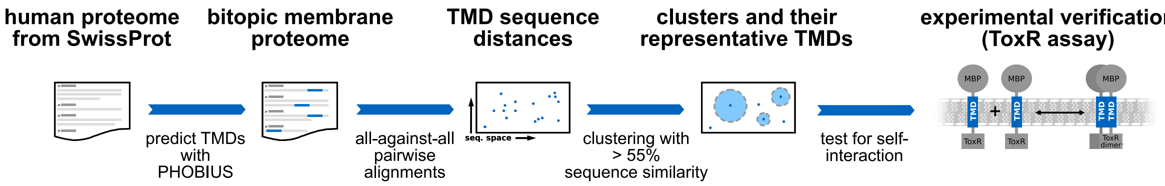

Figure 2.1: Strategy for classification of human bitopic TMDs using sequence similarity clustering. Proteins were clustered for their TMD sequence similarities. Representative TMDs of clusters that included > 5 members were tested for self-interaction using the ToxR system.

2.3.1 Clustering TMDs

The human bitopic membrane proteins were clustered for their TMD sequence similarity. First, the database of distinct bitopic TMDs had to be created. Then, pairwise all-against-all alignments were calculated and used to compute similarity scores. Finally, the scores were used to group similar TMD sequences and generate clusters.

2.3.1.1 Construction of a human bitopic TMD database

For the prediction of TMDs within membrane proteins the standard parameters of Phobius [39] (version 1.01, 2.2.1.1, page 22) were used. Only proteins containing one TMD and annotated as “single-pass membrane protein” in UniProtKB [152] were selected. From the resulting dataset of 3,534 bitopic membrane proteins the TMD sequences were extracted. Identical TMDs were retained in a single instance, yielding a database of 2,205 distinct TMDs.

2.3.1.2 Pairwise alignments of unique TMDs

Since a global alignment algorithm cannot calculate scores between sequences of different length and simultaneously exclude gaps, the module ’water’ of the EMBOSS package (2.2.1.2, page 23) was utilized to calculate local pairwise alignments and Smith-Waterman bit scores [143] between TMD sequences. The following parameters were changed from default: gapopen=100.00, gapextend=10.00, datafile=PHAT7573, aformat=score. The calculation of pairwise bit scores required an amino acid substitution matrix. Since default substitution matrices were designed from soluble proteins, the PHAT substitution matrix (2.1.2.1, page 22) was used instead. The parameters for gap opening and extension prevent the insertion of gaps in alignments. The prevention of gaps is important because interaction motifs are highly dependent on the position of involved amino acids. Gaps would stretch or tear such amino acid patterns.

2.3.1.3 Score/selfscore ratios

The ’water’ module (2.2.1.2, page 23) performed pairwise gapless alignments of a TMD sequence against the complete human bitopic TMD database (2.3.1.1, page 24). By repeating the procedure for each sequence in the dataset a matrix of all-against-all pairwise bit scores was obtained. Due to the dependence of the scores on the local alignment length, calculated bit scores cannot be used directly as a measurement for the distance between sequences of different length. To adjust the scores for TMD sequence length the entire score matrix was normalized by computing the ratios of the scores divided by their selfscore according to equation 2.1:

$$ssr_{s1,s2}[\%] = \frac{S_{s1,s2} + S_{s2,s1}}{S_{s1} + S_{s2}} \cdot 100\% \tag{2.1}$$

where $S_{s1,s2}$ and $S_{s1,s1}$ are the bidirectional bit scores of TMD sequence 1 and sequence 2, respectively. S_{s1} is the bit score of sequence 1 against itself as is S_{s2} for sequence 2. $ssr_{s1,s2}$ represents the score/selfscore ratio (ssr) of sequence 1 and 2 in percent. Those ratios are independent of the alignment length. They are also comparable to all other ssr.

2.3.1.4 Randomization of TMDs

In order to compare the similarities between TMDs to randomly occurring resemblance, the database of human bitopic transmembrane proteins had to be randomized. The length and amino acid composition of randomized sequences needed to stay equal to the initial non-randomized dataset to remain comparable. Therefore, the amino acid positions of each TMD sequence were shuffled resulting in a randomized set of 2,205 protein sequences. For the comparison of occurrences of putative interaction motifs within randomized and natural TMDs, the procedure was repeated 1,000 times to increase the accuracy of randomization.

2.3.2 Sequence characterization

TMD sequences were analyzed for their amino acid composition and putative interaction motifs. By comparison of two datasets of TMDs odds ratios were calculated. Further, the similarity of complete protein sequences was compared to TMD similarities by calculating $\mathrm{ssr}_{\mathrm{TMD}}/\mathrm{ssr}_{\mathrm{complete}}$ similarity ratios.

2.3.2.1 Odds ratios

TMD sequences were searched for the enrichment of specific amino acids or putative interaction motifs. Therefore, the amino acids of all investigated sequence were counted and compared to their expected occurrences. Putative interaction motifs were only counted once per TMD. In most cases the counts of motifs or amino acids were compared between two subsets of TMDs and odds ratios were calculated according to equation 2.2:

$$odds\ ratio \quad = \quad \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \tag{2.2}$$

where p_1 and p_2 are the probabilities of occurrence within the first and the second subset of TMDs, respectively. Odds ratios above 1.0 indicate that the occurrence of the specific motifs or amino acids is more likely in the first set of TMDs, whereas a odds ratio below 1.0 indicate enrichment in the second set of TMDs.

2.3.2.2 Similarity ratios

To compare TMDs and complete protein sequence similarities of each protein in the dataset, pairwise ssr values (2.3.1.3, page 25) at the level of TMD ($\mathrm{ssr}_{\mathrm{TMD}}$) and complete sequences ($\mathrm{ssr}_{\mathrm{complete}}$) between cluster members and their most representative sequence were calculated. These ssr were used to compute $\mathrm{ssr}_{\mathrm{TMD}}/\mathrm{ssr}_{\mathrm{complete}}$ ratios according to equation 2.3:

$$similarity\ ratio \quad = \quad \frac{ssr_{TMD}}{ssr_{complete}} \tag{2.3}$$

where ssr_{TMD} is the score/selfscore ratio between a member TMD and the representative TMD of a cluster, and $ssr_{complete}$ is the score/selfscore ratio between a complete protein sequence of a cluster's member and the complete protein sequence which belongs to the most representative TMD of that cluster. Since complete protein sequences mainly consist of extramembranous parts, $\mathrm{ssr}_{\mathrm{TMD}}/\mathrm{ssr}_{\mathrm{complete}}$ ratios above 1.0 signify that TMDs are more similar than their corresponding extramembranous domains, whereas ratios below 1.0 indicate that TMDs are less similar than the extramembranous domains. The $\mathrm{ssr}_{\mathrm{TMD}}/\mathrm{ssr}_{\mathrm{complete}}$ ratios can be compared between different sets of sequences.

2.3.3 Analysis of ToxR reporter activities

In general, relative β-Galactosidase (β-Gal) activities are presented as box plots which are generated by R 2.10.0. To compare ToxR reporter activity between different orientations of most representative TMDs, values for orientation-dependence were calculated. In addition, β-Gal activities also served for the computation of the impact of specific mutations on self-interaction of TMDs.

2.3.3.1 Orientation-dependence

In case a TMD was analyzed for self-interaction in different orientations, a numeric value for the orientation-dependence of the TMD was calculated from the highest and lowest interacting orientation of the TMD. The following equation 2.4 was used:

$$ORD \quad = \quad 1 - \frac{median(ToxR)_{min}}{median(ToxR)_{max}} \tag{2.4}$$

where $median(ToxR)_{min}$ and $median(ToxR)_{max}$ are the lowest and highest median β-Gal activities, respectively. ORD is the orientation-dependence between 0 and 1. Values close to 0 represent low dependence of the β-Gal activity on the TMD's facing motifs, whereas values close to 1 indicate a large difference in activity between different TMD orientations.

2.3.3.2 Impact of point mutations

After mutating specific amino acids within a TMD, the ToxR reporter activity was measured for the mutated construct. To indicate the effect of the mutation, the decrease or increase of reporter activity was calculated in respect to the wild-type TMD self-interaction. Therefore, the impact of point mutations (IPM) was calculated according to the following formula 2.5:

$$IPM = \left|\frac{median(ToxR)_{mut}}{median(ToxR)_{wt}} - 1\right| \quad (2.5)$$

where $median(ToxR)_{wt}$ and $median(ToxR)_{mut}$ are the median β-Gal activities for the wild-type TMD and the mutated form of the TMD. IPM is the impact of that specific point mutation in a range between 0 and 1. Values close to 1 indicate a large influence of mutations, whereas values close to 0 represent low impact of the mutation on the self-association of the investigated TMD. The MIM is the maximal impact of point mutations for an TMD that was mutated at different positions.

2.4 Laboratory materials

If not mentioned differently, general chemicals were obtained from the companies Applichem, Roth, and Sigma-Aldrich. Solutions, buffer, and media were prepared with distilled (dH_2O) or deionized water. The recipes of required solutions and buffers are listed at their specific methods. If necessary they were filtered sterile (0.45 µm) or autoclaved.

2.4.1 Media

The media down below were used to grow *E.coli* cells:

LB medium (pH 7.0):		
1	%	(w/v) Tryptone
0.5	%	(w/v) Yeast extract
171	mM	NaCl
Adjust pH with NaOH.		

SOB medium (pH 7.0):		
2	%	(w/v) Tryptone
0.5	%	(w/v) Yeast extract
8.6	mM	NaCl
2.5	mM	KCl

All other required media are listed at their corresponding methods. For the preparation of solid media 1.5% (w/v) agar was used. All media were autoclaved.

2.4.2 Plasmids and bacterial strains

In order to measure the interaction of different TMDs the insertion of each individual TMD into the ToxR-TMD-MalE fusion protein is required. This was realized by exchanging the TMD coding sequence from the starting vector pToxRV αV mut [123] via cassette cloning. Figure 2.2 on page 30 depicts the pToxRV αV mut plasmid and its essential components.

The following *E.coli* strains and plasmids in table 2.1 were used in this work:

Table 2.1: *E. coli* strains and the starting plasmid used in this work.

Label	Resist.	Genotype	Application	Reference
FHK12	Amp^R	*ctx::lacZ*	interaction analysis	[120, 119]
PD28	Tet^R	*MalE-*	integration test	[161]
DH5α	-	*supE44 ΔlacU169 (ϕ80lacZΔM15) hsdR17 recA1endA1 gyrA96 thi-1 relA1*	cloning, mutagenesis, plasmid preparation	-
XL1-Blue	Tet^R	-	cloning, mutagenesis	Stratagene
TOP10	Str^R	-	cloning, mutagenesis	Invitrogen
pToxRV αV mut	Km^R	*araBAD::toxR/tmd/ malE(Δ367-370) /myc/his6* ColE1 origin	interaction analysis	[123]

For the long term storage of bacterial strains at -80 °C overnight cultures were mixed with 15% sterile glycerin and frozen in liquid nitrogen.

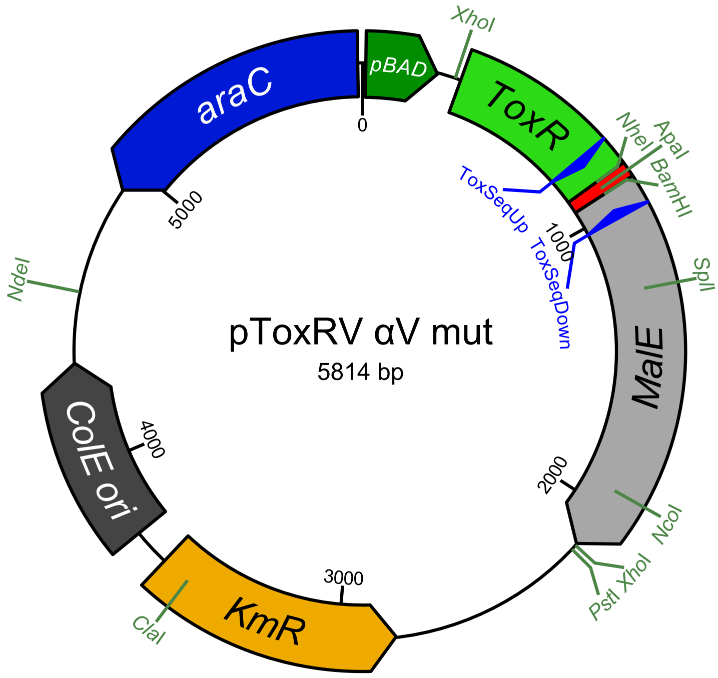

Figure 2.2: The pToxRV αV mut plasmid contains a coding region for the ToxR-TMD-MalE fusion protein consisting of the ToxR domain (green), the αV mut TMD (red), and the maltose binding domain MalE (gray). The fusion protein is under the control of the *pBAD* operator/promoter (dark green) which can be repressed by the *araC* (blue) gene product. The AraC protein can be bound by arabinose preventing its binding to DNA and enabling the expression of the ToxR-TMD-MalE fusion protein. The plasmid codes the high copy *E. coli* origin *ColE* (black) and the kanamycin resistance gene *KmR* (orange). The binding locations for sequencing primers are marked in blue.

2.4.3 Antibiotics

The following antibiotics in table 2.2 were used for the selection of bacterial strains and transformed cells:

Table 2.2: Antibiotics used to select for transformed *E. coli* cells.

Antibiotic	Final concentration	Selection
Ampicillin	100 µg/ml	FHK12 cells
Tetracycline	12.5 µg/ml	PD28 and XL1-Blue cells
Kanamycin	33 µg/ml	pToxRV plasmid
Streptomycin	100 µg/ml	TOP10 cells

2.4.4 Enzymes and antibodies

All restriction enzymes and DNA modifying enzymes (i.e. T4 DNA ligase and T4 polynucleotide kinase) were obtained from Fermentas. The *PfuUltraII* polymerase was purchased from Stratagene and the RNase A from Applichem. The following table 2.3 lists the utilized antibodies:

Table 2.3: Antibodies used for the detection of the maltose binding protein.

Antibody	Dilution	Source
Rabbit anti MBP antiserum	1:10,000	New England Biolabs
Anti rabbit IgG AP conjugate	1: 7,500	Promega

2.4.5 Oligonucleotides

Sequencing primers are 5' fluorescence tagged (2.5.9, page 39) and were purchased from MWG-Biotech. They were solved in dH_2O and stored with a concentration of 100 pmol/µl at -20 °C. Further oligonucleotides required for the cassette cloning (2.5.7, page 37) were desalted and not modified. Those were ordered from Invitrogen, diluted to a concentration of 100 pmol/µl in dH_2O, and stored at -20 °C. Primers for mutagenesis (2.5.8, page 38) were created accordingly to the information given from the manufacturer of QuikChange™ Site-Directed Mutagenesis Kits and also purchased from Invitrogen. The following table 2.4 lists the used sequencing primers:

Table 2.4: The nucleotide sequences of sequencing primers in 5' to 3' orientation.

Primer	Sequence	Application
ToxSeqUp	`CGCAGAATCAAGCAGTGTGCC`	Binds to the ToxR domain for sequencing the TMD in sense direction (2.5.9, page 39)
ToxSeqDown	`CCGTTATAGCCTTTATCGCCG`	Binds to the MalE domain for sequencing the TMD in anti-sense direction (2.5.9, page 39)

The sequences of oligonucleotides to create TMD cassettes are designed individually as described in section 2.5.7 on page 37.

2.4.6 Size standards

The size standards listed below in table 2.5 were used to mark the band spacing in DNA and protein detecting gels:

Table 2.5: Gel size standards for detecting DNA and proteins.

Size standard	Utilization	Source
GeneRuler™ DNA Ladder Mix	Agarose gel electrophoresis	Fermentas
GeneRuler™ 1 kb DNA Ladder	Agarose gel electrophoresis	Fermentas
Page Ruler Unstained Protein Ladder	SDS-PAGE	Fermentas
Page Ruler Prestained Protein Ladder	SDS-PAGE	Fermentas

2.4.7 Kit systems and prepared material

The following prepared kit systems and materials in table 2.6 were used in this work:

Table 2.6: Kit systems for DNA purification and sequencing.

Label	Utilization	Source
NucleoSpin® Plasmid	Plasmid extraction	Macherey-Nagel
NucleoSpin® Extract II	DNA extraction from agarose gel	Macherey-Nagel
QuikChange™ Site-Directed Mutagenesis Kit	Position specific mutagenesis	Stratagene
SequiTherm EXCEL II DNA Sequencing Kit-LC	DNA sequencing	Epicentre Biotechnologies
dNTP-Mix (10 mM)	Position specific mutagenesis	Fermentas
Long Ranger® Gel Solution 50%	Sequencing gels	Biozym
30% acryl-bisacrylamide mix	SDS-PAGE	Applichem
Filter paper (FN 7a, 200 g/m^2)	Western blots	Munktell & Filtrak GmbH
Nitro cellulose membrane	Western blots	GE Healthcare

2.4.8 Equipment

The specialized devices listed in table 2.7 were used in this work.

Table 2.7: Laboratory equipment and their manufacturers.

Device	Type	Manufacturer
ELISA Reader	Thermomax microplate reader	Molecular Devices
Sequencer	DNA Sequencer LONG READIR 4200 incl. Base ImageIR Image Analysis 4.0	LI-COR Biosciences
Sequence editor	BioEdit Sequence Alignment Editor 7.0.5.3	[162]
Thermocycler	Mastercycler	Eppendorf
Electrophoresis device	PerfectBlue double gel system Twin S, PerfectBlue double gel system Twin ExW S	peqlab
Centrifuges	Centrifuge Z 513 K	Hermle
	Table centrifuge Z 233 MK-2	Hermle
	Vacuum centrifuge Univapo 100 H	UniEquip
Photometer	Ultrospec 3100pro photometer	GE Healthcare Bio-Sciences

2.5 Molecular biological methods

This section comprises methods to design, produce, purify, and manipulate genetic material for the generation of required plasmid DNA to measure TMD interaction using the ToxR system.

2.5.1 Preparation of competent cells

The *E. coli* strains FHK12, PD28 and TOP10 were transformed by heat shock (2.5.2.1, page 35) whereas the strain XL1-Blue was transformed by electroporation (2.5.2.2, page 35). Chemically competent and electro-competent cells are prepared as described below.

2.5.1.1 Preparation of chemically competent cells

Chemically competent cells are required for the heat shock transformation of *E. coli* cells. The preparation of chemically competent cell was performed using the method of Hanahan [163]. 50 ml SOC medium were inoculated with 1 ml of an overnight culture of the desired strain and subsequently incubated at 37 °C and 250 rpm. The optical density of the culture was measured at 600 nm at regular intervals. After reaching an OD_{600} of 0.4 to 0.6 the cells were harvested at 4,000 rpm (Hermle Centrifuge Z 513 K) and 4 °C for 10 min. During all procedures the cells were kept cold to slow down their growth. The pellet was resuspended in 15 ml ice-cold transformation buffer and incubated for further 10 min on ice. Afterwards, the cells were centrifuged again at 4,000 rpm and 4 °C for 10 min and resuspended in 5 ml ice-cold transformation buffer. 175 µl DMSO were added under gentle rocking and the mixture was incubated for 5 min on ice. Again, 175 µl DMSO was added similarly (to a final concentration of 7% (v/v)) and incubated for 5 min on ice. The cell suspension was aliquoted in 100 µl per tube, immediately frozen down in liquid nitrogen, and stored at -80 °C.

SOC medium (pH 7.0):		
		SOB medium (2.4.1, page 28)
10	mM	$MgCl_2$
10	mM	$MgSO_4$
20	mM	Glucose

Filter all solutions with 0.45 pore size.

Transformation buffer (pH 6.7):		
10	mM	PIPES
15	mM	$CaCl_2 \cdot 2H_2O$
250	mM	KCl

Adjust the pH with KOH.
Filter with 0.45 pore size.
Store at 4 °C.

2.5.1.2 Preparation of electro-competent cells

E. coli cells which have to be transformed via electroporation were prepared to become electro-competent. 400 ml of LB medium were inoculated with 1% (v/v) of a fresh overnight *E. coli* culture of the required strain. The cells were grown at 37 °C and 300 rpm to an OD_{600} of 0.6. Afterwards, they were chilled on ice for 15 min and all following steps were performed on ice as close as possible to 0 °C. The culture was span down at 4,000 rpm (Hermle Centrifuge Z 513 K) and 4 °C for 15 min. After discarding the supernatant the cells were washed three times with 100 ml, 50 ml, and 20 ml 10% (w/v) glycerol, respectively. Each washing step was followed by the centrifugation of the cells at 4,000 rpm and 4 °C for 10 min. Finally, the cells were resuspended in 10% (w/v) glycerol with a volume of 1% of the initial LB culture. The suspension was aliquoted in 50 µl per tube, immediately frozen in liquid nitrogen, and stored at -80 °C.

2.5.2 Transformation of competent cells

Competent *E. coli* cells were transformed with plasmid DNA either using heat shock transformation or electroporation. The number of transformed cells in case of the electroporation [164] is about 10-100 times higher than with heat shock transformation. Therefore, the electroporation method was preferably performed with plasmid DNA from ligated plasmids which usually are low in plasmid concentration.

2.5.2.1 Heat shock transformation

The competent cells were thawed either on ice or at 37 °C for 30 s. Approximately 100 ng plasmid DNA were added to 100 µl chemically competent cells (2.5.1.1, page 34) and incubated for 20 min on ice. The cell suspension was exposed to a heat shock for 60 s at 42 °C in a water bath followed by a 2 min cooling on ice. Afterwards, 400 µl LB medium were added to the cells which subsequently had to be agitated at 37 °C for 1 h to develop the antibiotic resistance. The cell suspension was either used to measure the ToxR interaction assay (2.7.1, page 42) or for the inoculation of liquid cultures (1:100) with the appropriate antibiotics (2.4.3, page 30).

2.5.2.2 Electroporation

Aliquots of 50 µl electro-competent (2.5.1.2, page 34) cells were thawed on ice. A maximum of 1 µl of DNA from a ligation solution or around 50 ng plasmid DNA were added to the cells and mixed carefully. After filling the mixture in a precooled 1 mm cuvette for electrical transformation the electroporation was performed using one single pulse of 1,800 V. 200 µl LB medium were added into the cuvette. The cell solution was transferred to a reaction tube and subsequently agitated at 37 °C for 1 h to develop antibiotic resistance. The cells were either centrifuged at 1,200 xg and room temperature for 2 min and then plated on agar medium with antibiotic or used to inoculate liquid cultures with the appropriate antibiotics (2.4.3, page 30).

2.5.3 Extraction of plasmid DNA

For the preparation of plasmid DNA the NucleoSpin® Plasmid purification kit from Macherey-Nagel was used accordingly to the manufacturers protocol for high copy plasmid purification. In case plasmid DNA had to be extracted from an agarose gel, the NucleoSpin® Extract II DNA extraction kit was utilized analogously to the manufacturers protocol.

2.5.4 Enzymatic restriction digestion

If not noted otherwise, the digestion of plasmid DNA using restriction enzymes was performed in accordance with the manufacturers information. For preparation of plasmid DNA 1 to 4 µg DNA was used. In case of control digestions only 0.1 to 0.5 µg plasmid was digested for 1 h. The following restriction enzymes were applied:

Table 2.8: Restriction enzymes and their usage.

Enzymes	Buffer	Application	Reference
*Nhe*I/*Bam*HI	Green (Fermentas)	Deletion of TMDs from pToxRV αV mut	Cassette cloning (page 37)
*Apa*I	Green (Fermentas)	Digestion of unchanged or religated vector	Cassette cloning (page 37)
*Apa*I/*Pst*I	Blue (Fermentas)	Control for deleted TMD αV mut	Cassette cloning (page 37)
*Dpn*I	PfuUltraII polymerase buffer	Digestion of parental supercoiled dsDNA	Mutagenesis (page 38)

2.5.5 Agarose gel electrophoresis

The size specific separation of DNA was performed in 1% (w/v) agarose gels with 0.5 µg/ml ethidium bromide in TBE buffer using a voltage of 80 V for 45 to 60 min. The DNA samples were mixed with loading dye (Fermentas). To estimate the size and concentration of DNA fragments a size standard (2.4.6, page 32) with defined fragment sizes was used. The DNA was visualized under UV light (312 nm).

1x TBE buffer (pH 8.0):		
89	mM	Tris
89	mM	Boric acid
2.5	mM	EDTA

Store at room temperature.

2.5.6 Determination of DNA concentration

The concentration of prepared DNA (2.5.3, page 35) was photometrically measured at 260 nm in a quartz cuvette with a dilution of 1:50. In addition, the absorption was also recorded at 280 nm to detect a possible contamination with protein. For a pure DNA solution an OD_{260} of 1 corresponds to a concentration of double stranded DNA of 50 µg/ml. The quality of the DNA solution was deduced from the ratio of OD_{260} to OD_{280}. A OD_{260}/OD_{280} in range between 1.8 and 2 the DNA was considered as pure.

2.5.7 Cassette cloning

In order to clone a specific TMD into the ToxR fusion protein the sense and anti-sense DNA oligomers had to be designed individually (figure 2.3, page 38) and were ordered from Invitrogen. All primers were optimized for *E. coli* t-RNAs as well as to avoid hairpin structures and self-annealing by introducing silent mutations which do not change the amino acid sequence. For each investigated TMD sense and anti-sense oligomers were phosphorylated and simultaneously the pToxRV αV mut vector was digested with *Nhe*I and *Bam*HI in the same reaction mixture for 1.5 h at 37 °C. The following scheme shows the initial reaction mixture composition for one cassette cloning:

Vector linearization and oligo phosphorylation:		
1	µg	pToxRV αV mut plasmid DNA
1	µg	Sense oligonucleotide DNA
1	µg	Anti-sense oligonucleotide DNA
1x		Buffer green (Fermentas)
0.5	mM	dNTPs
2.5	U	*Nhe*I
2.5	U	*Bam*HI
10	U	Polynucleotide kinase

Fill up to a volume of 20 µl with dH_2O.

After digestion the restriction enzymes were inactivated and both phosphorylated oligonucleotides were hybridized to a short double strand DNA cassette by incubating the mixture for 20 min at 80 °C and cooling it down for 1 h in the switched-off heating block. By adding 4 U T4 DNA ligase and 0.5 mM dNTPs, vector fragment and oligo cassette were ligated for 1 h at 22 °C. Another incubation step for 10 min at 65 °C inactivated the ligase to avoid religation after the *Apa*I digestion of religated source vector. The

```
Q9UN71-0 forward
          .........1.........2.........3.........4.........5.........6
 5' CTAGCTtgcagttttatctggtggtTgcTctggcgctgattagcgtgctgtttctggtggcgatgGG 3'
           L  Q  F  Y  L  V  V  A  L  A  L  I  S  V  L  F  L  V  A  M

Q9UN71-0 reverse
          6.........5.........4.........3.........2.........1.........
5' GATCCCcatcgccaccagaaacagcacgctaatcagcgccagAgcAaccaccagataaaactgcaAG 3'
           M  A  V  L  F  L  V  S  I  L  A  L  A  V  V  L  Y  F  Q  L
```

Figure 2.3: Exemplary creation of oligonucleotides for the TMDs cassette of protocadherin gamma-B4 (UniProtKB ID: Q9UN71). Sense and anti-sense oligomer are both written from 5' to 3'. The TMDs have a fixed length of 20 amino acids eventually resulting in 60 nucleotides. Triplets were optimized for *E. coli*. Triplets depicted in green were refined to avoid hairpin structures and self-annealing of the single strand oligonucleotides by exchanging nucleotides (upper case) without changing the amino acid sequence. Terminal nucleotides (blue) form sticky ends for *Nhe*I and *Bam*HI restriction sites with overlapping nucleotides (red). Both oligomers are annealed and form a TMD cassette which is inserted into the pToxRV plasmid.

*Apa*I step was performed by adding 5 U restriction enzyme and an incubation for 1 h at 30 °C.

1 µl of the ligation batch was used to transform XL1-Blue cells (2.5.2.2, page 35) which then were separated on LB agar medium with Kanamycin (33 µg/ml). After the preparation of plasmid DNA from single colony clones an *Apa*I/*Pst*I digestion was performed to control for the removed αV mut TMD. In case the agarose gel showed an 1.25 kb DNA fragment, the clone was transformed with the pToxRV αV mut plasmid without incorporation of the desired TMD. Before using the vector for the ToxR assay (2.7.1, page 42) the TMD region was sequenced using the ToxSeqUp primer (2.5.9, page 39).

2.5.8 Position specific mutagenesis

Point mutations of single amino acids were introduced with the QuikChange™ Site-Directed Mutagenesis Kit from Stratagene by using the manufacturers protocol. The mutagenesis primers were designed to reach a melting point of ≥78 °C and the mutation strand synthesis reaction was performed with 50 ng parental plasmid DNA and the *PfuUltraII* DNA polymerase. After 18 cycles of PCR a *Dpn*I digestion was carried out to cut methylated parental supercoiled dsDNA. The reaction mixture was used to transform XL1-Blue (2.5.2.2, page 35) or TOP10 *E. coli* cells (2.5.2.1, page 35) and to amplify the point mutated DNA constructs. The TMD region was validated by a sequencing reaction (2.5.9, page 39) with the ToxSeqUp or ToxSeqDown fluorescence labeled primer.

2.5.9 DNA sequencing

The sequencing of plasmid DNA was performed after the chain-termination method from Sanger [165] by using the SequiTherm EXCELII DNA sequencing kit from Epicentre Biotechnologies and processed according to the manufacturers cycle sequencing protocol for 25-41 cm gels without mineral overlay. The chain-termination reaction was realized with sequencing primers (2.4.5, page 31) which were 5' fluorescence tagged (IRD700, IRD800).

The separation of the obtained fluorescence marked fragments was carried out with the DNA sequencer LONG READIR 4200 (LI-COR Biosciences). 0.5 µl of each reaction was loaded on a 25 cm long and 0.25 mm thick polyacrylamid gel. During the electrophoresis at 1,200 V, 37 mA, 40 W, and 50 °C the fluorescence was detected after excitation with a laser. The lane pattern was analyzed with the Base ImageIR Image Analysis 4.0 software (LI-COR Biosciences) and the sequence was evaluated with Bioedit (Version 7.0.9.0).

Sequencing gel:

10.5	g	Urea
2.5	ml	10x TBE long run
3.75	ml	Formamide
4	ml	Long Ranger® Gel Solution 50%

Fill up to a volume of 25 ml with dH_2O.

25	µl	TEMED
175	µl	10% (w/v) APS

1x TBE long run buffer (pH 8.3-8.7):

134	mM	Tris
45	mM	Boric acid
2.5	mM	EDTA

Store at room temperature.

2.6 Protein and immunochemical methods

The following methods deal with the expression and detection of the ToxR-TMD-MalE fusion protein. Without proof for the presence of fusion protein the results of interaction measurements cannot be trusted.

2.6.1 SDS-PAGE

The separation of proteins according to their molecular size was performed after Laemmli [124]. 200 µl *E. coli* FHK12 were pelleted for 1 min at 11,800 xg and lysed in 50 µl 1x SDS sample buffer. To denaturate the proteins the solution was boiled for 10 min at 110 °C and equivalents of 40 to 100 µl (~10 µl boiled sample) were used for a denaturating SDS-PAGE.

The electrophoretic separation was carried out at 120 V while the samples ran within the 5% stacking gel and at 150 V upon reaching the 10% resolving gel. The separation was finished when the sample buffer ran out of the gel. To compare the protein molecular sizes either the unstained or prestained protein MW marker (Fermentas) was used.

Laemmli running buffer (pH 8.3):

20	mM	Tris
192	mM	Glycine
0.1	%	(w/v) SDS

Store at 4 °C.

2x SDS sample buffer (pH 6.8):

150	mM	Tris
1.2	%	(w/v) SDS
30	%	Glycerol
15	%	β-mercaptoethanol
0.0025	%	Bromphenol blue

Store at room temperature.

5% stacking gel: 20 ml

6.8	ml	dH_2O
1.7	ml	30% acryl-bisacryl-amide mix
1.25	ml	1.5 M Tris (pH 8.8)
0.1	ml	10% (w/v) SDS
0.1	ml	10% (w/v) APS
0.01	ml	TEMED

10% resolving gel: 10 ml

7.9	ml	dH_2O
6.7	ml	30% acryl-bisacryl-amide mix
5	ml	1.5 M Tris (pH 8.8)
0.2	ml	10% (w/v) SDS
0.2	ml	10% (w/v) APS
0.008	ml	TEMED

2.6.2 Western blotting

The Western blot was used to specifically detect and estimate the ToxR-TMD-MalE fusion proteins. First, the proteins were separated by performing a SDS-PAGE and then transferred to a nitrocellulose membrane. A primary antibody which attaches to the maltose binding domain was used to bind the chimeric proteins. The detection was realized with a second antibody coupled to alkaline phosphatase (AP) which catalyzes the staining reaction of BCIP as a substrate and NBT as an oxidant to form a slate precipitation.

The separated proteins were blotted with a semi-dry blotter (CTI Chemical Technologies International) onto a nitrocellulose membrane [125]. 4 layers filter paper soaked with blotting buffer were placed on the anode followed by the soaked nitrocellulose membrane, the facing up SDS gel, and another 4 sheets of soaked paper. To ensure optimal current flux any bubble between the layers was removed. The transfer was carried out at 1 mA/cm^2 for 1.5 h. In case a unstained protein MW marker was used, the membrane was stained reversible with PonceauS solution, the marker lanes were marked with a pencil, and the destaining took place by rinsing the membrane with dH_2O. To prevent unspecific binding of antibody, the membrane was blocked by an incubation in TBS with 3% (w/v) BSA for 1 h at room temperature or over night at 4 °C. Afterwards, the membrane was washed once with TBS for 5 min and then incubated with primary antibody (Rabbit anti MBP antiserum in TBS with 3% (w/v) BSA, 2.4.4, page 31) for 1 h. Unbound antibody was removed by washing three times with TBST for 5 min, before the secondary antibody (Anti rabbit IgG AP conjugate in TBS with 3% (w/v) BSA, 2.4.4, page 31) was applied for 1 h. Again, the membrane was washed 3 times in TBST for 5 min. Finally, the chimeric ToxR protein was detected with staining solution. The reaction was stopped with dH_2O as soon as the protein bands were clearly visible.

Blotting buffer:

1x		Laemmli (2.6.1, page 40)
20	%	(v/v) Methanol

Store at room temperature.

PonceauS solution:

3	%	(w/v) Trichloroacetic acid
0.3	%	(w/v) PonceauS

Store at room temperature.

TBS (pH 7.4):

20	mM	Tris
150	mM	NaCl

Store at room temperature.

TBST (pH 7.4):

		TBS
0.5	%	(v/v) Tween 20

Store at room temperature.

Staining solution:

20	ml	AP buffer
120	µl	NBT solution
60	µl	BCIP solution

Always freshly prepared.

AP buffer (pH 9.5):

100	mM	Tris
100	mM	NaCl
5	mM	$MgCl_2$

Store at room temperature.

BCIP solution:

		Dimethylformamide
5	%	(w/v) BCIP

Store at -20 °C.

NBT solution:

70	%	(v/v) Dimethylformamide
5	%	(w/v) NBT

Store at -20 °C.

2.7 Analysis of self-interacting transmembrane domains

To measure the self-interaction of TMDs the ToxR system was used. Since the chimeric ToxR protein can only interact when the fusion protein is expressed sufficiently, a Western blot was performed to verify the expression level. A second test experiment was required to review the correct insertion of the fusion protein into the inner *E. coli* membrane of PD28 cells. The principle mechanics of those methods are explained in section 1.3.3 on page 13.

2.7.1 Cultivation of ToxR protein expressing FHK12 cells

Chemically competent FHK12 cells were thawed on ice and transformed with pToxRV plasmid DNA accordingly to the transformation protocol in section 2.5.2.1, page 35. Four times 10 µl of the transformation solution was used to inoculate 1 ml FHK12 culture medium using 24-well tissue culture plates. The culture medium contained arabinose solution and IPTG. Usually each plasmid was transformed three times into FHK12 cells and measured in quadruplicates for 12 measurements of biological and technical replicates. Then, the plates were shaken at 200 rpm at 37 °C for 20 h. 5 µl of each culture was used for the ToxR assay and 50 µl each for the Western blot expression test (2.7.4, page 45).

FHK12 culture medium:

		LB medium (2.4.1, page 28)
0.0025	%	(w/v) Arabinose
0.4	mM	IPTG
33	µg/ml	Kanamycin

Always freshly prepared.

Arabinose solution:

1	%	(w/v) L-arabinose

Filter with 0.45 pore size.

Store at 4 °C.

2.7.2 ToxR interaction assay

After the 20 h expression of ToxR fusion proteins the resulting β-Gal activity of each FHK12 culture was measured. Therefore, 5 µl of each culture was transferred to a 96-well plate and diluted with 100 µl chloroform buffer. After mixing the cell density of each suspension was determined in an ELISA plate reader at OD_{600}. By adding 50 µl SDS buffer the cells were lysed at 28 °C for 10 min or until the suspension was cleared up completely. Afterwards, 50 µl ONPG buffer was added and the OD_{405} was measured immediately after mixing for a period of 20 min at 30 s intervals. From the alteration of the absorption at 405 nm per time point the velocity of the hydrolysis reaction of ONPG was calculated which again was used to ascertain the activity of β-Gal in Miller units (MU) [122] using the following formula 2.6:

$$Activity[MU] \quad = \quad \frac{\Delta OD_{405}}{\Delta t[min]} \div OD_{600} \cdot 1000 \tag{2.6}$$

where $\frac{\Delta OD_{405}}{\Delta t[min]}$ is the slope of the linear increase of the measured OD_{405} per min and OD_{600} is the cell density of *E. coli* FHK12 cells in the suspension.

From the activity of β-Gal the relative amount of interacting ToxR-TMD-MalE fusion proteins can be derived. Since only the relative interaction of TMDs can be measured the TMD of GpA and G83A are determined in each experiment as a positive and negative control, respectively. The membrane-spanning leucine zipper AZ2 was used as a reference to separate high interacting TMDs that exceed AZ2 reporter activity. As a negative control ΔTM lacks a TMD and therefore cannot self-interact

The data was acquired by the Softmax Pro 3.0 software from Molecular Devices. The subsequent analysis and evaluation was implemented with Python 2.5.2 and R 2.10.0 scripts. The scripts cover the least square linear regression to determine the slope of the OD_{405} increase, the calculation of means, standard deviations, medians, and quartiles for all samples, the normalization for of all data for the comparison to GpA, as well as the drawing of plots which depict and summarize the acquired data.

Z-buffer (pH 7.0):

60	mM	$Na_2HPO_4 \cdot 7H_20$
40	mM	$NaH_2PO_4 \cdot H_20$
10	mM	KCl
1	mM	$MgSO_4 \cdot 7H_20$

Store at room temperature.

Chloroform buffer:

		Z-buffer
10	%	(v/v) Chloroform
1	%	(v/v) β-mercaptoeth.

Vortex for 60 s. Centrifuge.
Take the aqueous supernatant.

SDS buffer:

		Z-buffer
1.6	%	(w/v) SDS (Sigma L-6026)

Store at room temperature.

ONPG buffer:

		Z-buffer
0.4	%	(w/v) ONPG

Always freshly prepared.

2.7.3 PD28 integration assay

The correct and efficient insertion of ToxR-TMD-MalE fusion proteins into the inner bacterial membrane was tested with the MalE deficient PD28 *E. coli* strain (2.4.2, page 29). Chemical competent PD28 cells were transformed (2.5.2.1, page 35) with pToxRV plasmid DNA similar to the ToxR assay (2.7.1, page 42). 2 ml of each 5 ml overnight liquid culture with tetracycline were centrifuged at 1,200 xg and 4 °C for 10 min to pellet the cells carefully. The supernatant was discarded and the pellet was washed trice in 1 ml PBS and centrifuged again at 1,200 xg and 4 °C for 10 min to remove remaining LB medium. After the last washing step, the cells were resuspended in 800 µl PBS. 20 µl were used to inoculate twice into 5 ml PD28 minimal medium which were incubated in total for 24 h at 37 °C. To determine the cell growth in PD28 minimal medium the OD_{600} was measured the first time after 16 h of incubation by transferring 200 µl of each culture to a 96-well plate. The whole growth kinetic was recorded by repeating the measurement every 2 h. The growth rate was calculated by determining the slope from a linear increase of the cell density over time on a logarithmic scale. Since the PD28 growth rates have to be compared between different experiments a construct including the GpA TMD and another construct without any TMD (ΔTMD) was measured in each experiment as a positive and a negative control, respectively.

As in the case of the ToxR data analysis, the PD28 data was acquired using the Softmax Pro 3.0 software from Molecular Devices and the subsequent analyses, normalization and plotting was completed with Python 2.5.2 and R 2.10.0 scripts.

PBS (pH 7.4):		
10	mM	Na_2HPO_4
1.75	mM	KH_2PO_4
137	mM	NaCl
2.7	mM	KCl

Autoclave.

M9 salts:		
48	mM	$Na_2HPO_4 \cdot 7H_20$
17	mM	KH_2PO_4
8.5	mM	NaCl
19	mM	NH_4Cl

Autoclave.

PD28 minimal medium [122]:		
		M9 salts
0.4	%	(w/v) Maltose
2	mM	$MgSO_4$ (sterile)
1	µg/ml	Thiamin
33	µg/ml	Kanamycin

Always freshly prepared.

Thiamin solution:		
1	mg/ml	Thiamin · HCl

Filter with 0.45 pore size.

Maltose solution:		
10	%	Maltose

Filter with 0.45 pore size.

2.7.4 Western blot expression analysis

In order to compare different expression levels of ToxR fusion proteins a sample of 50 µl of each 24-well FHK12 culture (2.7.2, page 42) was taken, combined for of equal samples, and stored at -20 °C. For each sample 200 µl cell suspension were used for SDS-PAGE (2.6.1, page 39). The proteins in the cell lysate were separated electrophoretically and detected using a Western blot (2.6.2, page 40). The expression levels of each ToxR-TMD-MalE fusion protein were compared manually by the intensity of their lanes.

3 Results

This chapter comprises the results of the systematic assessment of self-interaction of the human bitopic transmembrane domains (TMDs). The first part describes the identification of the similarity threshold and results of the TMD-based clustering of human bitopic membrane proteins based on the similarity threshold identified in this work. The classified proteins were characterized in terms of functional annotation and the occurrence of putative interaction motifs in the second part. The last part presents the results on the homotypic interaction of selected TMDs and includes the mutation of conserved amino acid motifs.

The collaborative character of this project enabled us to share the analytical work between the laboratories in Munich and Jerusalem. For that reason about half of all cloning, Western blots, ToxR assays, and PD28 measurements was performed by Miriam Krugliak in Israel and then collected and interpreted together with the results obtained by the author of this dissertation.

3.1 Classification of human bitopic TMDs

First, the UniProtKB database (2.1.1, page 21) was searched for human proteins resulting in 34,761 protein sequences. After the prediction of transmembrane regions with Phobius (2.2.1.1, page 22) a database of 3,534 bitopic TMDs was obtained, of which 2,205 were unique TMD sequences (2.3.1.1, page 24). The average TMD length was 21.5 amino acids (range: 16-34 amino acids).

3.1.1 Pairwise alignments of TMDs

In order to cluster all 2,205 unique TMD sequences for sequence similarity, pairwise alignments (2.3.1.2, page 25) were calculated for each possible pair of TMDs resulting in a score/selfscore ratio (ssr) matrix (2.3.1.3, page 25) of 2,205 $\times$ 2,205 scores. Each ssr takes a value between 0 and 100%. The mean ssr of TMDs was found to be ssr_{mean} = 37.3% and the median ssr_{median} = 36%.

3.1.2 Identification of the homology threshold for clustering TMDs

Since the level of sequence homology at which TMDs could be clustered in a biologically meaningful way was not known, at first a homology threshold was identified. An easy approach to determine biologically relevant similarities of naturally occurring TMD sequences is to compare the abundance of pairwise alignments of natural and randomized TMD sequences of the same amino acid composition. After randomizing the human bitopic TMD sequences (2.3.1.4, page 26) and calculating the all-against-all TMD similarity, the resulting ssr values were compared to the ssr matrix of the original, non-randomized TMD set. Figure 3.1 shows the superimposed distributions of ssr values for both sequence datasets.

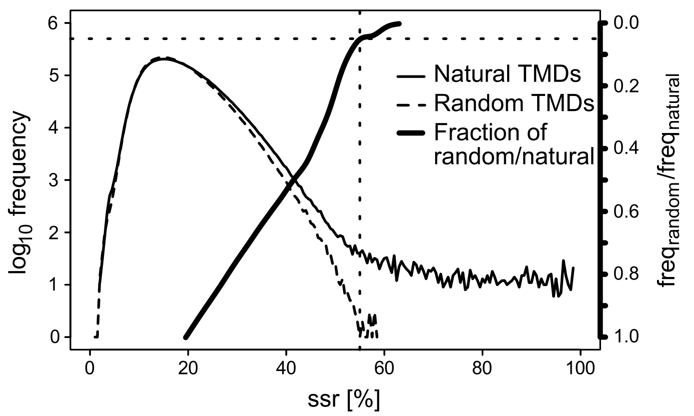

Figure 3.1: Establishing a TMD homology threshold for cluster building. Frequency distributions of ssr values characterizing pairwise alignments of natural TMDs from human bitopic proteins (solid curve) and their randomized counterparts (dashed curve). Calculating the fraction of the ssr values derived from aligning randomized TMDs from the ssr values derived from aligning natural TMDs (bold curve), reveals that the probability of aligning two natural TMDs by chance is $\leq$ 5% using a threshold of ssr = 55% (dotted line).

Both histograms of ssr values show similar distributions up to values of around 50%. In the randomized TMD sequence dataset the occurrence of similarities $> 50\%$ drops and similarity values of ssr $> 60\%$ were not observed. In contrast, the frequency of occurrence of aligned natural TMD pairs exceeds that of randomized TMD pairs by ≥ 20-fold above a similarity threshold of ssr $= 55\%$. In other words, the probability of aligning the average natural TMD pair at random at a $\geq 55\%$ homology is less than 5%.

3.1.3 Clustering TMDs

Clusters of TMDs that share $\geq 55\%$ homology were built by searching the ssr matrix (2.3.1.3, page 25) which was ordered by the appearances of proteins in the UniProtKB. The protein with the TMD that has the largest number of hits was retrieved from the TMD database along with its homologs. The query sequence was chosen as the "most representative" TMD sequence of the first cluster. The procedure was repeated on the reduced database to get the next most representative TMD and therefore the next cluster, followed by further cycles of reduction of the database until no further similarities $\geq 55\%$ could be found.

After clustering the database of human bitopic TMDs using a similarity of $\geq 55\%$, 278 clusters (> 2 members) contained 40.5% (893/2,205) of all human bitopic proteins. The 33 'top' clusters (C1-C3, C5-C31, and C33-C35) included 5 or more unique TMDs, thus covering 13.5% (298/2,205) of the human bitopic proteome. Each cluster contains one most representative TMD which is similar (ssr $\geq 55\%$) to all other members. Table 3.1 on page 50 lists some properties of the top 33 clusters.

Most (20/33) top clusters (C1-C3, C9, C10, C12, C14-C16, C18, C20-C22, C24-C29, and C33) mainly contain proteins of similar biological function (termed "functionally homogeneous") according to the respective annotation in UniProtKB [152] while 13/33 top clusters (C5-C8, C11, C13, C17, C19, C23, C30, C31, C34, and C35) contain sequences with predominantly heterogeneous biological function (in "functionally heterogeneous" clusters $> 40\%$ of the members have functions that deviate from the most prevalent function, table 3.1, page 50). Cluster C4 and C32 had to be removed completely caused by an update to the UniProtKB in October 2009 which revealed most of their members as non-transmembrane.

Table 3.1: The 33 top clusters of human bitopic TMDs.

Cluster	Representative protein[a]	Members[b]	Most prevalent functional annotation[c]	Functional diversity[d] [%]	Representative TMD sequence[e]
C1	Q9UN71	29	Protocadherin	0	lqfYLvvAlaliSvlflvam
C2	P01892	22	HLA class I α-chain	9	ipivGiiAGLvlfgavitga
C3	Q9ULB5	19	Cadherin	16	tgaliailacvltllvlill
C5	P78310	15	Integrin α	47	gliagaiigtLlalaligli
C6	Q6UWB1	15	No prevalent annotation	93	vlpgilflwglfllgcglsl
C7	Q8N967	12	Integrin β	83	gtviiaGvvcGvvcimmvva
C8	Q9BZ76	11	Contactin	55	AviGGviavvifillcitai
C9	P43629	11	Ig-like receptor	18	iliGtSVviilfilLlffll
C10	O75318	10	UDP-guanosyltransferase	0	dVIgFLLacVaTviFiitKf
C11	Q6ZV29	9	Phospholipase	77	ltGiavGallalalvgvlil
C12	P01908	9	HLA class II α-chain	22	vvcalgLsvGlvGivvGtvl
C13	Q9H1U4	8	Syntaxin	63	niiiltviiivvvllmgfvg
C14	Q8IYS5	8	Leukocyte Ig-like receptor	25	gnLvRlglAgLvLisLgalv
C15	Q9Y286	8	Sialic-acid-bind. Ig-like lectin	13	vllgavgGaGatAlvflsfc
C16	P56199	8	Integrin α	13	vplWvillsafaGllllmll
C17	Q8NC67	8	Neuropilin	63	hgtiiGitsgivlvlliisi
C18	P54710	8	Ion transport regulator	25	vrngGlifAglafivGllil
C19	P34810	7	Leucine-rich repeat containing	57	plIiglillgllalvliafC
C20	P13765	6	HLA class II β-chain	0	rkMLsGiaaFlLGLifllvG
C21	Q8NF91	5	Nesprin	20	raalPLqLLlLlliglacLv
C22	Q8IW52	6	SLIT and NTRK like protein	0	iLilsiLvvliltvfvafcl
C23	Q9UGN4	6	CMRF35-like molecule	67	plllsllalLlLllvgasll
C24	Q14DG7	5	Transm. Protein 132 family	0	ALLcVFClAIlvFLiNcvaF
C25	Q14954	5	Killer cell Ig-like receptor	0	HvLIGTSVVkipFtillFfL
C26	P23763	5	VAMP/ Synaptobrevin	20	nckmmImLGaICAiivvviv
C27	Q6PJG9	5	Fibronectin domain containing	0	GGTltvavGGvlVAalLVFt
C28	Q7L4S7	5	Armadillo-repeat containing	0	revGwmaAGlmigAGacYcv
C29	Q14126	5	Desmoglein	20	glgPaaialmilafllLlLv
C30	Q8IZU9	5	Kin of IRRE-like protein 3	60	maviiGvavGaGvaflVlma
C31	Q8IUN9	5	No prevalent annotation	80	pchlllslGlgllllviicv
C33	P32856	5	GRAM domain containing	20	rklmfiiicvivlLviLgii
C34	Q6UXC1	5	Lysosome associated protein	60	svPavvgsallllmllVLlg
C35	Q15262	5	Tyrosine-protein phosphatase	40	vkiagisaGilvfillllvv

[a] UniProtKB identifier of the protein containing the most representative TMD sequence of the cluster.
[b] Number of unique TMDs in the cluster.
[c] Most prevalent functional annotation of cluster members as annotated in UniProtKB.
[d] Fraction of proteins in the cluster which differs from the most prevalent functional annotation.
[e] Representative TMD sequence in optimal orientation for self-interaction (see figure 3.6, page 62). Uppercase amino acids are at least 90% conserved. Bold amino acids were selected for mutation analysis. Underlined SmxxxSm motifs are present in $\geq$ 60% of the members.

3.2 Characterization of human bitopic TMD clusters

After the clustering of TMDs of the human bitopic membrane proteome, clustered TMDs were compared to non-clustered TMDs with respect to enrichment of common amino acid patterns. Furthermore, the similarities of clustered TMDs and the respective complete protein sequences were compared to compare the respective levels of sequence conservation.

First, the frequencies of amino acids and common interaction motifs were calculated for the whole human bitopic TMD dataset including 2,205 sequences (92,413 amino acids). The motif occurrences there were compared to the ones in the randomized dataset (2,205 sequences, 1000 permutations). The results are listed in table 3.2.

Due to the low frequencies of charged residues, TMDs contain a reduced range of amino acids. The enrichment analysis revealed amino acid motifs known to facilitate

Table 3.2: Occurrences of amino acids in human bitopic TMDs and enrichment analysis of Small-xxx-Small motifs in all human bitopic TMDs vs. randomized controls.

Amino acid	Bitopic TMDs[a] $p \cdot 100$ [%]	Motif	Random[b] $p_2 \cdot 100$ [%]	Bitopic TMDs[b] $p_1 \cdot 100$ [%]	Odds ratio[c]
A	10.90	GxxxG	9.03	16.10	1.93*
C	3.18	GxxxA	13.30	20.00	1.63*
D	0.24	GxxxS	4.68	4.04	0.86
E	0.27	AxxxG	13.37	18.05	1.43
F	7.53	AxxxA	17.54	19.37	1.13
G	7.28	AxxxS	7.42	9.25	1.27
H	0.65	SxxxG	4.88	6.21	1.29
I	11.46	SxxxA	7.46	7.39	0.99
K	0.51	SxxxS	3.60	3.08	0.85
L	23.66	GxxxxG	8.56	9.57	1.13
M	2.60	GxxG	9.57	9.43	0.98
N	0.56	GxG	9.81	9.93	1.01
P	1.75	SmxxxxSm	50.96	53.38	1.10
Q	0.58	SmxxxSm	52.54	59.68	1.34
R	0.69	SmxxSm	54.10	56.64	1.11
S	4.66				
T	4.00				
V	14.25				
W	2.29				
Y	2.93				

[a] Percentage of the amino acids in the TMD sequence dataset.
[b] Percentage of TMDs containing the amino acid motif.
[c] Odds ratio calculated with p_1 and p_2 using the method 2.3.2.1 on page 26.
* Statistically significant odds ratio with $p < 0.05$ using a Chi-squared test and the Bonferroni correction for multiple comparisons.

TMD-TMD interaction overrepresented in natural bitopic TMDs compared to the randomized set of TMD sequences. Particularly, GxxxG and GxxxA motifs are significantly enriched in natural, non-randomized TMDs. As control GxxxxG, GxxG, and GxG motifs are not enriched.

3.2.1 Comparison of clustered with non-clustered TMDs

The amino acid composition was also compared between clustered (19,897 amino acids) and non-clustered TMDs (28,229 amino acids) as well as between top cluster TMDs (6,741 amino acids) and non-clustered TMDs. The resulting odds ratios for each amino acid frequency are listed in table 3.3. Sets of small (GAS), apolar (LIV), charged (DEKR), and polar (CHNQSTY) amino acids were also considered.

Table 3.3: Enrichment analysis of amino acids of clustered and non-clustered TMDs.

Amino acid	Non-clustered[a] $p_2 \cdot 100$ [%]	Clustered[a] $p_{1,cluster} \cdot 100$ [%]	Odds ratio[b]	Top clusters[a] $p_{1,top} \cdot 100$ [%]	Odds ratio[b]
A	10.40	11.09	1.07	11.48	1.12
C	3.45	2.95	0.85	2.58	0.74
D	0.24	0.09	0.35	0.00	0.00
E	0.24	0.13	0.53	0.06	0.24
F	8.37	6.56	0.77*	5.73	0.66*
G	6.79	7.98	1.19	9.48	1.44*
H	0.72	0.36	0.50*	0.30	0.41
I	10.68	13.82	1.34*	14.97	1.47*
K	0.51	0.38	0.73	0.46	0.89
L	23.79	23.89	1.01	25.62	1.10
M	2.63	2.51	0.95	2.14	0.81
N	0.57	0.36	0.63	0.15	0.26
P	1.80	1.27	0.70	0.90	0.50*
Q	0.60	0.45	0.75	0.24	0.40
R	0.70	0.38	0.54	0.33	0.47
S	5.08	3.85	0.75*	3.12	0.60*
T	4.46	3.71	0.83	2.95	0.65*
V	13.32	16.26	1.26*	17.30	1.36*
W	2.48	1.65	0.66*	0.79	0.31*
Y	3.15	2.32	0.73*	1.42	0.44*
GAS	22.27	22.92	1.04	24.08	1.11
LIV	47.79	53.97	1.28*	57.88	1.50*
DEKR	1.70	0.97	0.57*	0.85	0.49*
CHNQSTY	18.03	14.00	0.74*	10.76	0.55*

[a] Percentage of the amino acids in the TMD sequence dataset.
[b] Odds ratio calculated with p_1 and p_2 using the method 2.3.2.1 on page 26.
* Statistically significant odds ratio with $p < 0.05$ using a Chi-squared test and the Bonferroni correction for multiple comparisons.

The enrichment analysis revealed an increase in lipophilic amino acids, particularly isoleucine and valine, in clustered TMDs. Charged and polar amino acids are generally underrepresented in clusters. Aromatic amino acids (F, W, Y) as well as histidine and serine also occurred less frequently. It has to be noted that the frequency of amino acids in clustered TMDs is dependent on their overall percentage due the clustering procedure which preferably selects sequences with frequent amino acids. This was partially corrected by using the membrane protein specific substitution matrix, PHAT7573 (2.1.2.1, page 22). The significant odds ratios of H, I, V, W, and Y could thus be explained by their frequency in TMDs. Phenylalanine and serine are less frequently found in clustered TMDs independent of their overall occurrence.

SmxxxSm motifs are often found in TMDs capable of helix-helix interaction (1.3.2.2, page 11). To evaluate a possible connection of the clustering with potential self-interaction motifs, an enrichment analysis was performed. The results of the comparison of clustered (893 sequences) and non-clustered TMDs (1,312 sequences) as well as top cluster TMDs (298 sequences) and non-clustered TMDs are listed in table 3.4.

The probability of occurrence of specific motifs depends on the frequency of included amino acids. Therefore, the clustering procedure enriched motifs including glycine and

Table 3.4: Enrichment analysis of Small-xxx-Small motifs in clustered and non-clustered TMDs.

Motif	Non-clustered[a] $p_{2,single} \cdot 100$ [%]	Clustered[a] $p_{1,cluster} \cdot 100$ [%]	Odds ratio[b]	Top clusters[a] $p_{1,top} \cdot 100$ [%]	Odds ratio[b]
GxxxG	11.66	22.62	2.21*	36.24	4.31*
GxxxA	14.86	27.55	2.18*	37.58	3.45*
GxxxS	4.88	2.80	0.56	1.34	0.27
AxxxG	17.23	19.26	1.15	19.46	1.16
AxxxA	19.21	19.60	1.03	21.14	1.13
AxxxS	9.98	8.17	0.80	6.04	0.58
SxxxG	6.02	6.49	1.08	7.38	1.24
SxxxA	7.77	6.83	0.87	2.68	0.33
SxxxS	3.73	2.13	0.56	1.34	0.35
GxxxxG	7.77	12.21	1.65	17.11	2.45*
GxxG	8.46	10.86	1.32	14.77	1.87
GxG	8.84	11.53	1.34	14.77	1.79
SmxxxxSm	51.75	55.77	1.18	61.41	1.48
SmxxxSm	56.71	64.05	1.36	68.12	1.63
SmxxSm	55.11	58.90	1.17	70.81	1.98

[a] Percentage of TMDs containing the amino acid motif.
[b] Odds ratio calculated with p_1 and p_2 using the method 2.3.2.1 on page 26.
* Statistically significant odds ratio with $p < 0.05$ using a Chi-squared test and the Bonferroni correction for multiple comparisons.

depleted motifs including serine. Odds ratios above 1.0 were not surprising for Gx_nG motifs (including the significant enrichment of GxxxxG), because glycine was found more often in clustered than in non-clustered TMDs. However, GxxxG motifs were much more enriched than other Gx_nG motifs and thus are more frquent in clustered TMDs independent of the enrichment of glycines. Single or multiple GxxxG or GxxxA motifs were found twice as often in clustered sequences (21.7%) and even three times as often in top clusters (35.6%) compared to non-clustered TMD sequences (11.6%). In contrast, the sum of possible SmxxxSm motifs is in more than half of all TMDs nearly independent of clustering (64.1% in clusters, 56.7% in sequences not clustered).

Another method for revealing common amino acid motifs is the analysis of each cluster's alignment for conserved residues. This was accomplished manually using ClustalX2 (2.2.1.3, page 23) for visualization. Since a minimum count of sequences is required to calculate multiple alignments of TMDs, the analysis was performed only for the top 33 clusters (table 3.1, page 50). In 10 top clusters (C2, C5, C7, C11, C12, C17, C20, C28, C30, and C35) the GxxxG motif is conserved in at least 60% of their members. 20 out of 33 clusters (C2, C3, C5, C7-C12, C15-C18, C20, C26-C28, C30, C31, and C35) include a SmxxxSm motif conserved in 60% of the alignment.

3.2.2 Comparison of clustered TMDs with their soluble domains

To explore the relationship between sequence homology and the functional diversity within clusters, pairwise ssr values were calculated at the level of TMD (ssr_{TMD}) and complete sequences ($ssr_{complete}$) between cluster members and their most representative sequence. This comparison resulted in 615 $ssr_{TMD}/ssr_{complete}$ ratios as described in 2.3.2.2 on page 26. The results are depicted as histograms in figure 3.2 on page 55. The distributions of $ssr_{TMD}/ssr_{complete}$ ratios are compared for all clusters, the top clusters, and for functionally heterogeneous top clusters. To equalize the distributions they are plotted on a logarithmic scale.

The distribution of the $ssr_{TMD}/ssr_{complete}$ ratios shows two maxima. Proteins where the homology of TMDs and complete sequences are relatively similar ($ssr_{TMD}/ssr_{complete} < 2.5$) are clearly separated from those proteins where the complete sequences are much more diverse than the TMDs ($ssr_{TMD}/ssr_{complete} > 2.5$). Interestingly, most members of functionally heterogeneous clusters (functional diversity > 40% , table 3.1, page 50) fall into the second class, i.e. they are on average much more homologous at the level of the TMD (average ssr_{TMD} = 62.2%) than at the level of complete sequence ($ssr_{complete}$ = 13.6%). By comparison, functional homogeneous clusters are characterized by an

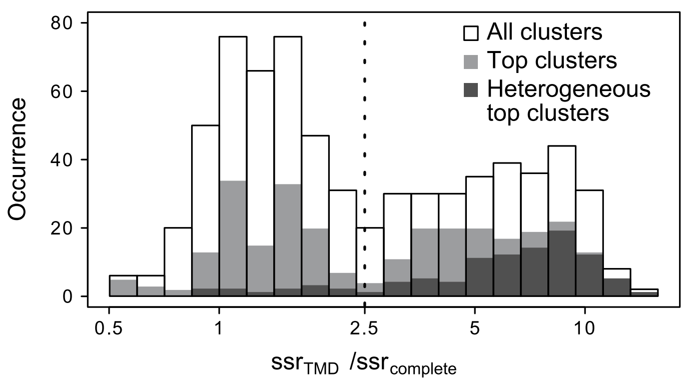

Figure 3.2: Comparison of TMD similarity with complete sequence similarity within TMD clusters. Pairwise alignments of TMDs from cluster members to their most representative TMD sequences were used to calculate TMD similarities for each cluster (ssr_{TMD}). Similarly, the corresponding complete protein similarities ($ssr_{complete}$) were calculated. If a TMD is more similar to the representative sequence than the complete protein sequence, the $ssr_{TMD}/ssr_{complete}$ ratio is > 1. The distributions of these ratios were compared for all 298 clusters (white bars), the 33 top clusters (light gray bars), and the subset of functionally heterogeneous top clusters (dark gray bars). Bars are superimposed and not cumulative. Note that the higher abundance of $ssr_{TMD}/ssr_{complete} > 2.5$ within functionally heterogeneous clusters indicates a much higher TMD similarity relative to complete sequence similarity in this subset.

average $ssr_{TMD} = 71.3\%$ and $ssr_{complete} = 38.8\%$. In contrast, the members of all top clusters or all clusters show no preference for higher TMD than complete sequence homology. Since the complete sequences consist mostly of extramembranous sequences, these results indicate that the extramembranous domains within functionally heterogeneous clusters are structurally non-homologous while those of functionally homogeneous clusters are. It has to be noted that the distribution of $ssr_{TMD}/ssr_{complete}$ ratios is shifted to larger values, since TMDs are less diverse in their amino acid composition compared to soluble domains.

3.2.3 Comparison of functionally homogeneous and heterogeneous clusters of TMDs

In general, functionally heterogeneous clusters are on average much less homologous at the level of the complete sequence than functionally homogeneous clusters (see 3.2.2, page 54). An possible correlation between sequence similarity and functional diversity of clusters was tested with Spearman's correlation test for not normally distributed data. By testing for correlation of similarity (average ssr) and functional diversity (table 3.1, page 50) for top clusters, the correlation coefficient was found to be ρ = -0.67

($p = 1.8 \times 10^{-5}$) which indicates that decreasing TMD similarity is linked to increasing functional diversity. In case of complete sequence similarity, the correlation was even stronger with $\rho = -0.71$ ($p = 3.0 \times 10^{-6}$). These results support not only that complete protein sequences but also TMD sequences are responsible for the function of proteins.

To investigate the differences between functional heterogeneous (113 sequences, 2,568 amino acids) and homogeneous clusters (185 sequences, 4,185 amino acids) clusters, the amino acid composition of their TMDs were compared in the top 33 clusters. The results are listed in table 3.5. Neither G, A, nor S was found significantly enriched or depleted in functional heterogeneous clusters.

Due to the low number of sequences in top clusters, the enrichment analysis of amino acid motifs in functional heterogeneous clusters yields no significant results. Yet, the

Table 3.5: Enrichment analysis of amino acids of heterogeneous and homogeneous top clusters of TMDs.

Amino acid	Functional homogeneous[a] $p_2 \cdot 100$ [%]	Functional heterogeneous[a] $p_1 \cdot 100$ [%]	Odds ratio[b]
A	11.88	10.98	0.92
C	2.84	2.02	0.71
D	0.00	0.00	-
E	0.00	0.16	-
F	7.00	3.58	0.49*
G	8.53	10.98	1.32
H	0.36	0.27	0.76
I	13.38	17.33	1.36
K	0.57	0.27	0.47
L	25.50	26.21	1.04
M	2.20	1.95	0.88
N	0.24	0.00	0.00
P	0.88	0.82	0.92
Q	0.36	0.08	0.22
R	0.38	0.23	0.61
S	3.13	3.08	0.98
T	3.15	2.73	0.86
V	17.51	17.21	0.98
W	0.62	0.93	1.51
Y	1.46	1.17	0.80
GAS	23.54	25.04	1.09
LIV	56.39	60.75	1.20
DEKR	0.96	0.66	0.69
CHNQSTY	11.54	9.35	0.79

[a] Percentage of the amino acids in the TMD sequence dataset.
[b] Odds ratio calculated with p_1 and p_2 using the method 2.3.2.1 on page 26.
* Statistically significant odds ratio with $p < 0.05$ using a Chi-squared test and the Bonferroni correction for multiple comparisons.

odds ratios delineate an increased incidence of GxxxG, AxxxG, SxxxG, and pooled SmxxxSm motifs. It has to be proven whether the existence of such motifs effect the self-interaction. By looking at the conserved GxxxG motifs in top clusters, 5/10 motifs were found in functional heterogeneous clusters (C5, C7, C17, C30, C35).

3.2.4 Extension of TMD clusters via complete sequence similarity

To compare the TMD-based clustering to the more traditional approach of using complete sequences, the clusters were extended by including bitopic proteins aligned on the basis of complete sequence similarity. A sequence conservation level of 25% is known to signify structural homology as shown by analyses of x-ray structures [146]. Since extramembranous domains do not include position dependent transmembrane interaction motifs, $ssr_{complete}$ were calculated (2.3.1.3, page 25) allowing gaps. If the $ssr_{complete}$ between a representative protein sequence and a non-clustered protein was $\geq 25\%$, the sequence was added to the existing cluster. Furthermore, new clusters of non-clustered complete sequences were created by calculating $ssr_{complete}$ between all complete protein sequences that were not a member of any cluster. A graphical representation of the human bitopic membrane protein clusters and their extension by complete sequences is given in figure 3.3 on page 58.

In total 95 bitopic transmembrane proteins were added to the existing clusters and 156 sequences formed new clusters based on the complete sequence similarity. The fraction of clustered proteins increased from 40.5% to 51.9% (1,144/2,205) of all human bitopic proteins by applying the combined clustering procedure using 55% similarity for TMD sequences and 25% for complete proteins. This small increase in coverage implies that our TMD-based clustering explores the bitopic protein sequence to a large proportion. We also compared the extension of functionally homogeneous and heterogeneous top clusters by complete sequence similarity. Homogeneous clusters gain 12% of new members while heterogeneous ones grow only by 4%.

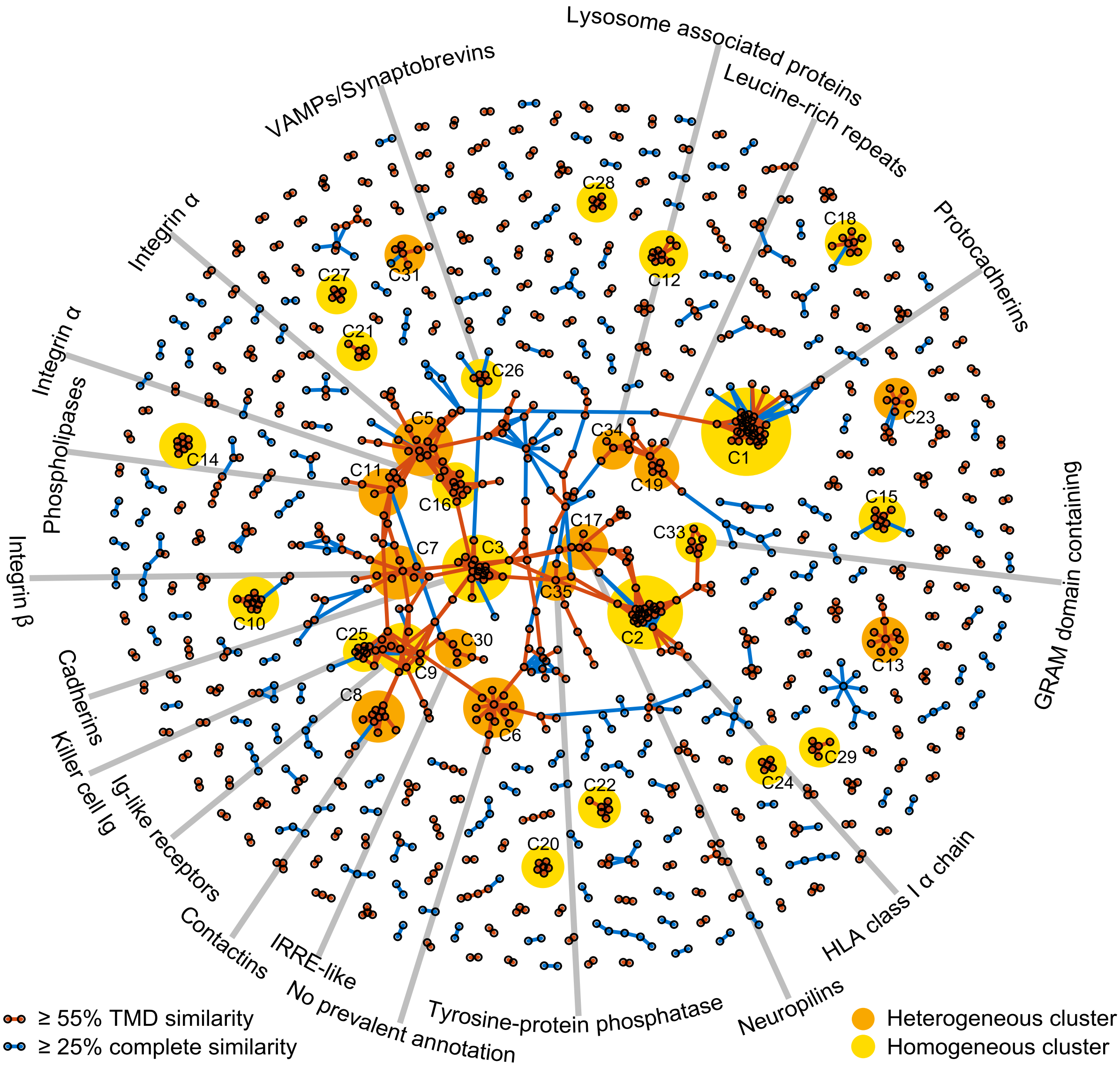

Figure 3.3: A graphical representation of human bitopic membrane protein clusters. Alignments based on TMD sequence similarities $\geq$ 55% (red connections) generated 298 clusters. These were extended by including proteins whose complete sequence similarities is $\geq$ 25% (blue connections). The 33 top TMD clusters are underlaid with circular areas whose diameters reflect the size of the TMD-based clusters and whose color reflects functional homogeneity (yellow) or heterogeneity (orange) of their members (see 3.1.3 on page 49 for details). The main annotation is assigned to each functionally homogeneous cluster. Note that heterogeneous clusters include proteins with functions differing from the main annotation. Connections between different clusters arise from similarities of TMDs or complete sequences which are homolog to either the connected clusters. The CLANS software (2.2.1.4, page 23) was used to calculate the coordinates of TMDs.

3.3 Homotypic TMD-TMD interaction

The ToxR system (1.3.3.1, page 13) was used to measure the self-interaction of TMDs via the activity of β-Galactosidase (β-Gal) reporter enzyme (2.7.2, page 42). TMD sequences were introduced into the ToxR chimeric protein by ligation of a respective oligonucleotide cassette (2.5.7, page 37) between the *Nhe*I and *Bam*HI sites of the pToxRV plasmid (figure 2.2, page 30). To keep the lengths of the TMDs comparable, inserted segments had a fixed length of 20 amino acids. The ToxR protein expression was tested by Western blot expression analysis (2.7.4, page 45). The efficiency of membrane integration was controlled by complementing the MalE deficiency of *E. coli* PD28 cells (2.7.3, page 43). ToxR proteins were considered sufficiently expressed and correctly integrated into the membrane when the slope of the growth curve was at least 50% of that of GpA.

3.3.1 Self-interaction of representative TMDs in different orientations

The representative TMDs from the 33 top clusters were now tested for self-interaction using the ToxR transcription activator assay (2.7.2, page 42). Affinities were determined relative to the high-affinity TMD of GpA and its non-interacting mutant G83A that is thought to produce a non-specific background signal [63]. Initially, the optimal orientation of the potentially interacting faces of the TMDs relative to the DNA-binding ToxR domain was determined. Assuming α-helicity of the TMDs, stepwise insertion of three additional residues at their N-terminus concurrent with the stepwise deletion of three residues at their C-termini rotates the potential TMD-TMD interface by up to $3 \times 100°$, i.e. almost a full helix turn, relative to the ToxR domain. Therefore, each representative TMD from table 3.1 on page 50 had to be cloned four times by shifting a 20 amino acid window along the TMD sequence as shown exemplarily in figure 3.4 on page 60.

The orientation influences the coupling of TMD-TMD interaction to transcription activation (1.3.2, page 9) in cases where the TMD interface is specific for one helix side. The results of the ToxR interaction assays are shown in figure 3.5 on page 61 and in table 7.1 on page 109. To determine a numeric value for the orientation-dependence of TMDs, the difference of ToxR reporter activity between highest and lowest self-interacting orientation was calculated as described in 2.3.3.1 on page 27. The results of the Western blot expression analysis showed sufficient protein expression (figure 7.6A,

```
Q9UN71|PCDGG_HUMAN
 n-...DPSDLQAELQFYLVVALALISVLFLVAMILAIALRLRRS...-c
                .........1.........2...

Q9UN71-0      LQFYLVVALALISVLFLVAM
Q9UN71-1       QFYLVVALALISVLFLVAMI
Q9UN71-2        FYLVVALALISVLFLVAMIL
Q9UN71-3         YLVVALALISVLFLVAMILA
```

Figure 3.4: Cloning strategy of TMD sequences exemplified by the protocadherin gamma-B4 sequence (UniProtKB ID: Q9UN71, representative TMD of cluster C1). The first line demonstrates a short fragment of the complete protein sequence of protocadherin gamma-B4 with the section predicted to be transmembranous (red). A segment of 20 amino acids length was selected for the introduction into the ToxR-TMD-MalE fusion protein preferably including highly conserved residues. Stepwise insertion of three additional residues at their N-terminus concurrent with the stepwise deletion of three residues at their C-termini rotate the potential TMD-TMD interfaces by up to 3 × 100°. By creating primers for the four selected TMD fragments (Q9UN71-[0-3]) the 17 amino acid long core sequence (blue) was introduced into the fusion protein coding sequence in four different orientations.

page 108). The PD28 integration assay results which control for membrane insertion are depicted in figure 7.1 on page 104 and show that all clones except C25 expressed correctly membrane-integrated ToxR proteins.

The relative affinities of 6 TMDs (C5, C11, C12, C15, C26, and C28) show a clear preference for one TMD orientation (figure 3.5 on page 61, dark shading, “strong orientation-dependence”). Another 9 TMDs (C6-C8, C10, C19, and C21-C24) display affinities with little dependence on orientation (light shading, “weak orientation-dependence”). The affinities of the other tested TMDs are independent of orientation (no shading, “no orientation-dependence”). The orientation-dependence was found to be independent of the cluster size (implied by the order of cluster numbers). Cluster C25 was removed, because of its insufficient membrane integration (figure 7.1, page 104) and was disregarded in further analyses. The orientations of representatives TMD exhibiting the highest relative affinity were considered to represent the optimal orientation.

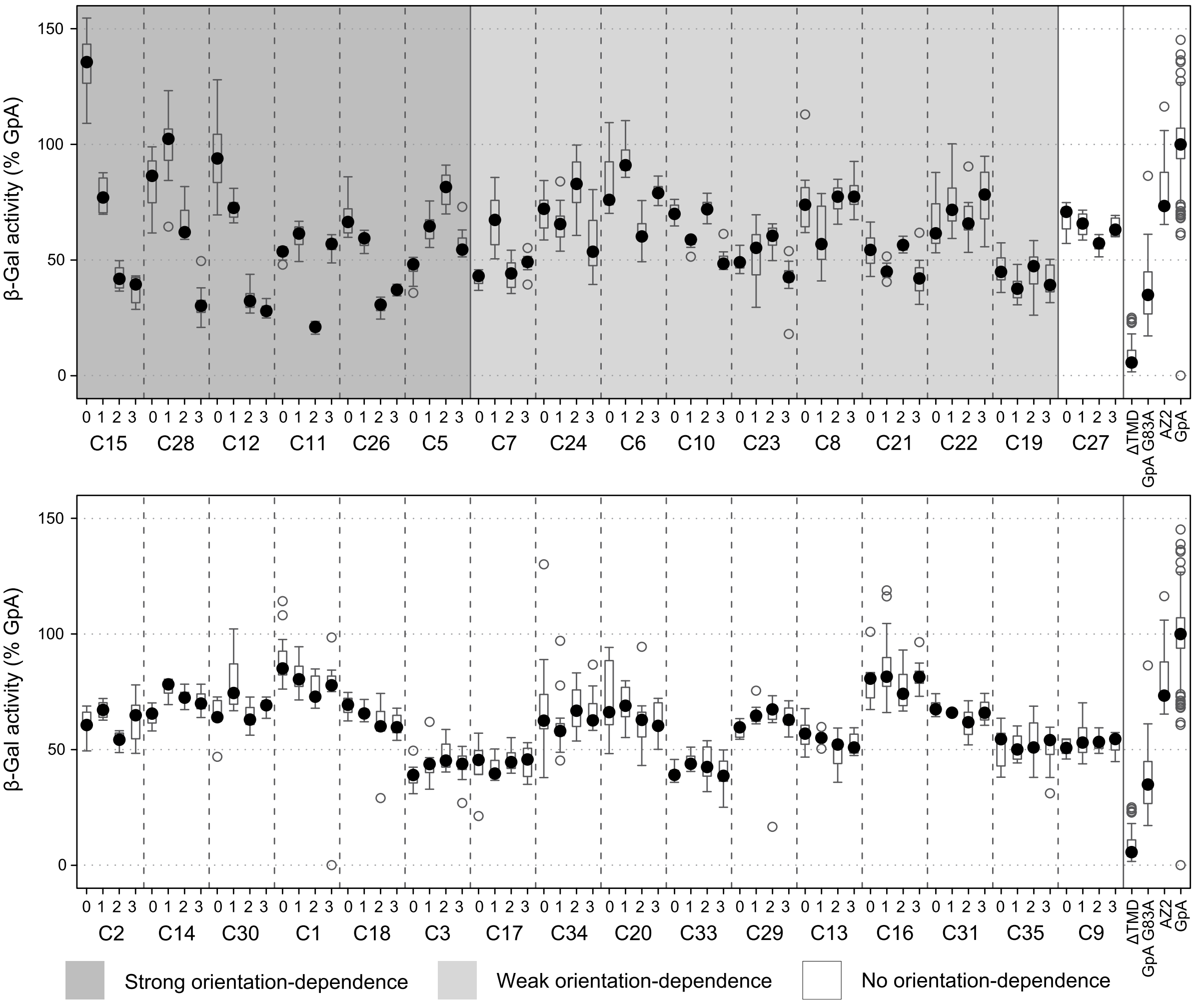

Figure 3.5: Dependence of self-interaction on the orientation of the TMD relative to the ToxR domain for each top cluster's most representative TMD. Data represent relative β-Gal activities (GpA = 100%) as measured with the ToxR system (dot: median, box: interquartile range (IQR), whiskers: upper/lower quartile with max. 1.5 × IQR). The results are sorted according to the orientation-dependence of interaction. 6 TMDs show > 40% difference in relative β-Gal activity demonstrating strong orientation-dependence (dark shading), 9 TMDs show weak orientation-dependence with 20-40% difference in relative β-Gal activity (light shading), whereas the self-interaction of 17 TMDs differs < 20% and is clearly unaffected by orientation (no shading). Precise numbers for relative affinities are listed in table 7.1 on page 109. The results of the PD28 assay that controls for membrane insertion are shown in figure 7.1 on page 104.

3.3.2 Self-interaction of representative TMDs in optimal orientation

The comparison of self-interaction of representative TMDs in optimal orientation is depicted in figure 3.6 and in the table 7.2 on page 111. The respective PD28 integration assay results can be found in figure 7.2 on page 105.

The relative affinities of TMDs in their optimal orientation cover a broad range from ~40% to ~135% of the GpA signal. All β-Gal activities exceed that of the GpA G83A mutant. Around one third (12/32) of representative TMDs elucidate β-Gal activities that exceed that of the membrane-spanning leucine zipper AZ2 and therefore are denoted as "strongly self-interacting". Affinities below that of AZ2 but above GpA G83A are denoted as "weakly self-interacting". Notably, high-affinity TMDs tend to be orientation-dependent (Spearman's correlation test between median β-Gal and ORD values, $\rho = 0.48$, p = 0.005).

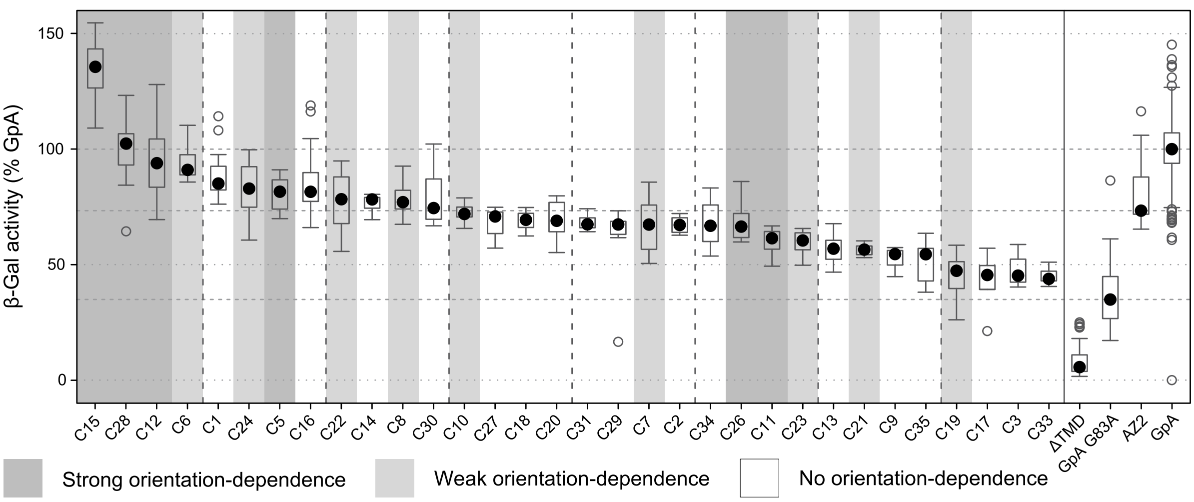

Figure 3.6: Self-interaction of most representative TMDs in optimal orientation. Data represent relative β-Gal activities (GpA = 100%) as measured with the ToxR system (dot: median, box: interquartile range (IQR), whiskers: upper/lower quartile with max. 1.5 $\times$ IQR). Self-interaction of TMDs in their optimal orientation, as identified in figure 3.5 on page 61, ordered by decreasing affinity. TMDs showing orientation-dependent self-interaction are shaded. TMDs used for reference are explained in the text. The results of the PD28 assay that controls for membrane insertion are shown in figure 3.5 on page 61

3.3.3 Comparison of orientation-dependent with orientation-independent self-interaction of TMDs

In order to compare TMDs that exhibit a clear preference for one TMD orientation with those that are weakly orientation-dependent, the top 32 clusters were categorized in two groups. The clusters containing TMDs with more than 40% loss in reporter activity between different orientations constitute the strongly orientation-dependent dataset (6 clusters, 1,113 amino acids). The remaining clusters comprise the dataset of weakly and not orientation-dependent TMDs (26 clusters, 5,470 amino acids). Both datasets were compared for their amino acid composition. The results are listed in table 3.6.

Table 3.6: Enrichment analysis of amino acids in clusters containing orientation-dependent representative TMDs.

Amino acid	Weakly and not orientation-dependent[a] (ORD < 0.4) $p_2 \cdot 100$ [%]	Strongly orientation-dependent[a] (ORD > 0.4) $p_1 \cdot 100$ [%]	Odds ratio[b]
A	11.37	13.03	1.20
C	2.05	3.05	1.23
D	0.00	0.00	-
E	0.05	0.09	1.64
F	5.85	4.04	0.74
G	8.34	15.45	2.01*
H	0.27	0.01	0.00
I	14.81	15.54	1.06
K	0.29	1.08	3.72
L	26.64	21.11	0.74
M	1.97	2.96	1.52
N	0.18	0.01	0.00
P	0.09	0.27	0.30
Q	0.26	0.18	0.70
R	0.38	0.09	0.23
S	3.18	2.61	0.81
T	2.96	2.61	0.88
V	17.99	14.73	0.79
W	0.71	0.99	1.39
Y	1.33	1.53	1.15
GAS	22.89	31.36	1.54*
LVI	59.43	51.39	0.72
DEKR	0.73	1.26	1.73
CHNQSTY	10.69	9.97	0.93

[a] Percentage of the amino acids in the dataset of TMDs.
[b] Odds ratio calculated with p_1 and p_2 using the method 2.3.2.1 on page 26.
* Statistically significant odds ratio with $p < 0.05$ using a Chi-squared test and the Bonferroni correction for multiple comparisons.

Only glycines and the pooled small amino acids of G, A, and S were significantly overrepresented in clusters of strongly orientation-dependent TMDs. Interestingly, 4/6 strongly orientation-dependent TMDs (C5, C11, C12, and C28) contain a GxxxG motif conserved in at least 60% of their homologs (table 3.1 on page 50). In contrast, only 6/26 weakly or not orientation-dependent TMDs (C2, C7, C17, C20, C30, and C35) share this pattern. Although the co-occurrence of GxxxG with strong orientation-dependence is not statistically significant due to the low number of cases (odds ratio = 6.20, p = 0.06, Fisher's exact test), orientation-dependent TMDs tend to contain conserved GxxxG motifs. Furthermore, conserved SmxxxSm motifs were also contained more often in orientation-dependent TMDs (6/6 in strongly orientation-dependent TMDs, 14/26 in weakly or not orientation-dependent TMDs).

Orientation-dependence and functional diversity were not found to be correlated (Spearman's correlation test between functional diversity and ORD values, ρ = -0.04, p = 0.84). Also, the orientation-dependence of TMDs was not correlated to the average ssr of top clusters (Spearman's correlation, p = 0.10) and not correlated to the $ssr_{TMD}/ssr_{complete}$ ratios of top clusters (Spearman's correlation, p = 0.86) of clusters.

3.3.4 Conserved self-interaction within clusters of TMDs

To examine whether the relative affinities of TMDs are conserved within clusters, 8 clusters (C3, C7-C9, C12, C15, C30, and C31) of different interaction strength, functional diversity, and GxxxG content were selected (table 3.7, page 65). From each of these clusters, the self-interaction of 2-5 TMDs was determined. The TMD sequences were chosen for various homologies (different ssr_{TMD}) to the cluster's representative TMD. Cluster C12 was analyzed by Felix Behr in the course of his Bachelor's thesis. The results are presented in figure 3.7 on page 65 as well as in table 7.3 on page 112. The Western blot expression analysis resulted in sufficient protein expression (figure 7.6B, page 108). The outcomes of the PD28 integration assay are depicted in supplementary figure 7.3 on page 106 and show correct membrane insertion for all clones.

In 6/8 chosen clusters (C3, C7-C9, C12, and C31) the relative affinity was comparable (± 20%) to the most representative sequences. There was no correlation between the average variation of relative affinities within clusters and the presence of putative interaction motifs (table 3.7, page 65). The average variation of relative affinities also was not correlated to orientation-dependence (Spearman's correlation test between AV and ORD values, p = 0.20) and not correlated to functional diversity of corresponding clusters (Pearson's correlation test between AV values and functional diversity, p = 0.82).

Table 3.7: Selected clusters for the comparison of relative affinities within clusters.

Cluster	Conserved motif[a]	Functional diversity[b]	ToxR median[c]	Orientation-dependence[d]	Members[e]	N[f]	AV[g]
C3	`GxxxA`	16	45.2	0.14	19	5	7.1
C9	-	18	54.6	0.07	11	4	7.4
C12	`GxxxG`	22	93.9	0.70	9	4	12.0
C8	`GxxxA`	55	77.1	0.27	11	3	13.3
C31	`SxxxG`	80	67.4	0.08	5	3	15.1
C7	`GxxxG`	83	67.3	0.36	12	5	19.5
C30	`GxxxG`	60	74.5	0.15	5	2	24.7
C15	-	13	135.6	0.71	8	2	34.9

[a] Conserved putative interaction motif in the TMD cluster alignment.
[b] The fraction of proteins in the cluster which differs from the main functional annotation.
[c] The median (>50%) of β-Gal activity measured for the TMD in % of GpA.
[d] The value for the orientation-dependence (ORD) between 0 and 1 for the four orientations (2.3.3.1, page 27). Small values indicate a low dependence of β-Gal activity on TMD orientation.
[e] The number of unique TMDs in the cluster.
[f] The number of TMDs compared to the representative TMD of the cluster.
[g] Average variance of median β-Gal activity of TMDs measured for this cluster as percentage of the median (>50%) of β-Gal activity of the representative TMD.

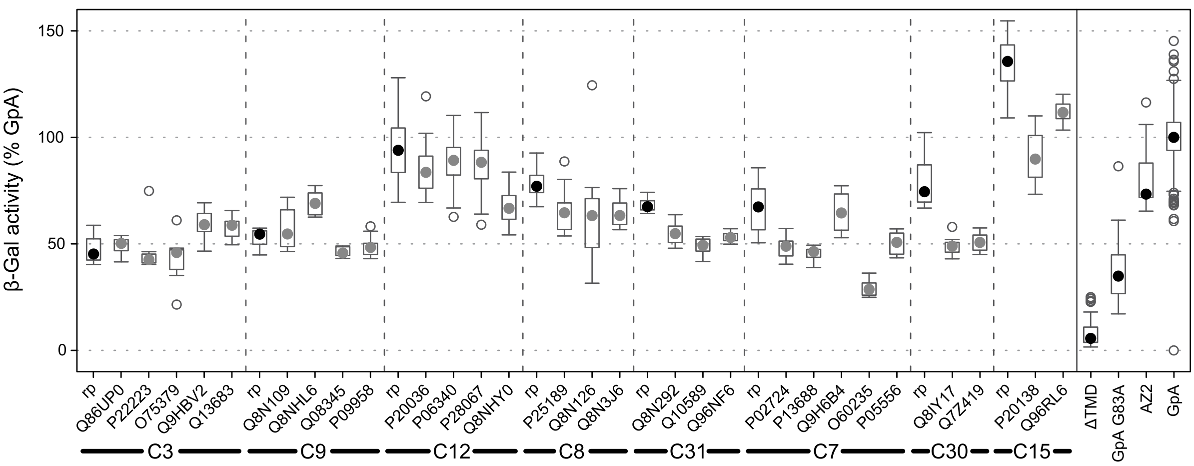

Figure 3.7: Conservation of self-interaction within exemplary clusters. Data represent relative β-Gal activities (GpA = 100%) as measured with the ToxR system (dot: median, box: interquartile range (IQR), whiskers: upper/lower quartile with max. 1.5 × IQR). Clusters are sorted in descending order of the average conservation of their member's self-interaction. TMDs within each cluster are ordered according to descending similarity to their representative TMD (rp) from left to right. TMDs used for reference are explained in the text. Precise numbers for relative affinities are listed in table 7.3 on page 112. The results of the PD28 assays that controls for membrane insertion are shown in figure 7.3 on page 106.

However, the extend by which cluster member's relative affinities vary increases with the median of relative affinities (Pearson's correlation test between AV values and median β-Gal activity, ρ = 0.80, p = 0.02), presumably because higher β-Gal activities offer more potential for variation.

3.3.5 Sequence-specificity of self-interacting TMDs

12 representative TMDs (C1, C2, C8, C12, C14-C16, C20, C26, C28, C30, and C34) were mutated to assess the sequence-specificity of self-interaction (table 3.8). The mutations target mainly SmxxxSm motifs and highly conserved amino acids (table 3.1, page 50, bold type). SmxxxSm motifs (including GxxxG) are known to facilitate or support self-interaction of TMDs (1.3.2.2, page 11). 8 representative TMDs (C2, C8, C12, C16, C20, C26, C28, C30) were selected for mutation analysis of such common interaction motifs. Another 4 representative TMDs (C1, C14, C15, C34) that contain highly conserved

Table 3.8: Representative TMDs selected for the mutation of putative interaction motifs and/or highly conserved residues.

TMD[a]	Average ssr[b]	Average ssr_{TMD}/ $ssr_{complete}$[c]	Functional diversity[d]	ToxR median[e]	Orienta-tion-de-pendence[f]	Mutated motif/ residues[g]	Maximal impact of mutation[h]
C15-0	68	66.3	13	135.6	0.71	`GxxA`	0.79
C1-0	67	74.9	0	85.1	0.14	`YxxxAxxxxS`	0.63
C8-3	71	42.0	55	77.1	0.27	`GxxxA`	0.60
C28-1	69	22.9	0	102.4	0.71	`GxxxG`	0.44
C12-0	75	64.1	22	93.9	0.70	`GxxxGxxGxxxG`	0.37
C20-1	73	74.9	0	69.0	0.13	`AFxxGxxF`	0.34
C30-1	67	47.0	60	74.5	0.15	`GxxxG`	0.24
C34-2	59	10.2	60	66.8	0.13	`PxxxG`	0.18
C14-1	82	25.7	25	78.2	0.16	`GxxG`	0.16
C2-1	75	106.4	9	67.2	0.19	`GxxxG`	0.14
C26-0	78	93.1	20	66.5	0.54	`GxxxA`	0.11
C16-1	67	23.3	13	81.5	0.09	`WxxxxSxxxG`	0.09

[a] Most representative TMD of the cluster.
[b] The average ssr calculated from ssr (2.3.1.3, page 25) between each cluster member TMD and the representative TMD.
[c] The average $ssr_{TMD}/ssr_{complete}$ ratio (2.3.2.2, page 26) for the cluster.
[d] The fraction of proteins in the cluster which differs from the main functional annotation.
[e] The median (>50%) of β-Gal activity measured for the TMD in % of GpA.
[d] The value for the orientation-dependence (ORD, 2.3.3.1, page 27) between 0 and 1 for the four orientations. Small values indicate a low dependence of β-Gal activity on TMD orientation.
[g] The conserved amino acids or the potential interaction motif which were mutated.
[h] Maximal impact of point mutations (MIM, 2.3.3.2, page 28) calculated from β-Gal activities of wild-type and mutated form of the representative TMD.

residues were investigated by mutation analysis. Mutagenesis primers were created with a minimum of exchanged nucleotides resulting in the desired amino acid mutation. The influences of mutation on β-Gal activity are shown in figure 3.8 and in table 7.4 on page 113. The outcomes of Western blot analysis exhibit sufficient protein expression (figure 7.6C, page 108). The PD28 integration assay results are depicted in figure 7.4 on page 106 and show correct membrane insertion for all clones.

Depending on the TMD and the type of targeted residue, mutation reduced the relative affinity by up to 79% of the wild-type TMD (IPM, equation 2.5, page 28). For 6/12 TMDs (C1, C8, C12, C15, C20, and C28) the relative affinity dropped by $\geq$ 30% (denoted "mutation-sensitive") while $<$ 30% reduction is seen for the other mutated TMDs ("mutation-insensitive"). Although mutating glycines had strong effects in 5/12 cases (C8, C12, C15, C20, and C28), several GxxxG or other SmxxxSm motifs are insensitive to mutation. One has to note that mutation-sensitivity is technically more difficult to establish for those TMDs where self-interaction is lower and therefore closer to the un-

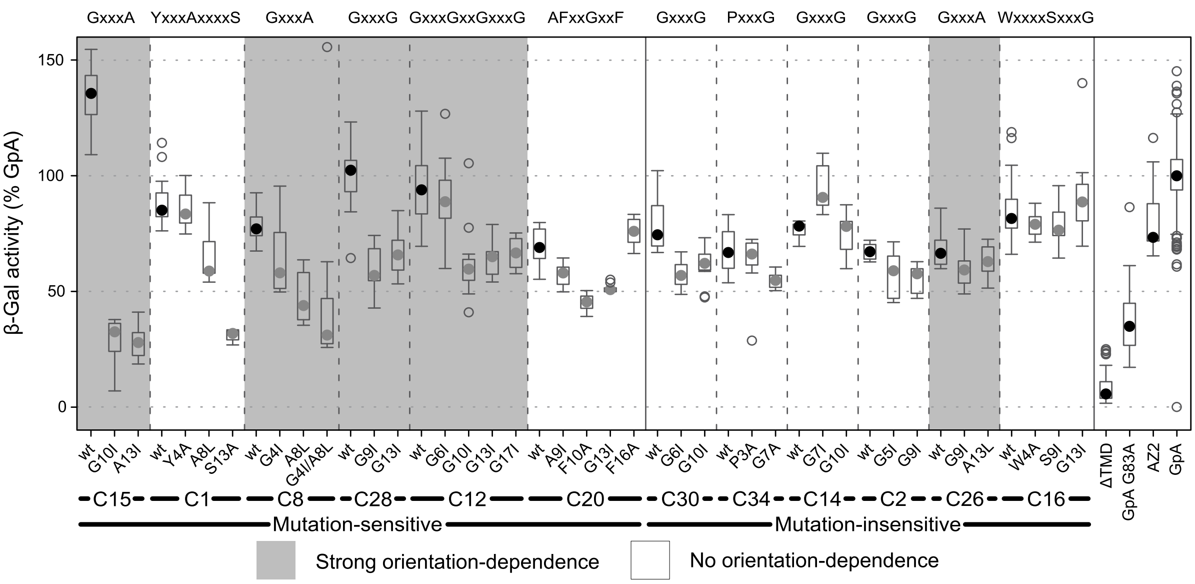

Figure 3.8: Sequence-specificity of self-interaction. Twelve exemplary sequences were mutated by exchanging the most conserved residues within the respective alignments. Data represent relative β-Gal activities (GpA = 100%) as measured with the ToxR system (dot: median, box: interquartile range (IQR), whiskers: upper/lower quartile with max. 1.5 $\times$ IQR). The wild-type TMDs (wt, black dots) are sorted from left to right in descending order of the maximal impact of the mutations on self-interaction. Putative interaction motifs are depicted on top. TMDs showing orientation-dependent self-interaction (see figure 3.5, page 61) are shaded. TMDs used for reference are explained in the text. Precise numbers for relative affinities are listed in table 7.4 on page 113. The results of the PD28 assay that controls for membrane insertion are shown in figure 7.4 on page 106.

specific signal elicited by GpA G83A. Thus, the maximal impact of point mutations was found to be correlated to the β-Gal activity of the wild-type TMD (Spearman's correlation test between MIM and median β-Gal activity, $\rho = 0.63$, p = 0.03). The maximal impact of mutations was not correlated to the average ssr of a cluster (Pearson's correlation test between MIM and average ssr values, p = 0.46) and also not correlated to the average $\mathrm{ssr_{TMD}}/\mathrm{ssr_{complete}}$ ratio (Pearson's correlation test between MIM and average $\mathrm{ssr_{TMD}}/\mathrm{ssr_{complete}}$ values, p = 0.79). Furthermore, there was no dependence between maximal impact of mutations and functional diversity (Spearman's correlation test between MIM and functional diversity values, p = 0.5). Although the maximal impact of mutations was not correlated with statistical significance to orientation-dependence (Spearman's correlation test between MIM and ORD values, $\rho = 0.43$, p = 0.15) due to the low number of cases, orientation-dependent TMDs tend to contain more mutation-sensitive amino acids than not orientation-dependent TMDs. These residues do not necessarily belong to GxxxG or other SmxxxSm motifs (e.g. C1 protocadherin).

3.3.6 Homotypic interaction of HLA class II α-chains

The HLA class II α-chain TMD (C12) was scanned systematically for critical amino acids. Two GxxxG motifs (table 3.1, page 50, C12) are located on one side of the HLA transmembrane helix. By mutating almost every amino acid of the representative TMD (HLA class II, DQ(1) α-chain), possible interaction motifs and the significance of GxxxG motifs were expected to show up. Therefore, glycines were exchanged for much larger β-branched isoleucines and their adjacent mostly large and aliphatic residues were mutated to alanine. The impact of mutations on self-interaction as well as the expression and membrane integration of the constructs were tested by Manuel Mohr in the course of his Bachelor's thesis. Figure 3.9 on page 69 and table 7.5 on page 114 presents the results of the ToxR assays. The results of the Western blots and PD28 assays are shown in figure 7.6C on page 108 and figure 7.5 on page 107, respectively.

The wild-type TMD of HLA class II α-chain exhibited a relative affinity of about 94% of GpA and therefore was the third-strongest interacting representative TMD of the 33 top clusters (figure 3.6, page 62). The mutation scanning analysis revealed G_{10}, G_{13}, and G_{17} to be the most important residues for self-interaction. G_6 is located at the N-terminus of the TMD and did not decrease di- or oligomerization when exchanged to alanine. All but one of the double, triple, and quadruple mutants of glycines reduced the β-Gal activity to about 65% of that of GpA. Surprisingly, the double mutant of the C-terminal $G_{13}xxxG_{17}$ motif restored the relative affinity to wild-type level.

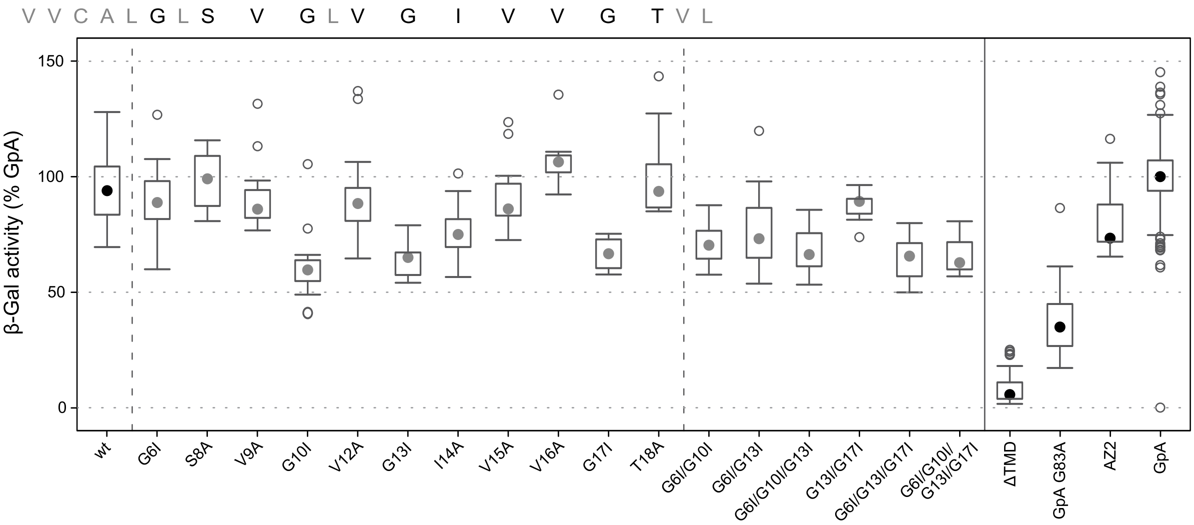

Figure 3.9: Specific self-interaction of the HLA class II α-chain TMD. The representative TMD of cluster C12 (wt) was mutated by exchanging most residues. Data represent relative β-Gal activities (GpA = 100%) as measured with the ToxR system (dot: median, box: interquartile range (IQR), whiskers: upper/lower quartile with max. 1.5 $\times$ IQR). The HLA class II α-chain TMD sequence is depicted on top showing mutated amino acids in black and non-mutated ones in gray. The mutated TMDs are sorted for the position of the exchanged amino acid. TMDs used for reference are explained in the text. Precise numbers for relative affinities are listed in table 7.5 on page 114. The results of the Western blots and PD28 assays are shown in figure 7.6C on page 108 and figure 7.5 on page 107, respectively.

3.3.7 Test for correlation of TMD affinity with membrane insertion

In this work, a total of 204 different constructs were tested for TMD-TMD self-interaction using the ToxR transcription activator assay (2.7.2, page 42). For control, each construct was tested further for its proper insertion into the inner bacterial membrane by determining its ability to complement the MalE deficiency of *E. coli* PD28 cells (see 2.7.3 on page 43 and figures 7.1 - 7.4 on page 104 - 106). Cluster C25 was removed because of insufficient membrane integration (figure 7.1, page 104). The relative growth of PD28 cells varied in a range of 50 to 200% of that of GpA tables 7.1 - 7.5 on page 109 - 114). To asses a possible correlation between relative ToxR activity and relative PD28 growth, all measured values of average β-Gal were plotted against the respective average PD28 growth in figure 3.10 on page 70.

The efficiency of membrane integration (as defined by the slope of the PD28 growth curve) of the protein that show successful membrane integration as defined here (slope $>$ 50% of GpA) does not correlate with TMD affinity. Thus, normalizing β-Gal activities for varying efficiencies of membrane integration would not improve the data.

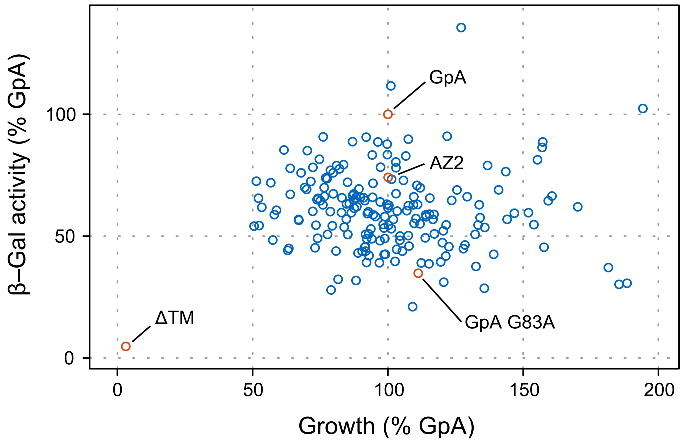

Figure 3.10: Correlating TMD-TMD interaction to membrane insertion. The measured median β-Gal activities are plotted against the respective PD28 growth kinetics reflecting membrane insertion. Data is taken from tables 7.1 - 7.5 on page 109 - 114. Note that TMD affinity is not dependent on the different levels of membrane integration observed with the ToxR/TMD/MalE hybrid proteins used in this work. The red circles represent the reference TMDs: GpA, the high-affinity TMD from human glycophorin A (GpA) forming strong dimers; AZ2, a membrane spanning leucine zipper with medium self-interaction; ΔTM, lacks a TMD and therefore cannot self-interact and cannot integrate into the membrane; G83A, a mutant of GpA which indicates the level of unspecific interaction.

4 Discussion

This work aims to improve the general understanding of the code underlying the interaction of TMDs. It intends to group transmembrane helices from bitopic proteins, to identify representative TMD sequences, and to characterize the most common helix-helix interfaces within the human bitopic proteome. The characterization of TMDs in terms of functional diversity, orientation-dependence, and the impact of mutations, aims to reveal general features related to self-interaction. Additionally, the clusters of TMDs from HLA class II antigens are presented as a new example for the possible role of interaction patterns in homotypic and heterotypic interaction.

The major conclusions are drawn from results obtained from bioinformatic procedures and laboratory measurements in the following sections.

4.1 The similarity threshold for TMD clustering

Structural homology between two soluble proteins is assumed for sequence identity in range of 25-35% [146]. For membrane spanning helices, such a range is most likely shifted towards higher values for some reasons: First, the alignments of TMD sequences are shorter than those of complete protein sequences. Second, TMDs are structurally more similar than soluble domains since they share an α-helical secondary structure. Third, TMDs have to adopt to the lipid bilayer and therefore contain on average 50% of hydrophobic residues (L, I, and V) which leads to coincidental similarity between TMDs. To avoid clusters which originate from such random similarity, a statistical approach for identifying a meaningful similarity threshold was required.

By comparison of natural and randomized TMDs a statistical similarity threshold was found at ssr = 55%, above which pairwise TMD alignments are considered non-random at a 95% level of confidence. Considering the reduced amino acid diversity within TMDs, this threshold of 55% seems reasonable for clustering structural homologue TMDs. However, the short length of TMD sequences raised the question whether the informational content of TMDs is sufficient to group proteins. Most similarities of TMD sequences were expected originate from late gene duplication events without subsequent diversification in function. In such cases, the TMD-TMD similarity may merely represent a functionally and structurally irrelevant echo of evolution. Alternatively, it could be or functional significance. One possibility for functional significance could be TMD-TMD interaction. It could also relate to protein/lipid interaction or backbone dynamics.

4.2 TMD-based clustering of human bitopic membrane proteins

To cluster the bitopic membrane proteins of the human proteome based on TMD sequence similarity, a simple algorithm was applied for two reasons: First, the amount of pairwise alignment scores handled at a time is limited by computational resources, i.e. time and memory. Second, complex algorithms like machine learning approaches require a minimum amount of informational content within the distance measurement between compared sequences. TMDs cannot provide high informational content due to their short length and low amino acid diversity. For example, p-values between complete protein sequences usually span hundreds of orders of magnitude, whereas they only range from $\sim$0.001 to 1.0 for TMDs of the human bitopic proteome. Hierarchical clustering procedures were also ineligible for TMDs, since they often merge clusters which are connected via few single similarities. The resulting clusters would be very large, low in number, and adopt a chain-like structure (figure 3.3, page 58, interconnection of different clusters by TMD similarity). By using the matrix-based approach (3.1.3, page 49), TMDs were grouped into clusters of closely related sequences and a list of TMDs each representing a cluster was obtained. Representative sequences of clusters can implicitly be assumed to represent their homologs in terms of potential TMD dimer structures.

The clustering of TMD sequences based on the similarity threshold of 55% groups $\sim$40% of the human bitopic membrane proteins. Since this coverage only increased from 40.5% to 51.9% after including proteins with a complete sequence homology of $\geq$ 25%, the TMD-based clustering explores the bitopic protein sequence space to a large proportion. Compared to the former analysis performed by Almén *et al.* [6] which assigns 59% of all membrane proteins into functional families based on their complete sequence homology, this percentage is surprisingly high. However, one has to keep in mind that some alignments may be generated by misaligning TMDs of low amino acid diversity. The alignments of cluster C6, C13, and C23 contain very high amounts of the hydrophobic residues L, I, and V, and could be examples for randomly clustered TMDs.

Clustered membrane proteins rarely share common extramembranous domains without similarity of their TMDs (3.2, page 55, $\mathrm{ssr_{TMD}/ssr_{complete}} < 1$). Therefore, the function of bitopic transmembrane proteins is commonly reflected by similarity of the TMD and not only by similarity of the extramembranous domain. Although TMDs are on average more similar than complete protein sequences [52], a large proportion of clustered proteins contains TMDs that are vastly more similar than the complete sequences ($\mathrm{ssr_{TMD}/ssr_{complete}} > 2.5$). In such cases, duplication followed by divergent evolution may be rather unlikely and a functional advantage or strong structural homology of such TMDs is suggested. The majority of the top clusters are functionally rather homogeneous as they contain mainly paralogs with similar annotation. This confirms that the similarity threshold of 55% is meaningful for functional categorization and again highlights the significance of TMDs for the function of membrane proteins. However, 90/265 pairwise TMD alignments that generated the top clusters also suggest relationships between TMDs that belong to proteins being apparently unrelated in function. Part of these alignments could arise from random similarity, in particular between TMDs containing reduced diversity of amino acids. In other cases, TMD-based clustering of functionally unrelated proteins might reflect convergent evolution of TMDs towards structures with similar properties. This is supported by the finding that the members of functionally heterogeneous clusters contain similar TMDs, yet exhibit a much greater diversity in terms of their complete protein sequences, and are enriched in SmxxxSm motifs. Also, members of functionally heterogeneous clusters tend to have fewer paralogs as indicated by the existence of fewer homologs at the level of complete sequence similarity compared to clusters of mostly functional homogeneous proteins. This finding suggests that convergent evolution of TMDs of functionally unrelated proteins may require less duplication events.

4.3 Self-interaction of clustered TMDs

The objective of this work was the systematic assessment of self-interaction of TMDs from the human bitopic proteome which were clustered by sequence homology. Therefore, the efficiency of TMD-TMD self-interaction was examined under comparable condi-

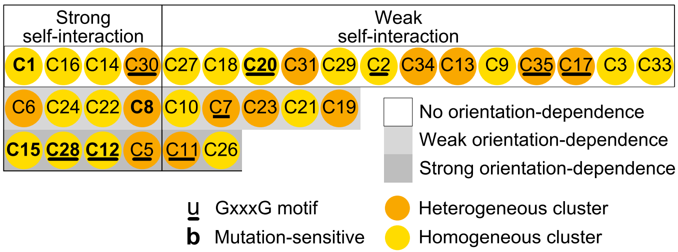

Figure 4.1: A graphical overview showing how the relative affinity of the most representative TMDs connects to their orientation-dependence, the maximal impact of point mutations, the presence of conserved GxxxG motifs, and the functional homogeneity of the respective top clusters. Clusters are divided into strong and weak self-interaction (left and right) defined by relative ToxR activities above or below AZ2, respectively. TMDs interacting in a orientation dependent manner are shaded in gray. Some representative sequences contain GxxxG motifs (underlined). TMDs sensitive for the mutation of conserved residues (bold) show decreased relative affinity by at least 30%.

tions, e.g. protein expression, lipid composition of the host membrane, and identity and density of other membrane proteins. By investigating the representative TMDs of each top cluster, a broad range of relative affinities was found (figure 7.2, page 105). A significant fraction (12/32) of the representative TMDs exhibits high relative affinity within the range of GpA. The interaction of high-affinity TMDs tends to exhibit orientation-dependence, which indicates preferential helix-helix interfaces, and mutation-sensitivity, which signifies well-packed interfaces (figure 4.1). Orientation-dependent self-interaction often includes specific amino acid motifs like GxxxG. In addition, such motifs contain less apolar residues like leucine and valine. Those interfaces may require smaller contact areas and thus may be more specific and less promiscuous. The specific motifs were often found to be mutation-sensitive which may indicate that the exchange of a single amino acid has large impact on the TMD-TMD self-interaction. This finding is supported by former analyses of randomized TMD libraries where high-affinity TMDs contain specific interaction pattern of few conserved amino acids [74, 91, 121, 128]. While 3/6 orientation-dependent TMDs were mutation-sensitive, only 3/26 weakly or not orientation-dependent TMDs were mutation-sensitive. Thereby, mutation- and orientation-independent high-affinity TMDs are likely to interact via non-specific mechanisms. Taken together, the co-occurrence of high relative affinity, orientation-dependence and mutation-sensitivity indicates a strong trend for specific and efficient self-interaction. In some cases, mutating various SmxxxSm motifs strongly reduced affinity, while no significant effect was seen in other cases. Thus, the presence of GxxxG or related motifs does not predict self-interaction.

Table 4.1: Comparison of measured self-interaction and interaction previously described in the literature.

Cluster	Most prevalent functional annotation[a]	ToxR median[b]	ORD[c]	MIM[d]	Known interaction[e]	Relevant HO_{TMD}[f]
C1	Protocadherin	$\underline{85.1}$	0.14	$\underline{0.63}$	$HO_{TMD/Ex}$ [166]	Yes [166]
C5	Integrin α	$\underline{81.6}$	$\underline{0.41}$	-	HO_{TMD}	Discussed
C16	Integrin α	$\underline{81.5}$	0.09	0.09	[130, 167, 168, 169, 170],	[107, 169,
C7	Integrin β	67.3	0.36	-	HE_{TMD} [130, 170, 171, 172]	173, 174]
C15	Sialic-acid-b. Ig-like lectin	$\underline{135.6}$	$\underline{0.71}$	$\underline{0.79}$	HO_{Ex} [175]	Unknown
C28	Armadillo-repeat containing	$\underline{102.4}$	$\underline{0.71}$	$\underline{0.44}$	Unknown	Unknown
C2	HLA class I α-chain	67.2	0.19	0.14	HE_{Ex} [176]	Unknown
C12	HLA class II α-chain	$\underline{93.9}$	$\underline{0.70}$	$\underline{0.37}$	HE_{Ex} [176]	Unknown
C20	HLA class II β-chain	69.0	0.13	$\underline{0.34}$		
C8	Contactin	$\underline{77.1}$	0.27	$\underline{0.60}$	HE_{Ex} [177, 178]	Unknown
C14	Leukocyte Ig-like receptor	$\underline{78.2}$	0.16	0.16	HO_{Ex} [179]	Unknown
C10	UDP-guanosyltransferase	71.9	0.33	-	HO_{Ex} [180]	Unknown
C3	Cadherin	45.2	0.14	-	$HO_{TMD/Ex}$ [181, 182]	Yes [181]
C13	Syntaxin	56.9	0.11	-	HO_{TMD} [183], HE_{TMD} [184]	Yes [185]
C26	VAMP/ Synaptobrevin	66.5	$\underline{0.54}$	0.11	HO_{TMD} [186, 187, 188], HE_{TMD} [183, 184, 189]	Yes [190, 191]

[a] Most prevalent functional annotation of cluster members as annotated in UniProtKB.
[b] The median (>50%) of β-Gal activity measured in % of GpA for the representative TMD in optimal orientation. In case of strong self-interaction (as defined in 3.3.2 on page 62), the value is underlined.
[c] The value for the orientation-dependence (ORD, 2.3.3.1, page 27) between 0 and 1 for the four orientations. Small values indicate a low dependence of β-Gal activity on TMD orientation. Values signifying strong orientation-dependence (as defined in 3.3.1 on page 59) are underlined.
[d] Maximal impact of point mutations (MIM, 2.3.3.2, page 28) calculated from β-Gal activities of wild-type and mutated form of the representative TMD. In case the TMD was denoted as mutation-sensitive (as defined in 3.3.5 on page 66), the value is underlined.
[e] Homotypic (HO) and/or heterotypic (HE) interaction mediated by either the TMD or extramembranous domains (Ex) as described in the literature.
[f] Functional relevance of homotypic interaction mediated by the TMD as described in the literature.

Previously, TMD-TMD interactions have been demonstrated for members of some clusters with high-affinity representative TMDs (table 4.1). For example, the cluster members of C1 (protocadherin) as well as C5 and C16 (integrin α) have been described to exhibit sequence specific homotypic TMD-TMD interactions [166, 167, 168]. In addition to that, the results indicate previously unknown efficient and specific self-interaction of other bitopic protein TMDs, including those of sialic-acid-binding Ig-like lectin (C15), armadillo-repeat containing protein (C28), and HLA class II α (C12). Western-blot analyses of sialic-acid-binding Ig-like lectin revealed a disulphide-linked dimer structure [175]. TMD-TMD interaction might facilitate the formation of covalent dimers and specific disulphide bridges. For armadillo-repeat containing proteins, an oligomerization is not yet known. In the case of HLA class II, an α-chain interacts with an β-chain to constitute a heterodimer which is also referred as MHC class II molecule (see 4.4 on page 77 and [176]). Therefore, the homodimerization of HLA-chains seems to be rather counter-intuitive but could maybe play a role in arranging the premature forms within the endoplasmic reticulum to prevent promiscuous heterotypic interaction. In some cases, self-interaction may be a by-product of functionally relevant heterotypic interaction. Besides HLA class II α- (C12) and β-chains (C20), the TMDs of integrin αIIb (C5 and C16) and β3 (C7) form a transmembrane complex [171]. Integrins also have been reported to perform sequence-specific homotypic interactions [167, 168, 169] but their functional relevance is unclear [173].

The representative bitopic TMDs of some other clusters were found to interact in a less specific way (weak orientation-dependence or weak impact of mutations, table 4.1, page 75) but still exhibit medium affinity. The corresponding proteins either interact heterotypically, like contactin (C8) [177, 178], or homotypically mediated by their extramembranous domains, i.e. leukocyte Ig-like receptor (C14) [179] and UDP-guanosyltransferase (C10) [180]. According to literature, contactins laterally interact with other cell surface proteins also suggesting the medium self-interaction as a by-product of functionally relevant heterotypic interaction. The crystal structure of leukocyte Ig-like receptor and the structural homology model of UDP-guanosyltransferase demonstrate the formation of homodimers but do not include the TMDs [179, 180]. Maybe such proteins prefer neither TMD-TMD nor TMD-lipid interaction which would be reflected by only medium ToxR reporter activity.

The relative affinity exhibited by several representative TMDs is rather low, yet still above that of the non-specific signal of GpA G83A (table 4.1, page 75). However, the TMDs in some of these cases are known for sequence-specific homotypic interactions, i.e. cadherin (C3) [181, 182], integrin β (C7) [168, 169, 170], syntaxin (C13) [183, 185], and synaptobrevin (C26) [187, 188, 190, 192]. These discrepancies may be explained by the differences between the tested representative TMDs and the homologs described in literature. In detail, the TMD of E-cadherin (cadherin-1) is described to strongly self-interact mediated by a leucine zipper side chain packing that supports lateral clustering which is important for cell-cell adhesion [181]. Since the representative TMD of C3 is cadherin-7, the slight differences between the TMD sequences of cadherin-1 from the literature and cadherin-7 may cause large differences in β-Gal activities and thus indicate low affinity for cadherin-7. Similarly, the high-affinity TMD of integrin β3 [168] is not a member of cluster C7. However, the TMD of integrin β1 is a member of C7 and exhibits medium affinity comparable to the literature [170]. For syntaxin, the described TMDs [183, 185] are also no members of cluster C13 which explains their significantly different affinity from the cluster's most representative TMD. As a member of cluster C26, synaptobrevin II exhibits medium affinity comparable to the results from the literature [183]. Synaptobrevin II also interacts with preferential helix-helix interfaces [190] which is confirmed by high orientation-dependence. The GxxxG-like motif of synaptobrevin II does not assist interaction [184] and thus explains the low maximal impact of mutations. It should be also borne in mind, that low affinity detected under the standardized conditions used in this work does not exclude TMD-based self-interaction of proteins expressed at higher concentrations in a eukaryotic cell. In addition, the lipid composition of the host membrane may affect affinity without compromising specificity of an interaction. Furthermore, orientation-dependence and mutation-sensitivity are technically more difficult to establish for those TMDs where self-interaction is lower since mutations and unfavored orientations exert weaker effects.

The TMDs of HLA class II α exhibited high affinity, strong orientation-dependence, and strong impact of point mutations. Since α and β-chain were found in clusters of TMDs, the possible homo- and heterotypic interaction is discussed in the following.

4.4 TMD-TMD interaction of HLA class II

As mentioned before, the heterodimer of HLA class II α- and β-chain (C12 and C20, table 3.1, page 50) is referred to as MHC class II molecule. The function of MHC class II molecules is to present peptides generated in the intracellular vesicles of B cells, macrophages, and other antigen-presenting cells to CD4 T cells [176]. To enable MHC molecules to perform their essential function of signaling intracellular infection, it is important that the complex is stable allowing long-term display of antigens. Despite the heterodimerization of HLA class II TMDs is not known, such interactions may be favorable for the stability of MHC and MHC/antigen complexes. Although heterotypic TMD-TMD dimerization was not tested here, the residues being critical for homotypic interaction might also support heterotypic interaction as described below.

The results of the ToxR reporter assay indicate specific and efficient self-interaction for the representative TMD of the α-chain cluster (C12) of HLA class II. This interaction may be induced by the two conserved GxxxG motifs which are located on the same side of the TMD (figure 4.2, page 79, left section). The replacement of either G_{10}, G_{13}, or G_{17} with isoleucine resulted in decreases of relative affinity and highlights the importance of these for self-interaction. Such glycines may reduce the distance between the helix backbones and thus facilitate hydrogen bond formation between C_α-hydrogens and the backbone of the partner helix [93]. In addition, the β-branched residue of I_{14} could cooperate with the $G_{13}xxxG_{17}$ motif to create a flat helix surface resulting in a well-packed interface which leads to more favorable van der Waals interactions [193]. Surprisingly, the simultaneous mutation of both glycines of the $G_{13}xxxG_{17}$ motif restored the affinity back to wild-type level. A possible explanation could be a switch of the interaction interface from $G_{13}xxxG_{17}$ to the N-terminal G_6xxxG_{10} motif as in the case of ErbB receptor tyrosine kinase [103, 104]. However, this would not explain the reduced affinity of the single G_{17} mutant, in which case the interaction should also switch to the N-terminal GxxxG motif.

The homotypic interaction of the representative TMD of HLA class II β was found to exhibit medium affinity with no orientation-dependence. This might signify self-interaction of low strength and specificity. Yet, the mutation of the conserved GxxxG-like motif and F_{10} significantly decreased affinity. The location of such residues (A_9, F_{10}, and G_{13}) suggest a similar to the α-chain TMD albeit less efficient mechanism of interaction (figure 4.2, page 79, right section). The AxxxG motif might reduce the distance between both interacting helices and induce well-packed interfaces similar to GxxxG motifs. Since the mutation of F_{16} did no reduce the β-Gal activity, an aromatic interaction between F_{10} and F_{16} of both partner helices is unlikely [73, 75]. However, F_{10} could enlarge the interface surface and thus increase van der Waals interaction. The contribution of other residues to the homotypic interaction of HLA class II β TMD was not tested.

The putative TMD-TMD interfaces of HLA class II α- and β-chain TMDs may adopt similar structures and thus may be also involved in heterotypic interaction using a related mechanism to that of homotypic interaction. Figure 4.2 on page 79 depicts a potential model of heterotypic TMD interaction between the HLA class II α- and β-chains. By placing the GxxxG-like motifs facing towards their partner helices, the model of heterotypic interaction suggests close packing of interfaces as in the cases of homotypic interactions. This is supported by the almost identical positioning of $G_{13,\alpha}$ and $G_{13,\beta}$ as

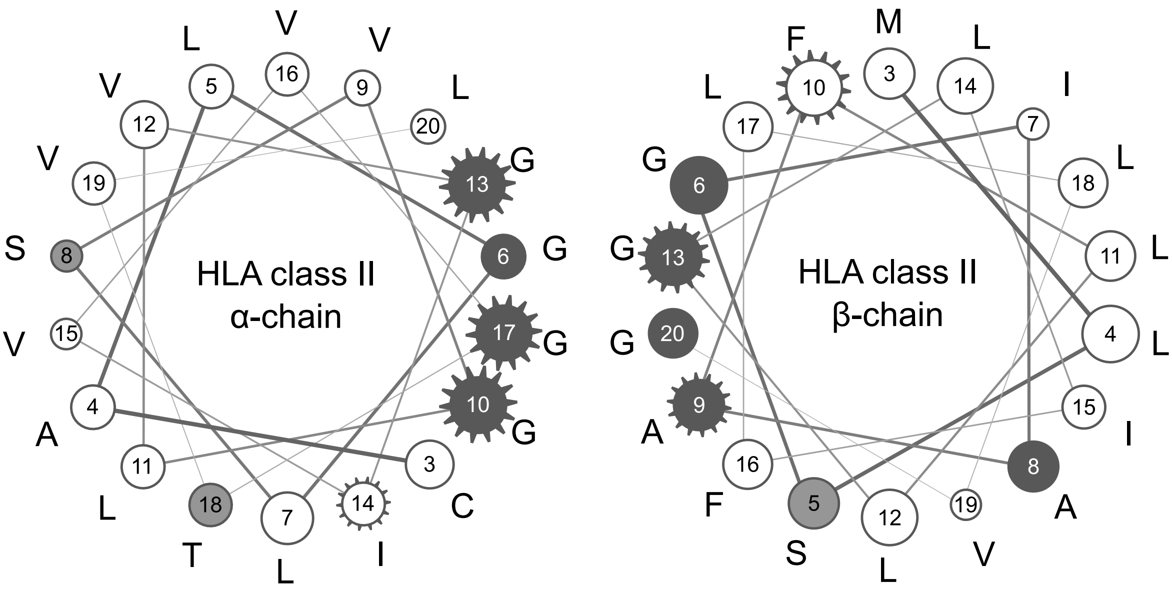

Figure 4.2: The helical wheel representations of the representative TMDs of HLA class II α- and β-chain. The membrane spanning helices contain many hydrophobic residues (white), few polar residues (gray), and some glycines and alanines (dark). The conservation of each residue is depicted as the circle size. Toothed circles indicate residues whose mutations were found to decrease self-interaction. All glycines of the HLA class II α TMD (C12) are located on one face of the helix in two GxxxG motifs. The β-chain (C20) contains even more glycines than the α-chain and includes one AxxxG motif. By facing the GxxxG-like motifs towards their partner helix, a possible heterotypic TMD interaction is illustrated. $G_{13,\alpha}$ and $G_{13,\beta}$, $G_{10,\alpha}$ and $A_{9,\beta}$, and $G_{17,\alpha}$ and $G_{20,\beta}$ could reduce the distance between both helices and thus facilitate backbone hydrogen bond formation. $I_{14,\alpha}$ and $F_{10,\beta}$ could extend the surface area of the interfaces and therefore increase van der Waals interaction.

well as $G_{10,\alpha}$ and $A_{9,\beta}$ within both TMDs. Due to the conservation of $G_{20,\beta}$ within all members of cluster C20, $G_{17,\alpha}$ and $G_{20,\beta}$ could also play a role in heterotypic interaction, yet $G_{20,\beta}$ was not found mutation-sensitive for homotypic interaction. Furthermore, $I_{14,\alpha}$ could take a similar role than $F_{10,\beta}$ which would be the increase of the interface surface. The heterodimerization of α- and β-chain TMDs might assist MHC complex stability.

4.5 Self-interaction of the human bitopic membrane proteome

The self-interaction of homologous TMDs was mostly found to be conserved in clusters (3.3.4, page 64) yet small differences can have a huge impact. If self-interaction is functionally important, the similar affinity of a cluster's members suggests that the corresponding proteins are expressed at similar concentrations and experience comparable influences from other factors like lipids, crowding effects, etc. In contrast, for those members of clusters whose affinity deviate, evolution could have shaped their interfaces in a similar fashion, but other influencing factors are less comparable in the host membrane environment than in measurements.

Does the TMD-based clustering also select for proteins with self-interacting TMDs from the pool of bitopic protein TMDs? Presently, this question cannot be answered with certainty since data for a representative set of non-clustered TMDs are lacking. However, Ried *et al.* recently found that almost half of the TMDs from human bitopic proteins exhibit non-random unilateral sequence conservation in alignments with their orthologs [53]. Since unilateral conservation correlates with the efficiency of self-interaction, this suggests that a major fraction of human bitopic protein TMDs can self-interact. Therefore, the broad distribution of affinities exhibited by top cluster TMDs may reflect self-interaction of human bitopic membrane proteins independent of clustering.

Generally, GxxxG motifs are enriched in clustered human bitopic TMDs but not necessarily involved in homotypic TMD-TMD interaction as discussed before. Exclusive homotypic or heterotypic interaction may be more easily achieved within non-clustered TMDs. Since TMDs with exclusive interaction are not abundant in the membrane proteome they do not share common interfaces.

4.6 The meaning of functionally heterogeneous clusters

Since domain recombination is not common in membrane proteins [43], convergent evolution by mutation may have generated the homology of the TMDs from functionally heterogeneous clusters. These TMDs are enriched for putative oligomerization motifs, i.e. GxxxG and other SmxxxSm. Also, GxxxG motifs were found to be more often conserved in at least 60% of the members of functionally heterogeneous clusters (6/13, Figure 4.1, page 74) than in functional homogeneous clusters (4/20). Evolution of these short motifs requires fewer mutations than that of more complex helix-helix interfaces and they could therefore develop rather rapidly by convergent evolution. On the other hand, TMDs of functionally heterogeneous clusters do not exhibit a generally higher relative affinity, orientation-dependence, or mutation sensitivity compared to TMDs from mainly homogeneous clusters (Figure 4.1, page 74). Therefore, clustering of TMDs from functionally diverse proteins is no stronger predictor for homotypic interaction than clustering of TMDs from functionally related proteins. In addition to TMD-TMD interaction, sequence conservation could reflect conserved heterotypic interaction, interaction with lipids, and/or co-factors. To avoid promiscuous heterotypic interactions between paralogs, individual protein subtypes could be located within membranes of different organelles or within various cell types.

5 Conclusion

This work aimed to systematically assess the self-interaction of the human bitopic TMDs clustered by sequence homology. It focused on revealing indicators for specific and efficient TMD self-interaction and addressed questions about the significance of TMD-based clusters, the distribution of self-interaction in the human bitopic membrane proteome, and the evolution of TMDs.

For the first time, the human bitopic membrane proteins were clustered based on the similarity of their TMD sequences only. Surprisingly, this clustering explored the bitopic protein sequence space to a large proportion. The majority of the top clusters were found to be functionally rather homogeneous as they contain mainly paralogs with similar functional annotation. However, many pairwise TMD alignments that generated the top clusters also suggest relationships between TMDs that belong to proteins being apparently unrelated in function. The TMD-based clustering of functionally unrelated proteins might reflect convergent evolution of TMDs towards structures with similar properties. This was supported by the finding that the members of functionally heterogeneous clusters contained similar TMDs, yet exhibited a much greater diversity in terms of their complete protein sequences similarity.

The efficiency of TMD-TMD self-interaction was examined with the ToxR transcription activator system under comparable conditions. In collaboration with the laboratory of Prof. Arkin in Jerusalem, the representative TMD for each major cluster was tested. This resulted in a broad range of relative affinities with a significant fraction of the representative TMDs exhibiting high relative affinity within the range of GpA. Such high-affinity TMDs tended to exhibit orientation-dependence and mutation-sensitivity. The co-occurrence of these criteria were assumed to indicated specific and efficient self-interaction. In some cases, mutating various GxxxG-like motifs strongly reduced affinity, while no significant effect was seen in other cases. As expected, the presence of GxxxG or related motifs did not predict self-interaction. The results of this work also reveal previously unknown efficient self-interaction of some bitopic protein TMDs. Such self-interaction could be a by-product of functionally relevant heterotypic interaction as it was suggested in the case of HLA class II complexes. To probe for such a possible heterogeneous interaction, the heterotypic ToxR assay [128, 140] could be used as a follow-up. Since the specific interaction mechanism of most top cluster TMDs is not clear, a characterization of their TMD-TMD interfaces could be advanced by predicted models for TMD-TMD assemblies using molecular modeling and molecular dynamics simulations. The suggested 3D models would allow for the preselection of residues for further mutagenesis. To improve the understanding of sequence motif contribution towards stabilization of the dimer structures, site-specific FTIR spectroscopy could be applied to determine the backbone structures and orientation of transmembrane helices in the lipid bilayer.

The question whether TMD-based clustering selects for proteins with self-interacting TMDs could not be answered yet. A comparison of relative affinities between clustered and non-clustered TMDs could reveal general differences in self-interaction. This would require further ToxR tests of non-clustered TMDs. However, TMDs of top clusters exhibited a broad distribution of affinities which might reflect self-interaction of human bitopic membrane proteins independent of clustering. As functionally homogeneous clusters mainly contained paralogs, such TMDs may have been evolved divergently by gene duplication and diversification. In contrast to functionally homogeneous clusters, convergent evolution may have generated the homology of the TMDs of functionally heterogeneous clusters. Such TMDs were found to be enriched for putative oligomerization motifs. Evolution of these short motifs would require fewer mutations than that of more complex helix-helix interfaces and thus they could have been developed rather rapidly in convergent evolution. However, TMDs from functionally heterogeneous clusters did not exhibit a generally higher relative affinity, orientation-dependence, or mutation sensitivity compared to TMDs from mainly homogeneous clusters. Therefore, clustering of TMDs from functionally diverse proteins was found to be no stronger predictor for homotypic interaction than clustering of TMDs from functionally related proteins.

An extension of this project could be a phylogenetic analysis to determine the evolutionary distances of gene duplication within functionally homogeneous and heterogeneous clusters. Such an investigation would require phylogenetic gene tree reconciliation of genes from different species including human. The results might reveal genes which were duplicated in the early or late stages of evolution and would allow for assignment of loss or retention of function and interaction of the corresponding TMDs. Further, one could study the potential evolution of TMDs from homogeneous towards heterogeneous interaction.

6 Bibliography

[1] S. H. White and W. C. Wimley. Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct*, **28**:319–365, 1999.

[2] S. J. Singer and G. L. Nicolson. The fluid mosaic model of the structure of cell membranes. *Science*, **175**(4023):720–731, 1972.

[3] K. Simons and E. Ikonen. Functional rafts in cell membranes. *Nature*, **387**(6633):569–572, 1997.

[4] E. Wallin and G. von Heijne. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci*, **7**(4):1029–1038, 1998.

[5] T. J. Stevens and I. T. Arkin. The effect of nucleotide bias upon the composition and prediction of transmembrane helices. *Protein Sci*, **9**(3):505–511, 2000.

[6] M. S. Almén, K. J. V. Nordström, R. Fredriksson, and H. B. Schiöth. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol*, **7**:50, 2009.

[7] C. M. Armstrong. Ionic pores, gates, and gating currents. *Q Rev Biophys*, **7**(2):179–210, 1974.

[8] E. Gouaux and R. Mackinnon. Principles of selective ion transport in channels and pumps. *Science*, **310**(5753):1461–1465, 2005.

[9] J. Schlessinger. Cell signaling by receptor tyrosine kinases. *Cell*, **103**(2):211–225, 2000.

[10] M. A. Lemmon and J. Schlessinger. Cell signaling by receptor tyrosine kinases. *Cell*, **141**(7):1117–1134, 2010.

[11] M. Takeichi. Cadherin cell adhesion receptors as a morphogenetic regulator. *Science*, **251**(5000):1451–1455, 1991.

[12] F. G. Giancotti and E. Ruoslahti. Integrin signaling. *Science*, **285**(5430):1028–1032, 1999.

[13] J. P. Dekker and E. J. Boekema. Supramolecular organization of thylakoid membrane proteins in green plants. *Biochim Biophys Acta*, **1706**(1-2):12–39, 2005.

[14] J. Vonck and E. Schäfer. Supramolecular organization of protein complexes in the mitochondrial inner membrane. *Biochim Biophys Acta*, **1793**(1):117–124, 2009.

[15] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal. Drug-target network. *Nat Biotechnol*, **25**(10):1119–1126, 2007.

[16] R. Worch, C. Bökel, S. Höfinger, P. Schwille, and T. Weidemann. Focus on composition and interaction potential of single-pass transmembrane domains. *Proteomics*, **10**(23):4196–4208, 2010.

[17] D. M. Engelman, T. A. Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, **15**:321–353, 1986.

[18] J. U. Bowie. Helix packing in membrane proteins. *J Mol Biol*, **272**(5):780–789, 1997.

[19] I. T. Arkin and A. T. Brunger. Statistical analysis of predicted transmembrane alpha-helices. *Biochim Biophys Acta*, **1429**(1):113–128, 1998.

[20] A. Senes, M. Gerstein, and D. M. Engelman. 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**(3):921–936, 2000.

[21] M. B. Ulmschneider and M. S. Sansom. Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta*, **1512**(1):1–14, 2001.

[22] W. C. Wimley and S. H. White. Designing transmembrane alpha-helices that insert spontaneously. *Biochemistry*, **39**(15):4432–4442, 2000.

[23] S. H. White and G. von Heijne. The machinery of membrane protein assembly. *Curr Opin Struct Biol*, **14**(4):397–404, 2004.

[24] J. Luirink, G. von Heijne, E. Houben, and J.-W. de Gier. Biogenesis of inner membrane proteins in Escherichia coli. *Annu Rev Microbiol*, **59**:329–355, 2005.

[25] A. J. M. Driessen and N. Nouwen. Protein translocation across the bacterial cytoplasmic membrane. *Annu Rev Biochem*, **77**:643–667, 2008.

[26] M. Müller, H. G. Koch, K. Beck, and U. Schäfer. Protein traffic in bacteria: multiple routes from the ribosome to and across the membrane. *Prog Nucleic Acid Res Mol Biol*, **66**:107–157, 2001.

[27] J. C. Samuelson, M. Chen, F. Jiang, I. Möller, M. Wiedmann, A. Kuhn, G. J. Phillips, and R. E. Dalbey. YidC mediates membrane protein insertion in bacteria. *Nature*, **406**(6796):637–641, 2000.

[28] T. A. Rapoport. Transport of proteins across the endoplasmic reticulum membrane. *Science*, **258**(5084):931–936, 1992.

[29] K. Wild, M. Halic, I. Sinning, and R. Beckmann. SRP meets the ribosome. *Nat Struct Mol Biol*, **11**(11):1049–1053, 2004.

[30] W. Wickner and R. Schekman. Protein translocation across biological membranes. *Science*, **310**(5753):1452–1456, 2005.

[31] G. Heijne. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J*, **5**(11):3021–3027, 1986.

[32] G. von Heijne. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol*, **225**(2):487–494, 1992.

[33] G. von Heijne and Y. Gavel. Topogenic signals in integral membrane proteins. *Eur J Biochem*, **174**(4):671–678, 1988.

[34] C. Lundin, H. Kim, I. Nilsson, S. H. White, and G. von Heijne. Molecular code for protein insertion in the endoplasmic reticulum membrane is similar for N(in)-C(out) and N(out)-C(in) transmembrane helices. *Proc Natl Acad Sci U S A*, **105**(41):15702–15707, 2008.

[35] S. U. Heinrich, W. Mothes, J. Brunner, and T. A. Rapoport. The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell*, **102**(2):233–244, 2000.

[36] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S. H. White, and G. von Heijne. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, **433**(7024):377–381, 2005.

[37] T. Hessa, N. M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S. H. White, and G. von Heijne. Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature*, **450**(7172):1026–1030, 2007.

[38] E. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, **6**:175–182, 1998.

[39] L. Käll, A. Krogh, and E. L. L. Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, **338**(5):1027–1036, 2004.

[40] H. Nielsen and A. Krogh. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol*, **6**:122–130, 1998.

[41] E. V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, **39**:309–338, 2005.

[42] V. Alva, M. Remmert, A. Biegert, A. N. Lupas, and J. Söding. A galaxy of folds. *Protein Sci*, **19**(1):124–130, 2010.

[43] Y. Liu, M. Gerstein, and D. M. Engelman. Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. *Proc Natl Acad Sci U S A*, **101**(10):3495–3497, 2004.

[44] I. T. Arkin, P. D. Adams, K. R. MacKenzie, M. A. Lemmon, A. T. Brünger, and D. M. Engelman. Structural organization of the pentameric transmembrane alpha-helices of phospholamban, a cardiac ion channel. *EMBO J*, **13**(20):4757–4764, 1994.

[45] J. L. Popot and D. M. Engelman. Helical membrane protein folding, stability, and evolution. *Annu Rev Biochem*, **69**:881–922, 2000.

[46] A. M. Dixon, B. J. Stanley, E. E. Matthews, J. P. Dawson, and D. M. Engelman. Invariant chain transmembrane domain trimerization: a step in MHC class II assembly. *Biochemistry*, **45**(16):5228–5234, 2006.

[47] S. P. Barwe, S. Kim, S. A. Rajasekaran, J. U. Bowie, and A. K. Rajasekaran. Janus model of the Na,K-ATPase beta-subunit transmembrane domain: distinct faces mediate alpha/beta assembly and beta-beta homo-oligomerization. *J Mol Biol*, **365**(3):706–714, 2007.

[48] F. Cymer, A. Veerappan, and D. Schneider. Transmembrane helix-helix interactions are modulated by the sequence context and by lipid bilayer properties. *Biochim Biophys Acta*, **1818**(4):963–973, 2012.

[49] D. Schneider. Rendezvous in a membrane: close packing, hydrogen bonding, and the formation of transmembrane helix oligomers. *FEBS Lett*, **577**(1-2):5–8, 2004.

[50] A. Fink, N. Sal-Man, D. Gerber, and Y. Shai. Transmembrane domains interactions within the membrane milieu: Principles, advances and challenges. *Biochim Biophys Acta*, **1818**(4):974–983, 2012.

[51] E. Li, W. C. Wimley, and K. Hristova. Transmembrane helix dimerization: beyond the search for sequence motifs. *Biochim Biophys Acta*, **1818**(2):183–193, 2012.

[52] M. Zviling, U. Kochva, and I. T. Arkin. How important are transmembrane helices of bitopic membrane proteins? *Biochim Biophys Acta*, **1768**(3):387–392, 2007.

[53] C. L. Ried, S. Kube, J. Kirrbach, and D. Langosch. Homotypic Interaction and Amino Acid Distribution of Unilaterally Conserved Transmembrane Helices. *J Mol Biol*, 2012.

[54] T. J. Stevens and I. T. Arkin. Substitution rates in alpha-helical transmembrane proteins. *Protein Sci*, **10**(12):2507–2517, 2001.

[55] T. A. Eyre, L. Partridge, and J. M. Thornton. Computational analysis of alpha-helical membrane protein structure: implications for the prediction of 3D structural models. *Protein Eng Des Sel*, **17**(8):613–624, 2004.

[56] Y. Mokrab, T. J. Stevens, and K. Mizuguchi. A structural dissection of amino acid substitutions in helical transmembrane proteins. *Proteins*, **78**(14):2895–2907, 2010.

[57] A. Fuchs, A. J. Martin-Galiano, M. Kalman, S. Fleishman, N. Ben-Tal, and D. Frishman. Co-evolving residues in membrane proteins. *Bioinformatics*, **23**(24):3312–3319, 2007.

[58] H. Hong and J. U. Bowie. Dramatic destabilization of transmembrane helix interactions by features of natural membrane environments. *J Am Chem Soc*, **133**(29):11389–11398, 2011.

[59] E. E. Matthews, M. Zoonens, and D. M. Engelman. Dynamic helix interactions in transmembrane signaling. *Cell*, **127**(3):447–450, 2006.

[60] D. T. Moore, B. W. Berger, and W. F. DeGrado. Protein-protein interactions in the membrane: sequence, structural, and biological motifs. *Structure*, **16**(7):991–1001, 2008.

[61] K. R. MacKenzie, J. H. Prestegard, and D. M. Engelman. A transmembrane helix dimer: structure and implications. *Science*, **276**(5309):131–133, 1997.

[62] M. A. Lemmon, J. M. Flanagan, H. R. Treutlein, J. Zhang, and D. M. Engelman. Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry*, **31**(51):12719–12725, 1992.

[63] D. Langosch, B. Brosig, H. Kolmar, and H. J. Fritz. Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. *J Mol Biol*, **263**(4):525–530, 1996.

[64] D. Langosch and J. Heringa. Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils. *Proteins*, **31**(2):150–159, 1998.

[65] F. X. Zhou, H. J. Merianos, A. T. Brunger, and D. M. Engelman. Polar residues drive association of polyleucine transmembrane helices. *Proc Natl Acad Sci U S A*, **98**(5):2250–2255, 2001.

[66] H. Gratkowski, J. D. Lear, and W. F. DeGrado. Polar side chains drive the association of model transmembrane peptides. *Proc Natl Acad Sci U S A*, **98**(3):880–885, 2001.

[67] J. P. Dawson, J. S. Weinger, and D. M. Engelman. Motifs of serine and threonine can drive association of transmembrane helices. *J Mol Biol*, **316**(3):799–805, 2002.

[68] S. E. Harrington and N. Ben-Tal. Structural determinants of transmembrane helical proteins. *Structure*, **17**(8):1092–1103, 2009.

[69] K. R. Mackenzie. Folding and stability of alpha-helical integral membrane proteins. *Chem Rev*, **106**(5):1931–1977, 2006.

[70] M. Eilers, S. C. Shekar, T. Shieh, S. O. Smith, and P. J. Fleming. Internal packing of helical membrane proteins. *Proc Natl Acad Sci U S A*, **97**(11):5796–5801, 2000.

[71] M. Eilers, A. B. Patel, W. Liu, and S. O. Smith. Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys J*, **82**(5):2720–2736, 2002.

[72] W. C. Wimley, K. Gawrisch, T. P. Creamer, and S. H. White. Direct measurement of salt-bridge solvation energies using a peptide model system: implications for protein stability. *Proc Natl Acad Sci U S A*, **93**(7):2985–2990, 1996.

[73] R. M. Johnson, K. Hecht, and C. M. Deber. Aromatic and cation-pi interactions enhance helix-helix association in a membrane environment. *Biochemistry*, **46**(32):9208–9214, 2007.

[74] A. Ridder, P. Skupjen, S. Unterreitmeier, and D. Langosch. Tryptophan supports interaction of transmembrane helices. *J Mol Biol*, **354**(4):894–902, 2005.

[75] S. Unterreitmeier, A. Fuchs, T. Schäffler, R. G. Heym, D. Frishman, and D. Langosch. Phenylalanine promotes interaction of transmembrane domains via GxxxG motifs. *J Mol Biol*, **374**(3):705–718, 2007.

[76] J. Ren, S. Lew, J. Wang, and E. London. Control of the transmembrane orientation and interhelical interactions within membranes by hydrophobic helix length. *Biochemistry*, **38**(18):5905–5912, 1999.

[77] E. Sparr, W. L. Ash, P. V. Nazarov, D. T. S. Rijkers, M. A. Hemminga, D. P. Tieleman, and J. A. Killian. Self-association of transmembrane alpha-helices in model membranes: importance of helix orientation and role of hydrophobic mismatch. *J Biol Chem*, **280**(47):39324–39331, 2005.

[78] A. Holt and J. A. Killian. Orientation and dynamics of transmembrane peptides: the power of simple models. *Eur Biophys J*, **39**(4):609–621, 2010.

[79] J. A. Killian. Hydrophobic mismatch between proteins and lipids in membranes. *Biochim Biophys Acta*, **1376**(3):401–415, 1998.

[80] A. G. Lee. Lipid-protein interactions in biological membranes: a structural perspective. *Biochim Biophys Acta*, **1612**(1):1–40, 2003.

[81] V. Anbazhagan and D. Schneider. The membrane environment modulates self-association of the human GpA TM domain–implications for membrane protein folding and transmembrane signaling. *Biochim Biophys Acta*, **1798**(10):1899–1907, 2010.

[82] S. Mall, R. Broadbridge, R. P. Sharma, J. M. East, and A. G. Lee. Self-association of model transmembrane alpha-helices is modulated by lipid structure. *Biochemistry*, **40**(41):12379–12386, 2001.

[83] A. K. Chamberlain, Y. Lee, S. Kim, and J. U. Bowie. Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *J Mol Biol*, **339**(2):471–479, 2004.

[84] W. M. Yau, W. C. Wimley, K. Gawrisch, and S. H. White. The preference of tryptophan for membrane interfaces. *Biochemistry*, **37**(42):14713–14718, 1998.

[85] G. van Meer, D. R. Voelker, and G. W. Feigenson. Membrane lipids: where they are and how they behave. *Nat Rev Mol Cell Biol*, **9**(2):112–124, 2008.

[86] J. C. Malinverni and T. J. Silhavy. An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane. *Proc Natl Acad Sci U S A*, **106**(19):8009–8014, 2009.

[87] B. Grasberger, A. P. Minton, C. DeLisi, and H. Metzger. Interaction between proteins localized in membranes. *Proc Natl Acad Sci U S A*, **83**(17):6258–6262, 1986.

[88] D. Langosch and I. T. Arkin. Interaction and conformational dynamics of membrane-spanning protein helices. *Protein Sci*, **18**(7):1343–1358, 2009.

[89] R. F. S. Walters and W. F. DeGrado. Helix-packing motifs in membrane proteins. *Proc Natl Acad Sci U S A*, **103**(37):13658–13663, 2006.

[90] C. B. Anfinsen. Advances in Protein Chemistry. Academic Press, 1988.

[91] W. P. Russ and D. M. Engelman. The GxxxG motif: a framework for transmembrane helix-helix association. *J Mol Biol*, **296**(3):911–919, 2000.

[92] B. J. Bormann, W. J. Knowles, and V. T. Marchesi. Synthetic peptides mimic the assembly of transmembrane glycoproteins. *J Biol Chem*, **264**(7):4033–4037, 1989.

[93] A. Senes, I. Ubarretxena-Belandia, and D. M. Engelman. The Calpha —H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc Natl Acad Sci U S A*, **98**(16):9056–9061, 2001.

[94] E. Arbely and I. T. Arkin. Experimental measurement of the strength of a C alpha-H...O bond in a lipid bilayer. *J Am Chem Soc*, **126**(17):5362–5363, 2004.

[95] S. O. Smith, D. Song, S. Shekar, M. Groesbeek, M. Ziliox, and S. Aimoto. Structure of the transmembrane dimer interface of glycophorin A in membrane bilayers. *Biochemistry*, **40**(22):6553–6558, 2001.

[96] H. R. Treutlein, M. A. Lemmon, D. M. Engelman, and A. T. Brünger. The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. *Biochemistry*, **31**(51):12726–12732, 1992.

[97] K. R. MacKenzie and K. G. Fleming. Association energetics of membrane spanning alpha-helices. *Curr Opin Struct Biol*, **18**(4):412–419, 2008.

[98] A. K. Doura, F. J. Kobus, L. Dubrovsky, E. Hibbard, and K. G. Fleming. Sequence context modulates the stability of a GxxxG-mediated transmembrane helix-helix dimer. *J Mol Biol*, **341**(4):991–998, 2004.

[99] A. Senes, D. E. Engel, and W. F. DeGrado. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr Opin Struct Biol*, **14**(4):465–479, 2004.

[100] D. Schneider and D. M. Engelman. Motifs of two small residues can assist but are not sufficient to mediate transmembrane helix interactions. *J Mol Biol*, **343**(4):799–804, 2004.

[101] B.-H. Luo and T. A. Springer. Integrin structures and conformational signaling. *Curr Opin Cell Biol*, **18**(5):579–586, 2006.

[102] F. Cymer and D. Schneider. Transmembrane helix-helix interactions involved in ErbB receptor signaling. *Cell Adh Migr*, **4**(2):299–312, 2010.

[103] C. Escher, F. Cymer, and D. Schneider. Two GxxxG-like motifs facilitate promiscuous interactions of the human ErbB transmembrane domains. *J Mol Biol*, **389**(1):10–16, 2009.

[104] S. J. Fleishman, J. Schlessinger, and N. Ben-Tal. A putative molecular-activation switch in the transmembrane domain of erbB2. *Proc Natl Acad Sci U S A*, **99**(25):15937–15940, 2002.

[105] E. V. Bocharov, K. S. Mineev, P. E. Volynsky, Y. S. Ermolyuk, E. N. Tkach, A. G. Sobol, V. V. Chupin, M. P. Kirpichnikov, R. G. Efremov, and A. S. Arseniev. Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state. *J Biol Chem*, **283**(11):6950–6956, 2008.

[106] K. S. Mineev, E. V. Bocharov, Y. E. Pustovalova, O. V. Bocharova, V. V. Chupin, and A. S. Arseniev. Spatial structure of the transmembrane domain heterodimer of ErbB1 and ErbB2 receptor tyrosine kinases. *J Mol Biol*, **400**(2):231–243, 2010.

[107] D. Schneider and D. M. Engelman. Involvement of transmembrane domain interactions in signal transduction by alpha/beta integrins. *J Biol Chem*, **279**(11):9840–9846, 2004.

[108] E. V. Bocharov, Y. E. Pustovalova, K. V. Pavlov, P. E. Volynsky, M. V. Goncharuk, Y. S. Ermolyuk, D. V. Karpunin, A. A. Schulga, M. P. Kirpichnikov, R. G. Efremov, I. V. Maslennikov, and A. S. Arseniev. Unique dimeric structure of BNip3 transmembrane domain suggests membrane permeabilization as a cell death trigger. *J Biol Chem*, **282**(22):16256–16266, 2007.

[109] M. E. Call, K. W. Wucherpfennig, and J. J. Chou. The structural basis for intramembrane assembly of an activating immunoreceptor complex. *Nat Immunol*, **11**(11):1023–1029, 2010.

[110] N. Sal-Man, D. Gerber, and Y. Shai. The composition rather than position of polar residues (QxxS) drives aspartate receptor transmembrane domain dimerization in vivo. *Biochemistry*, **43**(8):2309–2313, 2004.

[111] E. K. O'Shea, J. D. Klemm, P. S. Kim, and T. Alber. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science*, **254**(5031):539–544, 1991.

[112] R. Gurezka, R. Laage, B. Brosig, and D. Langosch. A heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments. *J Biol Chem*, **274**(14):9265–9270, 1999.

[113] R. Gurezka and D. Langosch. In vitro selection of membrane-spanning leucine zipper protein-protein interaction motifs using POSSYCCAT. *J Biol Chem*, **276**(49):45580–45587, 2001.

[114] W. Ruan, V. Becker, U. Klingmüller, and D. Langosch. The interface between self-assembling erythropoietin receptor transmembrane segments corresponds to a membrane-spanning leucine zipper. *J Biol Chem*, **279**(5):3273–3279, 2004.

[115] N. A. Noordeen, F. Carafoli, E. Hohenester, M. A. Horton, and B. Leitinger. A transmembrane leucine zipper is required for activation of the dimeric receptor tyrosine kinase DDR1. *J Biol Chem*, **281**(32):22744–22751, 2006.

[116] K. Skorupski and R. K. Taylor. Control of the ToxR virulence regulon in Vibrio cholerae by environmental stimuli. *Mol Microbiol*, **25**(6):1003–1009, 1997.

[117] V. L. Miller and J. J. Mekalanos. Synthesis of cholera toxin is positively regulated at the transcriptional level by toxR. *Proc Natl Acad Sci U S A*, **81**(11):3471–3475, 1984.

[118] V. L. Miller, R. K. Taylor, and J. J. Mekalanos. Cholera toxin transcriptional activator toxR is a transmembrane DNA binding protein. *Cell*, **48**(2):271–279, 1987.

[119] H. Kolmar, C. Frisch, G. Kleemann, K. Götze, F. J. Stevens, and H. J. Fritz. Dimerization of Bence Jones proteins: linking the rate of transcription from an Escherichia coli promoter to the association constant of REIV. *Biol Chem Hoppe Seyler*, **375**(1):61–70, 1994.

[120] H. Kolmar, F. Hennecke, K. Götze, B. Janzer, B. Vogt, F. Mayer, and H. J. Fritz. 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**(16):3895–3904, 1995.

[121] E. Lindner, S. Unterreitmeier, A. N. J. A. Ridder, and D. Langosch. An extended ToxR POSSYCCAT system for positive and negative selection of self-interacting transmembrane domains. *J Microbiol Methods*, **69**(2):298–305, 2007.

[122] J. H. Miller. Experiments in Molecular Genetics. Cold Spring Harbor Laboratory Press, U.S, 1972.

[123] E. Lindner. Identifikation heterotypischer TMD-TMD Interaktionen. Ph.D. thesis, Technische Universität München, 2006.

[124] U. K. Laemmli. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**(5259):680–685, 1970.

[125] W. N. Burnette. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate–polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Anal Biochem*, **112**(2):195–203, 1981.

[126] H. A. Shuman. Active transport of maltose in Escherichia coli K12. Role of the periplasmic maltose-binding protein and evidence for a substrate recognition site in the cytoplasmic membrane. *J Biol Chem*, **257**(10):5455–5461, 1982.

[127] B. Brosig and D. Langosch. The dimerization motif of the glycophorin A transmembrane segment in membranes: importance of glycine residues. *Protein Sci*, **7**(4):1052–1056, 1998.

[128] J. R. Herrmann, J. C. Panitz, S. Unterreitmeier, A. Fuchs, D. Frishman, and D. Langosch. Complex patterns of histidine, hydroxylated amino acids and the GxxxG motif mediate high-affinity transmembrane domain interactions. *J Mol Biol*, **385**(3):912–923, 2009.

[129] W. P. Russ and D. M. Engelman. TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc Natl Acad Sci U S A*, **96**(3):863–868, 1999.

[130] D. Schneider and D. M. Engelman. GALLEX, a measurement of heterologous association of transmembrane helices in a biological membrane. *J Biol Chem*, **278**(5):3105–3111, 2003.

[131] L. Chen, L. Novicky, M. Merzlyakov, T. Hristov, and K. Hristova. Measuring the energetics of membrane protein dimerization in mammalian membranes. *J Am Chem Soc*, **132**(10):3628–3635, 2010.

[132] J. Tang, H. Yin, J. Qiu, M. J. Tucker, W. F. DeGrado, and F. Gai. Using two fluorescent probes to dissect the binding, insertion, and dimerization kinetics of a model membrane peptide. *J Am Chem Soc*, **131**(11):3816–3817, 2009.

[133] E. Psachoulia, D. P. Marshall, and M. S. P. Sansom. Molecular dynamics simulations of the dimerization of transmembrane alpha-helices. *Acc Chem Res*, **43**(3):388–396, 2010.

[134] J. Zhang and T. Lazaridis. Transmembrane helix association affinity can be modulated by flanking and noninterfacial residues. *Biophys J*, **96**(11):4418–4427, 2009.

[135] C.-P. Chng and S.-M. Tan. Leukocyte integrin alpha-L-beta-2 transmembrane association dynamics revealed by coarse-grained molecular dynamics simulations. *Proteins*, **79**(7):2203–2213, 2011.

[136] N. Sal-Man, D. Gerber, I. Bloch, and Y. Shai. Specificity in transmembrane helix-helix interactions mediated by aromatic residues. *J Biol Chem*, **282**(27):19753–19761, 2007.

[137] O. S. Soumana, N. Garnier, and M. Genest. Molecular dynamics simulation approach for the prediction of transmembrane helix-helix heterodimers assembly. *Eur Biophys J*, **36**(8):1071–1082, 2007.

[138] O. S. Soumana, N. Garnier, and M. Genest. Insight into the recognition patterns of the ErbB receptor family transmembrane domains: heterodimerization models through molecular dynamics search. *Eur Biophys J*, **37**(6):851–864, 2008.

[139] D. W. Sammond, D. E. Bosch, G. L. Butterfoss, C. Purbeck, M. Machius, D. P. Siderovski, and B. Kuhlman. Computational design of the sequence and structure of a protein-binding peptide. *J Am Chem Soc*, **133**(12):4190–4192, 2011.

[140] J. R. Herrmann, A. Fuchs, J. C. Panitz, T. Eckert, S. Unterreitmeier, D. Frishman, and D. Langosch. Ionic interactions promote transmembrane helix-helix association depending on sequence context. *J Mol Biol*, **396**(2):452–461, 2010.

[141] C. A. Wilson, J. Kreychman, and M. Gerstein. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, **297**(1):233–249, 2000.

[142] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**(3):443–453, 1970.

[143] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, **147**(1):195–197, 1981.

[144] W. R. Pearson. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*, **183**:63–98, 1990.

[145] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, **215**(3):403–410, 1990.

[146] B. Rost. Twilight zone of protein sequence alignments. *Protein Eng*, **12**(2):85–94, 1999.

[147] R. F. Doolittle. Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, CA, USA, 1986.

[148] Y. Liu, D. M. Engelman, and M. Gerstein. Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol*, **3**(10):research0054, 2002.

[149] A. Oberai, Y. Ihm, S. Kim, and J. U. Bowie. A limited universe of membrane protein families and folds. *Protein Sci*, **15**(7):1723–1734, 2006.

[150] S. Neumann, H. Hartmann, A. J. Martin-Galiano, A. Fuchs, and D. Frishman. Camps 2.0: exploring the sequence and structure space of prokaryotic, eukaryotic, and viral membrane proteins. *Proteins*, **80**(3):839–857, 2012.

[151] A. Fuchs and D. Frishman. Structural comparison and classification of alpha-helical transmembrane domains based on helix interaction patterns. *Proteins*, **78**(12):2587–2599, 2010.

[152] The UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res*, **36**(Database issue):D190–D195, 2008.

[153] P. C. Ng, J. G. Henikoff, and S. Henikoff. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*, **16**(9):760–766, 2000.

[154] S. Henikoff, J. G. Henikoff, and S. Pietrokovski. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**(6):471–479, 1999.

[155] T. Müller, S. Rahmann, and M. Rehmsmeier. Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics*, **17 Suppl 1**:S182–S189, 2001.

[156] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**(22):10915–10919, 1992.

[157] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, **16**(6):276–277, 2000.

[158] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**(21):2947–2948, 2007.

[159] T. Frickey and A. Lupas. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**(18):3702–3704, 2004.

[160] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res*, **14**(6):1188–1190, 2004.

[161] P. Duplay, S. Szmelcman, H. Bedouelle, and M. Hofnung. Silent and functional changes in the periplasmic maltose-binding protein of Escherichia coli K12. I. Transport of maltose. *J Mol Biol*, **194**(4):663–673, 1987.

[162] T. A. Hall. BioEdit:a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**:95–98, 1999.

[163] D. Hanahan. Studies on transformation of Escherichia coli with plasmids. *J Mol Biol*, **166**(4):557–580, 1983.

[164] W. J. Dower, J. F. Miller, and C. W. Ragsdale. High efficiency transformation of E. coli by high voltage electroporation. *Nucleic Acids Res*, **16**(13):6127–6145, 1988.

[165] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, **74**(12):5463–5467, 1977.

[166] X. Chen, C. Molino, L. Liu, and B. M. Gumbiner. Structural elements necessary for oligomerization, trafficking, and cell sorting function of paraxial protocadherin. *J Biol Chem*, **282**(44):32128–32137, 2007.

[167] R. Li, R. Gorelik, V. Nanda, P. B. Law, J. D. Lear, W. F. DeGrado, and J. S. Bennett. Dimerization of the transmembrane domain of Integrin alphaIIb subunit in cell membranes. *J Biol Chem*, **279**(25):26666–26673, 2004.

[168] R. Li, C. R. Babu, J. D. Lear, A. J. Wand, J. S. Bennett, and W. F. DeGrado. Oligomerization of the integrin alphaIIbbeta3: roles of the transmembrane and cytoplasmic domains. *Proc Natl Acad Sci U S A*, **98**(22):12462–12467, 2001.

[169] R. Li, N. Mitra, H. Gratkowski, G. Vilaire, R. Litvinov, C. Nagasami, J. W. Weisel, J. D. Lear, W. F. DeGrado, and J. S. Bennett. Activation of integrin alphaIIbbeta3 by modulation of transmembrane helix associations. *Science*, **300**(5620):795–798, 2003.

[170] B. W. Berger, D. W. Kulp, L. M. Span, J. L. DeGrado, P. C. Billings, A. Senes, J. S. Bennett, and W. F. DeGrado. Consensus motif for integrin transmembrane helix association. *Proc Natl Acad Sci U S A*, **107**(2):703–708, 2010.

[171] T.-L. Lau, C. Kim, M. H. Ginsberg, and T. S. Ulmer. The structure of the integrin alphaIIbbeta3 transmembrane complex explains integrin transmembrane signalling. *EMBO J*, **28**(9):1351–1361, 2009.

[172] C. Kim, T.-L. Lau, T. S. Ulmer, and M. H. Ginsberg. Interactions of platelet integrin alphaIIb and beta3 transmembrane domains in mammalian cell membranes and their role in integrin activation. *Blood*, **113**(19):4747–4753, 2009.

[173] W. Wang, J. Zhu, T. A. Springer, and B.-H. Luo. Tests of integrin transmembrane domain homo-oligomerization during integrin ligand binding and signaling. *J Biol Chem*, **286**(3):1860–1867, 2011.

[174] K. Parthasarathy, X. Lin, S. M. Tan, S. K. A. Law, and J. Torres. Transmembrane helices that form two opposite homodimeric interactions: an asparagine scan study of alphaM and beta2 integrins. *Protein Sci*, **17**(5):930–938, 2008.

[175] A. L. Cornish, S. Freeman, G. Forbes, J. Ni, M. Zhang, M. Cepeda, R. Gentz, M. Augustus, K. C. Carter, and P. R. Crocker. Characterization of siglec-5, a novel glycoprotein expressed on myeloid cells related to CD33. *Blood*, **92**(6):2123–2132, 1998.

[176] W. M. Janeway CA Jr, Travers P. Immunobiology: The Immune System in Health and Disease. 5th edition. New York: Garland Science; 2001. Available from: http://www.ncbi.nlm.nih.gov/books/NBK10757/. Garland Science, 2001.

[177] E. Peles, M. Nativ, M. Lustig, M. Grumet, J. Schilling, R. Martinez, G. D. Plowman, and J. Schlessinger. Identification of a novel contactin-associated transmembrane receptor with multiple domains implicated in protein-protein interactions. *EMBO J*, **16**(5):978–988, 1997.

[178] E. M. Krämer, C. Klein, T. Koch, M. Boytinck, and J. Trotter. Compartmentation of Fyn kinase with glycosylphosphatidylinositol-anchored molecules in oligodendrocytes facilitates kinase activation during myelination. *J Biol Chem*, **274**(41):29042–29049, 1999.

[179] Y. Chen, F. Gao, F. Chu, H. Peng, L. Zong, Y. Liu, P. Tien, and G. F. Gao. Crystal structure of myeloid cell activating receptor leukocyte Ig-like receptor A2 (LILRA2/ILT1/LIR-7) domain swapped dimer: molecular basis for its non-binding to MHC complexes. *J Mol Biol*, **386**(3):841–853, 2009.

[180] B. C. Lewis, P. I. Mackenzie, and J. O. Miners. Homodimerization of UDP-glucuronosyltransferase 2B7 (UGT2B7) and identification of a putative dimerization domain by protein homology modeling. *Biochem Pharmacol*, **82**(12):2016–2023, 2011.

[181] O. Huber, R. Kemler, and D. Langosch. Mutations affecting transmembrane segment interactions impair adhesiveness of E-cadherin. *J Cell Sci*, **112 (Pt 23)**:4415–4423, 1999.

[182] R. Iino, I. Koyama, and A. Kusumi. Single molecule imaging of green fluorescent proteins in living cells: E-cadherin forms oligomers on the free cell surface. *Biophys J*, **80**(6):2667–2677, 2001.

[183] R. Laage, J. Rohde, B. Brosig, and D. Langosch. A conserved membrane-spanning amino acid motif drives homomeric and supports heteromeric assembly of presynaptic SNARE proteins. *J Biol Chem*, **275**(23):17481–17487, 2000.

[184] A. Stein, G. Weber, M. C. Wahl, and R. Jahn. Helical extension of the neuronal SNARE complex into the membrane. *Nature*, **460**(7254):525–528, 2009.

[185] M. W. Hofmann, K. Peplowska, J. Rohde, B. C. Poschner, C. Ungermann, and D. Langosch. Self-interaction of a SNARE transmembrane domain promotes the hemifusion-to-fusion transition. *J Mol Biol*, **364**(5):1048–1060, 2006.

[186] K. G. Fleming and D. M. Engelman. Computation and mutagenesis suggest a right-handed structure for the synaptobrevin transmembrane dimer. *Proteins*, **45**(4):313–317, 2001.

[187] R. Laage and D. Langosch. Dimerization of the synaptic vesicle protein synaptobrevin (vesicle-associated membrane protein) II depends on specific residues within the transmembrane segment. *Eur J Biochem*, **249**(2):540–546, 1997.

[188] R. Roy, R. Laage, and D. Langosch. Synaptobrevin transmembrane domain dimerization-revisited. *Biochemistry*, **43**(17):4964–4970, 2004.

[189] P. Washbourne, G. Schiavo, and C. Montecucco. Vesicle-associated membrane protein-2 (synaptobrevin-2) forms a complex with synaptophysin. *Biochem J*, **305 (Pt 3)**:721–724, 1995.

[190] J. Tong, P. P. Borbat, J. H. Freed, and Y.-K. Shin. A scissors mechanism for stimulation of SNARE-mediated lipid mixing by cholesterol. *Proc Natl Acad Sci U S A*, **106**(13):5141–5146, 2009.

[191] D.-H. Kweon, C. S. Kim, and Y.-K. Shin. Regulation of neuronal SNARE assembly by the membrane. *Nat Struct Biol*, **10**(6):440–447, 2003.

[192] X. Lu, Y. Zhang, and Y.-K. Shin. Supramolecular SNARE assembly precedes hemifusion in SNARE-mediated membrane fusion. *Nat Struct Mol Biol*, **15**(7):700–706, 2008.

[193] D. Langosch, E. Lindner, and R. Gurezka. In vitro selection of self-interacting transmembrane segments–membrane proteins approached from a different perspective. *IUBMB Life*, **54**(3):109–113, 2002.

7 Appendix

The appendix includes tables and figures as supplementary results. The first part shows supplementary figures of PD28 integration assays and Western blot expression analyses. These are controls for ToxR assay results. The second part contains tables that list ToxR interaction assay reporter activities and PD28 growth values from which the figures result.

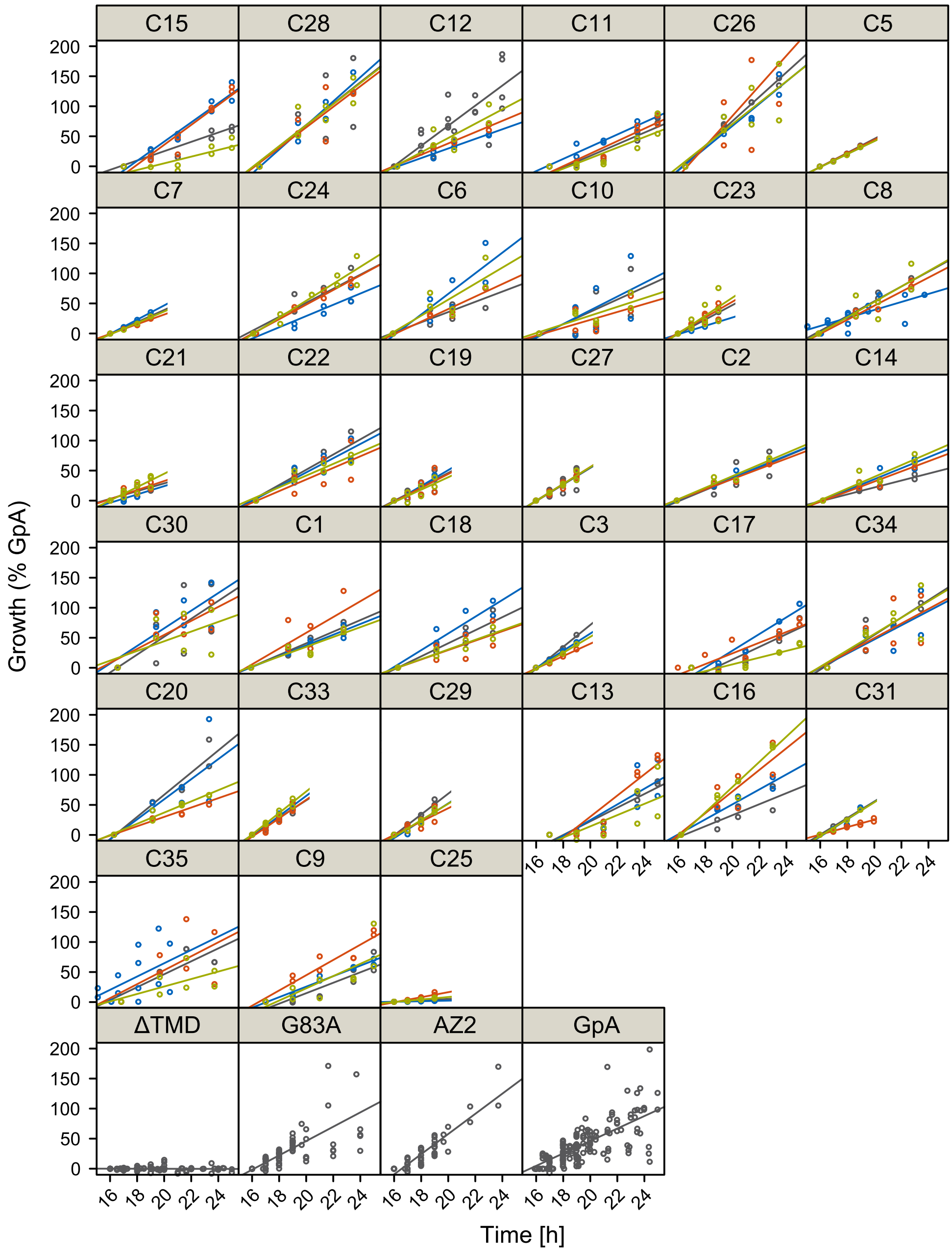

Figure 7.1: Control for membrane insertion of the ToxR-TMD-MalE fusion proteins tested for self-interaction in figure 3.5 on page 61. After incubation for at least 16 h in minimal medium containing 0.4% maltose as sole carbon source, the growth kinetics were obtained by measuring the OD_{600} for further 4-8 h and compared to that of the GpA construct (= 100%). For each representative sequence of the top clusters, the four different orientations were measured (-0 gray, -1 blue, -2 orange, -3 green). All clones except C25 were considered to express correctly membrane-integrated ToxR proteins since the slope of their growth curves is at least 50% of GpA.

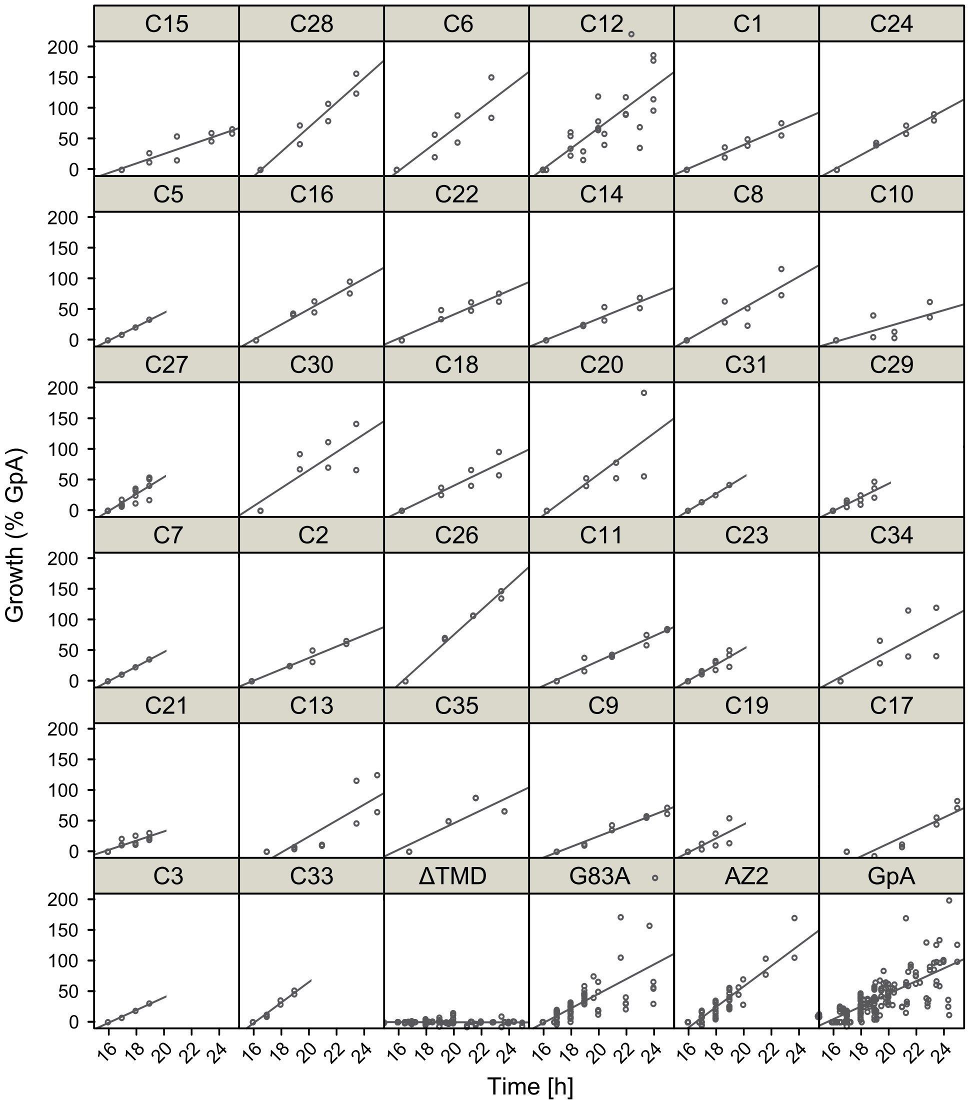

Figure 7.2: Control for membrane insertion of the ToxR-TMD-MalE fusion proteins tested for self-interaction in figure 3.6 on page 62. For each representative sequence of the top clusters, the optimal self-interaction orientation was measured. All clones were considered to express correctly membrane-integrated ToxR proteins since the slope of the growth curve is at least 50% of GpA.

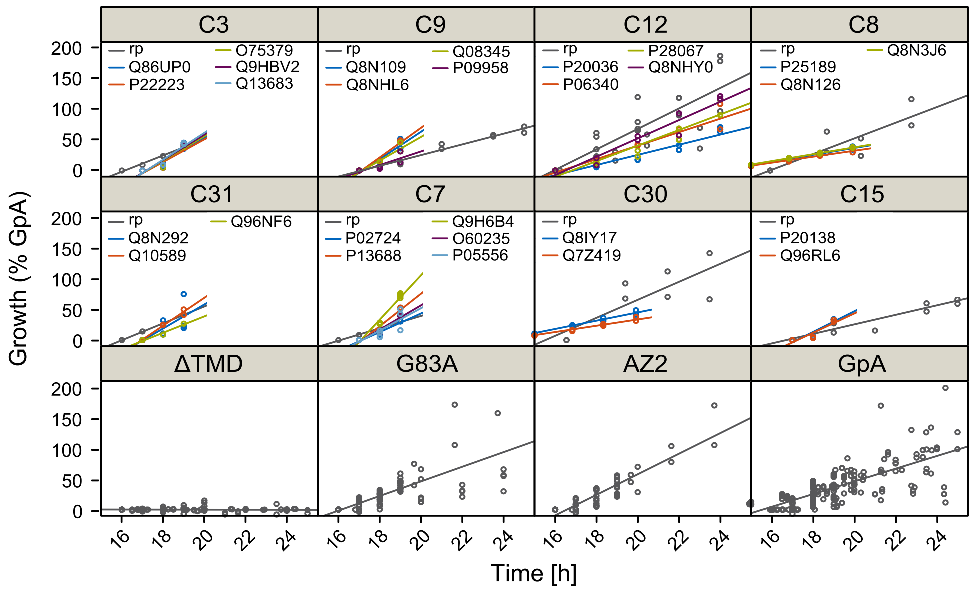

Figure 7.3: Control for membrane insertion of the ToxR-TMD-MalE fusion proteins tested for self-interaction in figure 3.7 on page 65. Each analyzed construct of exemplary clusters was tested. All clones were considered to express correctly membrane-integrated ToxR proteins since the slope of their growth curves is at least 50% of GpA.

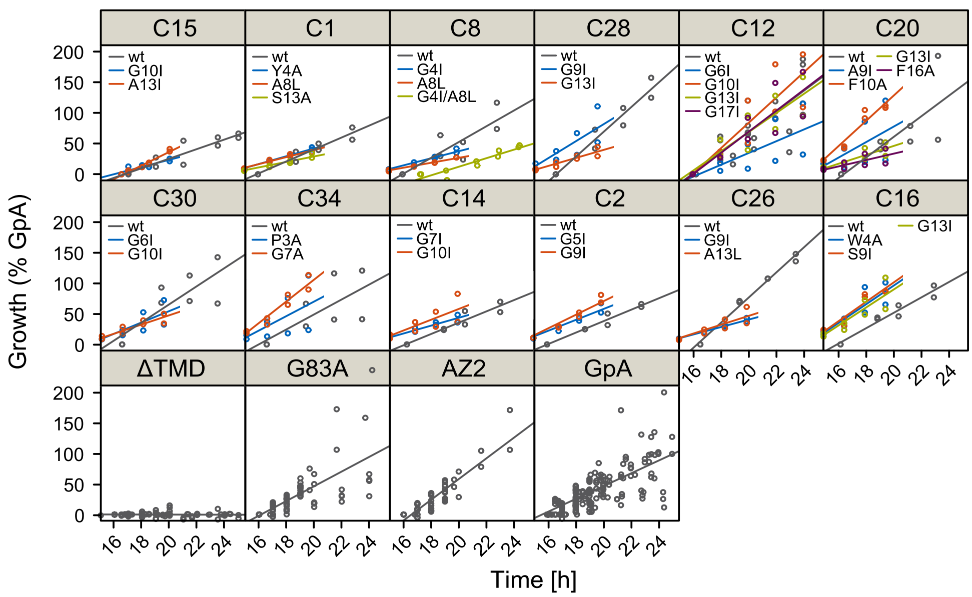

Figure 7.4: Control for membrane insertion of the ToxR-TMD-MalE fusion proteins tested for self-interaction in figure 3.8 on page 67. Each mutated construct was tested. All clones were considered to express correctly membrane-integrated ToxR proteins since the slope of their growth curves is at least 50% of GpA.

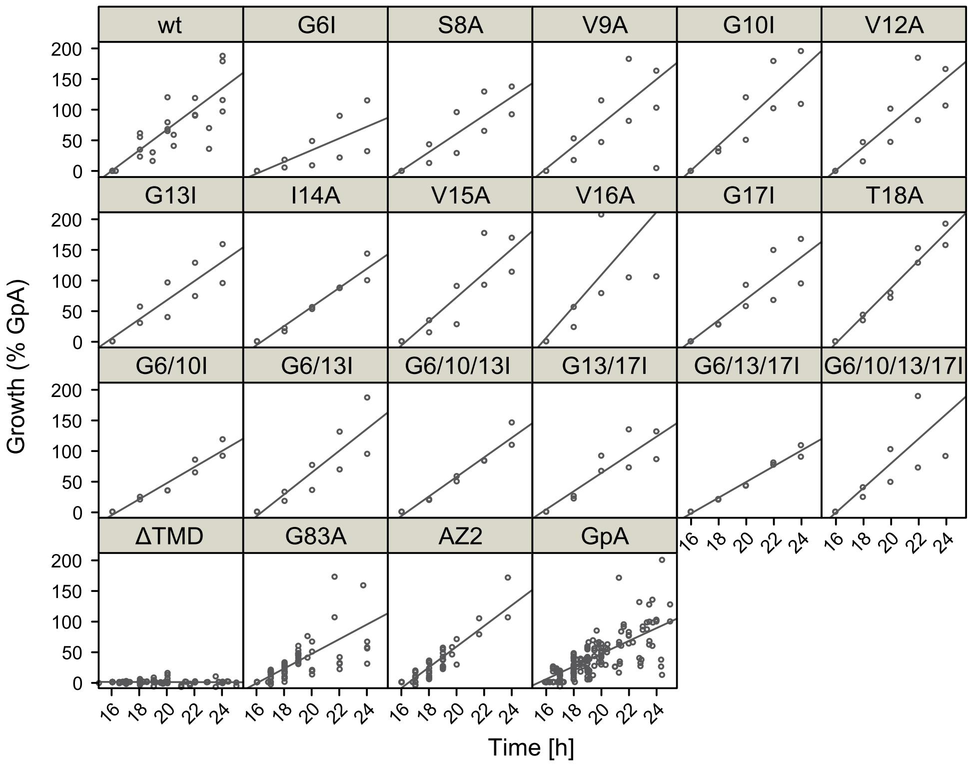

Figure 7.5: Control for membrane insertion of the ToxR-TMD-MalE fusion proteins tested for self-interaction in figure 3.9 on page 69. Each mutated construct of cluster C12 was tested. All clones were considered to express correctly membrane-integrated ToxR proteins since the slope of their growth curves is at least 50% of GpA.

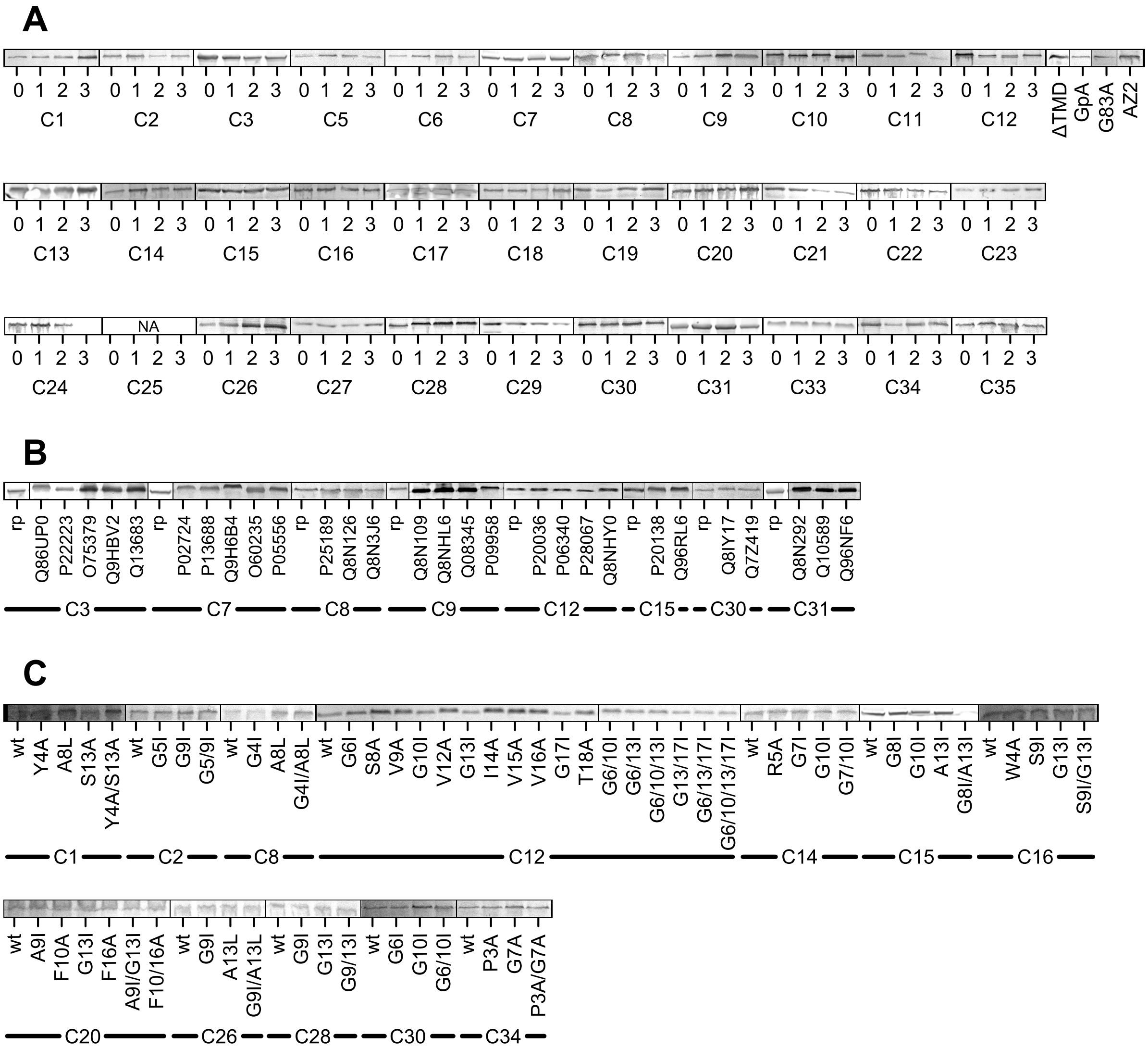

Figure 7.6: Western blot protein expression to test for sufficient ToxR-TMD-MalE fusion protein expression. (A) The protein expression of each representative sequence of the top clusters in four different orientations was tested (figure 3.5 and 3.6, page 61 and 62). All proteins were expressed sufficiently, except C25 and C24-3. The ToxR values of cluster 25 were not considered for interaction analysis due to insufficient expression and integration. Since the C24-3 proteins showed acceptable insertion into the membrane, the Western blot result was considered as incorrect. (B) The protein expression of fusion proteins of representative TMDs (rp) and similar TMD sequences within the same clusters (figure 3.7, page 65). Each construct led to sufficient protein expression. (C) The protein expression of wild type (wt) and mutated ToxR-TMD-MalE proteins (figure 3.8 and 3.9, page 67 and 69). All but one (C15-G8I/A13I) proteins were expressed adequately.

Table 7.1: Orientation-dependent homotypic interaction of representative TMDs.

	ToxR interaction assay							PD28 integration assay		
TMD[a]	Mean[b]	SD[c]	q25[d]	Median[e]	q75[f]	N[g]	ORD[h]	Growth[i]	log(growth)[i]	N[g]
C15-0	132.6	13.8	127.6	135.6	140.5	12	0.71	114.7	127.1	1
C15-1	77.1	7.4	70.6	77.0	85.3	12		63.6	78.7	4
C15-2	41.9	4.9	38.0	41.8	45.6	12		86.1	121.4	1
C15-3	37.1	5.7	32.1	39.4	42.4	12		80.6	119.5	1
C28-0	92.5	13.2	74.8	86.4	92.9	11	0.71	129.8	157.0	2
C28-1	109.2	17.3	93.1	102.4	106.7	11		146.7	194.3	2
C28-2	72.8	8.6	61.0	62.0	71.5	11		125.2	170.2	2
C28-3	34.3	8.4	27.3	30.2	32.5	11		128.2	185.5	2
C12-0	95.9	30.4	83.5	93.9	104.4	33	0.70	118.3	76.1	6
C12-1	71.1	4.7	68.5	72.6	74.8	11		61.7	51.4	2
C12-2	32.3	4.7	29.5	32.3	35.4	11		72.6	81.6	2
C12-3	27.7	2.6	26.0	27.9	30.0	11		92.1	79.0	2
C11-0	52.1	2.7	51.8	53.7	54.2	7	0.66	84.7	83.8	2
C11-1	59.1	6.0	56.7	61.4	64.3	9		94.4	87.6	1
C11-2	20.5	2.0	19.2	21.0	22.3	9		108.3	109.1	1
C11-3	55.6	3.7	54.6	57.0	59.0	12		87.1	84.6	6
C26-0	66.1	6.5	61.8	66.5	71.4	22	0.54	148.8	160.7	2
C26-1	57.5	3.0	56.3	59.4	60.0	11		135.5	146.8	2
C26-2	29.4	2.7	28.1	30.7	31.6	11		184.0	188.4	2
C26-3	36.4	1.6	35.6	37.1	38.1	11		130.6	181.5	2
C5-0	44.6	4.6	45.8	48.1	49.3	12	0.41	93.9	97.1	1
C5-1	62.1	6.0	59.3	64.6	67.2	12		87.2	75.0	1
C5-2	77.3	7.1	74.5	81.6	86.7	12		90.8	74.7	1
C5-3	54.8	5.9	54.1	54.5	59.0	12		86.7	73.4	1
C7-0	39.9	2.8	39.9	43.1	44.2	12	0.36	81.1	88.9	1
C7-1	64.0	11.5	57.9	67.3	75.6	12		95.4	87.4	1
C7-2	41.8	6.0	38.5	44.2	48.3	12		64.8	63.0	1
C7-3	46.0	3.6	47.2	49.2	50.1	12		77.0	74.0	1
C24-0	70.0	8.5	63.9	72.1	75.8	15	0.35	88.7	85.1	2
C24-1	64.9	8.3	60.0	65.5	69.0	15		71.1	74.2	2
C24-2	82.1	12.4	74.9	82.9	92.4	15		92.2	106.6	2
C24-3	56.2	13.3	47.6	53.6	67.1	15		107.3	135.8	2
C6-0	82.3	14.3	73.9	76.0	92.5	11	0.34	70.5	79.8	2
C6-1	92.5	7.6	88.9	91.0	97.6	11		142.0	121.9	2
C6-2	61.1	6.8	58.8	60.2	65.7	11		86.4	72.2	2
C6-3	78.5	4.0	78.1	79.0	81.8	11		110.4	81.0	2
C10-0	73.1	3.8	68.7	70.0	73.7	11	0.33	79.0	69.2	2
C10-1	59.6	2.7	56.7	58.8	59.5	11		88.3	58.0	2
C10-2	74.4	4.1	70.6	71.9	74.9	11		50.7	56.7	2
C10-3	51.7	4.6	47.5	48.4	51.6	11		56.3	57.4	2
C23-0	49.6	3.3	46.6	49.0	50.5	12	0.30	93.9	98.8	1
C23-1	52.3	12.1	43.6	55.3	61.0	11		54.4	109.7	1
C23-2	60.4	5.3	56.4	60.5	63.8	11		101.4	107.0	3
C23-3	41.1	8.6	40.7	42.5	44.9	12		116.3	139.1	3
C8-0	76.8	12.7	64.4	73.9	81.3	15	0.27	102.1	77.0	2
C8-1	59.5	12.5	50.3	56.9	73.3	15		76.1	67.0	4
C8-2	78.3	6.2	72.1	77.6	81.2	15		96.9	82.1	2
C8-3	79.8	6.5	74.3	77.1	81.1	12		105.9	77.5	2
C21-0	55.1	7.8	49.9	54.4	58.1	12	0.26	50.3	52.4	2
C21-1	46.0	3.0	44.3	45.0	46.5	11		54.1	63.4	2
C21-2	57.0	2.4	54.5	56.5	57.7	12		58.7	67.1	3
C21-3	42.9	8.6	37.3	42.0	46.0	12		87.1	93.0	3
C22-0	66.9	11.5	57.7	61.5	73.8	12	0.21	95.9	100.1	2
C22-1	75.6	11.7	68.0	71.7	79.7	12		89.7	94.5	2
C22-2	68.4	10.3	63.3	65.8	73.1	12		73.2	88.2	2
C22-3	82.5	13.2	68.8	78.3	86.4	10		73.8	97.3	2
C19-0	45.6	6.6	41.5	44.9	50.7	12	0.21	94.2	127.9	3
C19-1	37.6	5.6	33.9	37.6	39.8	12		107.5	132.6	2
C19-2	45.3	8.6	39.9	47.3	51.2	12		92.9	120.1	2
C19-3	41.0	6.5	36.6	39.2	47.8	12		84.2	92.1	2
C27-0	67.2	6.2	64.7	70.9	72.5	12	0.19	110.9	110.7	4
C27-1	64.1	4.9	61.0	65.8	69.1	12		110.5	91.0	1
C27-2	56.0	2.9	54.8	57.2	59.0	12		112.8	110.4	3
C27-3	43.8	46.0	60.5	63.1	67.6	12		112.3	99.6	2
C2-0	60.3	5.5	59.2	60.6	66.2	11	0.19	77.7	58.8	2
C2-1	65.9	3.6	64.0	67.2	70.3	11		76.7	63.9	2
C2-2	53.7	2.7	53.5	54.3	57.1	11		70.5	79.4	2
C2-3	62.3	8.5	55.2	64.9	69.1	32		79.5	87.6	2
C14-0	63.3	3.8	61.5	65.6	66.8	11	0.16	44.5	52.1	2
C14-1	75.3	3.8	74.4	78.2	79.1	11		70.7	72.3	2
C14-2	71.7	3.5	70.8	72.5	74.8	11		65.1	71.5	2
C14-3	69.6	4.3	67.7	69.9	74.1	11		75.3	76.4	2

Table continues on next page.

Table 7.1: Continued

TMD[a]	ToxR interaction assay Mean[b]	SD[c]	q25[d]	Median[e]	q75[f]	N[g]	ORD[h]	PD28 integration assay Growth[i]	log(growth)[i]	N[g]
C30-0	71.7	8.3	62.7	64.0	71.2	11	0.15	109.0	97.2	2
C30-1	87.0	13.7	69.6	74.5	87.1	11		107.5	112.0	2
C30-2	70.2	5.6	60.9	63.0	68.2	11		85.3	111.0	2
C30-3	75.8	3.2	67.5	69.2	71.2	11		59.0	70.0	2
C1-0	86.6	9.8	82.3	85.1	92.6	21	0.14	80.2	70.2	2
C1-1	80.1	6.6	77.3	80.4	86.1	11		74.0	102.9	2
C1-2	74.0	6.1	71.2	72.9	81.8	11		108.6	105.8	2
C1-3	71.9	24.9	76.2	77.8	81.8	11		67.6	63.9	2
C18-0	66.9	3.9	66.1	69.4	72.2	11	0.14	82.4	89.3	2
C18-1	50.5	46.1	64.1	65.6	67.4	11		105.6	115.5	2
C18-2	58.3	11.4	58.3	60.1	66.5	11		59.4	78.7	2
C18-3	59.0	4.8	57.3	59.7	64.8	11		60.9	82.8	2
C3-0	38.1	4.8	36.3	39.0	40.5	12	0.14	146.4	112.4	1
C3-1	43.1	7.1	40.1	43.7	45.5	11		114.8	90.5	1
C3-2	46.5	6.4	42.8	45.2	50.7	12		82.7	73.1	1
C3-3	42.4	6.5	41.6	43.9	46.3	12		106.6	91.8	1
C17-0	43.8	9.0	39.3	45.5	49.0	12	0.13	75.2	91.5	1
C17-1	41.0	4.5	37.9	39.6	45.3	12		93.6	102.6	2
C17-2	45.4	4.7	42.6	44.6	48.4	12		93.3	103.2	1
C17-3	44.2	6.7	38.7	45.7	50.3	12		76.9	91.7	3
C34-0	73.1	27.3	59.0	62.5	73.9	11	0.13	105.7	107.5	2
C34-1	66.0	15.5	56.0	58.0	61.2	11		86.2	95.4	2
C34-2	72.3	10.8	60.0	66.8	75.8	11		88.8	92.7	2
C34-3	71.6	9.5	61.4	62.6	69.9	11		99.6	100.4	2
C20-0	72.3	16.4	61.8	66.2	88.3	12	0.13	139.9	129.3	2
C20-1	70.5	8.5	65.0	69.0	76.4	12		126.7	125.5	2
C20-2	63.4	13.4	57.8	62.9	65.6	12		59.0	76.0	2
C20-3	64.3	7.0	59.7	60.3	70.0	12		70.7	103.4	2
C33-0	39.2	3.4	37.5	39.1	41.0	11	0.12	115.8	96.9	3
C33-1	44.5	3.2	43.0	43.8	47.2	11		134.1	105.4	2
C33-2	44.0	7.6	39.7	42.5	51.0	12		123.3	98.6	3
C33-3	40.6	7.2	36.8	38.7	45.0	12		143.2	115.2	2
C29-0	57.3	3.0	55.4	59.6	60.6	12	0.12	134.9	152.0	3
C29-1	63.4	3.9	62.5	64.7	66.1	12		112.5	123.6	1
C29-2	61.4	14.8	63.3	67.4	68.5	12		86.5	79.8	3
C29-3	61.9	5.2	60.9	62.9	66.8	12		106.3	133.8	2
C13-0	56.4	6.1	52.9	56.9	60.5	10	0.11	83.3	102.0	2
C13-1	54.3	2.2	54.0	55.1	55.8	12		102.8	111.9	1
C13-2	49.4	8.1	47.2	52.2	53.8	12		100.0	120.5	1
C13-3	52.2	4.6	48.5	50.9	56.7	11		75.0	117.2	2
C16-0	78.4	12.6	72.4	80.7	83.3	5	0.09	70.9	61.6	2
C16-1	81.9	12.4	77.5	81.5	89.8	22		96.4	83.6	2
C16-2	74.0	9.0	69.0	74.1	82.5	11		139.3	100.1	2
C16-3	79.5	6.0	78.7	81.4	83.9	11		162.3	155.3	2
C31-0	66.4	3.0	66.1	67.4	69.8	12	0.08	110.5	103.0	1
C31-1	64.1	0.5	65.6	66.0	66.1	3		115.3	94.3	1
C31-2	60.5	5.6	56.6	61.8	66.2	21		48.4	53.4	2
C31-3	65.8	4.4	63.3	65.9	69.9	12		112.6	89.8	1
C35-0	52.3	8.9	42.9	54.5	57.0	11	0.08	76.5	100.2	2
C35-1	50.5	5.6	46.0	50.1	55.5	12		132.0	104.5	4
C35-2	55.7	10.0	49.8	50.9	61.5	11		82.6	92.9	2
C35-3	51.6	9.0	47.9	54.1	55.4	11		44.2	50.5	2
C9-0	50.6	3.2	48.7	50.7	54.3	12	0.07	73.6	77.7	2
C9-1	54.2	8.4	49.7	53.1	56.6	12		115.7	101.3	3
C9-2	52.7	3.2	51.3	53.3	54.1	12		106.5	98.5	3
C9-3	52.5	4.5	50.6	54.6	55.7	12		161.1	133.3	2
ΔTMD	8.3	6.8	3.8	5.6	10.9	63		2.2	3.1	26
G83A	37.2	14.7	26.8	34.9	44.0	42		102.8	111.2	23
AZ2	82.9	17.3	71.9	73.4	87.0	22		106.5	101.4	17
GpA	100.0	15.3	93.9	100.0	107.0	350		100.0	100.0	45

[a] Most representative TMD of the cluster in 0-3 orientation.
[b] The mean of β-Gal activity measured for the TMD in % of GpA.
[c] The standard deviation of β-Gal activity measured for the TMD in % of GpA.
[d] The lower quartile (>25%) of β-Gal activity measured for the TMD in % of GpA.
[e] The median (>50%) of β-Gal activity measured for the TMD in % of GpA.
[f] The upper quartile (>75%) of β-Gal activity measured for the TMD in % of GpA.
[g] The number of measurements of ToxR activity or PD28 growth.
[h] The value for the orientation-dependence between 0 and 1 for the four orientations (2.3.3.1, page 27). Small values indicate a low dependence of β-Gal activity on TMD orientation.
[i] The PD28 integration assay growth in % of GpA calculated from the growth curve with linear regression or logarithmic regression.

Table 7.2: Self-interaction of most representative TMDs in optimal orientation.

	ToxR interaction assay							PD28 integration assay		
TMD[a]	Mean[b]	SD[c]	q25[d]	Median[e]	q75[f]	N[g]	ORD[h]	Growth[i]	log(growth)[i]	N[g]
C15-0	132.6	13.8	127.6	135.6	140.5	12	0.71	114.7	127.1	1
C28-1	109.2	17.3	93.1	102.4	106.7	11	0.71	146.7	194.3	2
C12-0	95.9	30.4	83.5	93.9	104.4	33	0.70	118.3	76.1	6
C6-1	92.5	7.6	88.9	91.0	97.6	11	0.34	142.0	121.9	2
C1-0	86.6	9.8	82.3	85.1	92.6	21	0.14	80.2	70.2	2
C24-2	82.1	12.4	74.9	82.9	92.4	15	0.35	92.2	106.6	2
C5-2	77.3	7.1	74.5	81.6	86.7	12	0.41	90.8	74.7	1
C16-1	81.9	12.4	77.5	81.5	89.8	22	0.09	96.4	83.6	2
C22-3	82.5	13.2	68.8	78.3	86.4	10	0.21	73.8	97.3	2
C14-1	75.3	3.8	74.4	78.2	79.1	11	0.16	70.7	72.3	2
C8-3	79.8	6.5	74.3	77.1	81.1	12	0.27	105.9	77.5	2
C30-1	87.0	13.7	69.6	74.5	87.1	11	0.15	107.5	112.0	2
C10-2	74.4	4.1	70.6	71.9	74.9	11	0.33	50.7	56.7	2
C27-0	67.2	6.2	64.7	70.9	72.5	12	0.19	110.9	110.7	4
C18-0	66.9	3.9	66.1	69.4	72.2	11	0.14	82.4	89.3	2
C20-1	70.5	8.5	65.0	69.0	76.4	12	0.13	126.7	125.5	2
C31-0	66.4	3.0	66.1	67.4	69.8	12	0.08	110.5	103.0	1
C29-2	61.4	14.8	63.3	67.4	68.5	12	0.12	86.5	79.8	3
C7-1	64.0	11.5	57.9	67.3	75.6	12	0.36	95.4	87.4	1
C2-1	65.9	3.6	64.0	67.2	70.3	11	0.19	76.7	63.9	2
C34-2	72.3	10.8	60.0	66.8	75.8	11	0.13	88.8	92.7	2
C26-0	66.1	6.5	61.8	66.5	71.4	22	0.54	148.8	160.7	2
C11-1	59.1	6.0	56.7	61.4	64.3	9	0.66	94.4	87.6	1
C23-2	60.4	5.3	56.4	60.5	63.8	11	0.30	101.4	107.0	3
C13-0	56.4	6.1	52.9	56.9	60.5	10	0.11	83.3	102.0	2
C21-2	57.0	2.4	54.5	56.5	57.7	12	0.26	58.7	67.1	3
C9-3	52.5	4.5	50.6	54.6	55.7	12	0.07	161.1	133.3	2
C35-0	52.3	8.9	42.9	54.5	57.0	11	0.08	76.5	100.2	2
C19-2	45.3	8.6	39.9	47.3	51.2	12	0.21	92.9	120.1	2
C17-0	43.8	9.0	39.3	45.5	49.0	12	0.13	75.2	91.5	1
C3-2	46.5	6.4	42.8	45.2	50.7	12	0.14	82.7	73.1	1
C33-1	44.5	3.2	43.0	43.8	47.2	11	0.12	134.1	105.4	2
ΔTMD	8.3	6.8	3.8	5.6	10.9	63		2.2	3.1	26
G83A	37.2	14.7	26.8	34.9	44.0	42		102.8	111.2	23
AZ2	82.9	17.3	71.9	73.4	87.0	22		106.5	101.4	17
GpA	100.0	15.3	93.9	100.0	107.0	350		100.0	100.0	45

[a] Most representative TMD of the cluster.
[b] The mean of β-Gal activity measured for the TMD in % of GpA.
[c] The standard deviation of β-Gal activity measured for the TMD in % of GpA.
[d] The lower quartile (>25%) of β-Gal activity measured for the TMD in % of GpA.
[e] The median (>50%) of β-Gal activity measured for the TMD in % of GpA.
[f] The upper quartile (>75%) of β-Gal activity measured for the TMD in % of GpA.
[g] The number of measurements of ToxR activity or PD28 growth.
[h] The value for the orientation-dependence between 0 and 1 for the four orientations (2.3.3.1, page 27). Small values indicate a low dependence of β-Gal activity on TMD orientation.
[i] The PD28 integration assay growth in % of GpA calculated from the growth curve with linear regression or logarithmic regression.

Table 7.3: Conservation of self-interaction within exemplary clusters of TMDs.

	ToxR interaction assay							PD28 integration assay		
TMD[a]	Mean[b]	SD[c]	q25[d]	Median[e]	q75[f]	N[g]	AV[h]	Growth[i]	log(growth)[i]	N[g]
C3-rp	46.5	6.4	42.8	45.2	50.7	12	7.1	106.63	91.76	1
C3-Q86UP0	48.5	3.7	47.4	50.2	51.8	12		102.17	107.53	5
C3-P22223	44.7	9.4	41.3	42.6	44.8	12		98.80	99.04	1
C3-O75379	42.3	9.4	38.3	45.9	47.1	12		107.23	107.82	2
C3-Q9HBV2	57.2	6.6	56.0	59.0	64.0	12		111.14	115.38	2
C3-Q13683	55.3	5.0	54.2	58.7	60.5	12		116.87	113.97	2
C9-rp	52.5	4.5	50.6	54.6	55.7	12	7.4	115.74	101.27	3
C9-Q8N109	56.8	9.9	48.7	54.6	64.0	12		119.08	121.54	3
C9-Q8NHL6	68.7	5.7	63.7	69.0	73.7	12		128.92	140.93	2
C9-Q08345	45.9	2.3	44.5	45.7	48.7	9		103.01	122.54	2
C9-P09958	48.2	4.8	45.0	48.3	50.3	11		59.34	104.42	2
C12-rp	95.9	30.4	83.5	93.9	104.4	33	12.0	118.32	76.07	6
C12-P20036	86.5	10.8	77.3	83.6	90.6	22		64.84	94.22	2
C12-P06340	89.8	12.9	82.4	89.2	95.2	22		86.12	96.28	2
C12-P28067	88.9	16.8	80.7	88.3	93.8	22		99.46	99.73	2
C12-Q8NHY0	68.8	8.9	61.8	66.7	72.6	22		117.16	76.51	2
C8-rp	79.8	6.5	74.3	77.1	81.1	12	13.3	105.92	77.51	2
C8-P25189	65.6	11.0	56.8	64.6	69.2	11		79.77	91.48	2
C8-Q8N126	63.1	24.9	48.3	63.3	71.2	11		70.02	84.68	2
C8-Q8N3J6	64.3	6.6	59.1	63.3	69.2	11		80.40	85.75	2
C31-rp	66.4	3.0	66.1	67.4	69.8	12	15.1	110.46	103.00	1
C31-Q8N292	54.2	5.2	50.9	54.8	57.6	12		104.76	98.27	3
C31-Q10589	48.3	3.6	46.7	49.3	52.1	12		121.43	113.81	2
C31-Q96NF6	52.5	2.4	51.6	53.0	54.5	12		71.43	93.68	1
C7-rp	64.0	11.5	57.9	67.3	75.6	12	19.5	81.12	88.94	1
C7-P02724	46.0	4.9	44.7	48.9	51.3	12		80.95	94.02	1
C7-P13688	43.3	3.1	44.7	46.4	47.2	12		141.57	128.47	1
C7-Q9H6B4	61.7	8.7	56.7	64.5	72.4	12		193.37	159.29	4
C7-O60235	28.5	3.6	26.0	28.6	31.4	12		105.42	135.64	2
C7-P05556	49.0	5.0	45.3	50.7	55.0	12		96.99	132.24	3
C30-rp	87.0	13.7	69.6	74.5	87.1	11	24.7	107.45	111.96	2
C30-Q8IY17	48.7	4.1	46.1	49.0	50.7	11		96.93	92.12	2
C30-Q7Z419	50.7	4.6	47.0	50.7	54.1	11		73.67	85.17	2
C15-rp	132.6	13.8	127.6	135.6	140.5	12	34.9	114.65	127.09	1
C15-P20138	89.3	11.7	81.5	89.8	99.7	12		90.06	107.56	2
C15-Q96RL6	110.7	4.9	109.1	111.7	114.4	12		84.04	101.08	2
ΔTMD	8.3	6.8	3.8	5.6	10.9	63		2.22	3.08	26
G83A	37.2	14.7	26.8	34.9	44.0	42		102.84	111.22	23
AZ2	82.9	17.3	71.9	73.4	87.0	22		106.51	101.36	17
GpA	100.0	15.3	93.9	100.0	107.0	350		100.00	100.00	45

[a] Most representative TMD (rp) of the cluster or UniProtKB identifier of a selected cluster member.
[b] The mean of β-Gal activity measured for the TMD in % of GpA.
[c] The standard deviation of β-Gal activity measured for the TMD in % of GpA.
[d] The lower quartile (>25%) of β-Gal activity measured for the TMD in % of GpA.
[e] The median (>50%) of β-Gal activity measured for the TMD in % of GpA.
[f] The upper quartile (>75%) of β-Gal activity measured for the TMD in % of GpA.
[g] The number of measurements of ToxR activity or PD28 growth.
[h] Average variance of median β-Gal activity of TMDs measured for this cluster as percentage of the median (>50%) of β-Gal activity of the representative TMD.
[i] The PD28 integration assay growth in % of GpA calculated from the growth curve with linear regression or logarithmic regression.

Table 7.4: Sequence-specificity of TMD self-interaction.

TMD[a]	ToxR interaction assay							PD28 integration assay		
	Mean[b]	SD[c]	q25[d]	Median[e]	q75[f]	N[g]	IPM[h]	Growth[i]	log(growth)[i]	N[g]
C15-wt	132.6	13.8	127.6	135.6	140.5	12		114.7	127.1	1
C15-G10I	29.1	9.6	24.6	32.5	35.5	12	0.76	46.0	68.9	2
C15-A13I	28.3	7.4	22.5	27.9	31.2	12	0.79	86.5	103.3	2
C1-wt	86.6	9.8	82.3	85.1	92.6	21		80.2	70.2	2
C1-Y4A	83.9	8.4	79.6	83.5	91.6	11	0.02	91.0	99.8	2
C1-A8L	64.3	12.4	57.9	58.8	71.5	11	0.31	85.6	94.6	2
C1-S13A	30.4	2.3	29.0	31.8	33.0	11	0.63	63.8	88.2	2
C8-wt	79.8	6.5	74.3	77.1	81.1	12		105.9	77.5	2
C8-G4I	58.5	17.2	51.3	58.0	75.5	7	0.25	82.3	98.9	2
C8-A8L	42.4	10.1	37.8	43.9	57.6	10	0.43	57.5	80.8	2
C8-G4IA8L	43.6	37.8	27.4	31.1	46.9	9	0.60	64.5	120.6	2
C28-wt	109.2	17.3	93.1	102.4	106.7	11		146.7	194.3	2
C28-G9I	59.7	10.0	54.6	56.9	68.5	11	0.44	168.6	144.1	2
C28-G13I	64.1	9.6	59.2	65.8	72.1	11	0.36	79.7	83.0	2
C12-wt	95.9	30.4	83.5	93.9	104.4	33		118.3	76.1	6
C12-G6I	84.8	17.3	81.6	88.8	98.1	21	0.05	74.0	86.9	2
C12-G10I	56.8	15.6	54.8	59.6	63.8	11	0.37	158.5	86.5	2
C12-G13I	61.2	7.3	57.4	65.0	67.2	11	0.31	119.8	73.8	2
C12-G17I	64.4	6.6	60.4	66.6	72.8	11	0.29	131.8	86.9	2
C20-wt	70.5	8.5	65.0	69.0	76.4	12		126.7	125.5	2
C20-A9I	56.3	5.0	53.1	58.0	60.5	9	0.16	154.4	114.1	2
C20-F10A	44.8	3.5	43.0	45.5	47.6	10	0.34	251.5	157.7	2
C20-G13I	50.7	1.9	50.0	50.8	51.5	9	0.26	88.7	91.7	2
C20-F16A	74.5	5.5	71.4	76.0	80.1	10	0.10	63.5	67.9	2
C30-wt	87.0	13.7	69.6	74.5	87.1	11		107.5	112.0	2
C30-G6I	57.5	6.1	53.0	56.9	61.6	11	0.24	113.0	116.7	2
C30-G10I	61.5	8.3	59.2	62.1	66.1	11	0.17	91.7	88.5	2
C34-wt	72.3	10.8	60.0	66.8	75.8	11		88.8	92.7	2
C34-P3A	63.3	12.6	61.4	66.2	70.9	11	0.01	144.3	109.6	2
C34-G7A	54.7	3.5	51.8	54.7	57.0	11	0.18	216.0	154.0	2
C14-wt	75.3	3.8	74.4	78.2	79.1	11		70.7	72.3	2
C14-G7I	91.9	9.7	87.3	90.7	104.3	11	0.16	91.5	92.0	2
C14-G10I	72.4	8.9	68.2	78.1	80.3	11	0.00	125.2	103.0	2
C2-wt	65.9	3.6	64.0	67.2	70.3	11		76.7	63.9	2
C2-G5I	56.5	9.6	48.0	58.9	63.9	10	0.12	126.8	117.0	2
C2-G9I	54.7	5.8	49.6	57.6	59.7	10	0.14	157.8	134.1	2
C26-wt	66.1	6.5	61.8	66.5	71.4	22		148.8	160.7	2
C26-G9I	58.8	8.2	54.1	59.3	62.9	10	0.11	86.5	92.4	2
C26-A13L	62.6	6.7	58.7	62.9	69.2	11	0.05	102.7	109.4	2
C16-wt	81.9	12.4	77.5	81.5	89.8	22		96.4	83.6	2
C16-W4A	76.7	5.2	74.7	79.0	82.3	11	0.03	186.0	136.9	2
C16-S9I	76.5	9.0	74.2	76.5	84.3	11	0.06	192.8	143.6	2
C16-G13I	88.7	18.2	80.5	88.7	96.3	11	0.09	179.1	157.3	2
ΔTMD	8.3	6.8	3.8	5.6	10.9	63		2.1	2.9	28
G83A	37.2	14.7	26.8	34.9	44.0	42		102.8	111.2	23
AZ2	82.9	17.3	71.9	73.4	87.0	22		106.5	101.4	17
GpA	100.0	15.3	93.9	100.0	107.0	350		100.0	100.0	47

[a] Most representative TMD of the cluster (wt) or mutated TMD.
[b] The mean of β-Gal activity measured for the TMD in % of GpA.
[c] The standard deviation of β-Gal activity measured for the TMD in % of GpA.
[d] The lower quartile (>25%) of β-Gal activity measured for the TMD in % of GpA.
[e] The median (>50%) of β-Gal activity measured for the TMD in % of GpA.
[f] The upper quartile (>75%) of β-Gal activity measured for the TMD in % of GpA.
[g] The number of measurements of ToxR activity or PD28 growth.
[h] Impact of point mutations calculated from β-Gal activities of wild-type and mutated form of the TMD (2.3.3.2, page 28).
[i] The PD28 integration assay growth in % of GpA calculated from the growth curve with linear regression or logarithmic regression.

Table 7.5: Specific self-interaction of the HLA class II α-chain TMD.

	ToxR interaction assay							PD28 integration assay		
TMD[a]	Mean[b]	SD[c]	q25[d]	Median[e]	q75[f]	N[g]	IPM[h]	Growth[i]	log(growth)[i]	N[g]
C12-wt	95.9	30.4	83.5	93.9	104.4	33		118.3	76.1	6
C12-G6I	84.8	17.3	81.6	88.8	98.1	21	0.05	74.0	86.9	2
C12-S8A	96.4	12.2	88.1	99.1	107.3	10	0.05	115.2	87.0	2
C12-V9A	83.7	15.2	82.1	85.9	94.2	11	0.09	140.0	96.7	2
C12-G10I	56.8	15.6	54.8	59.6	63.8	11	0.37	158.5	86.5	2
C12-V12A	85.2	18.6	80.8	88.4	95.1	21	0.06	144.6	102.8	2
C12-G13I	61.2	7.3	57.4	65.0	67.2	11	0.31	119.8	73.8	2
C12-I14A	70.2	11.7	69.5	75.0	81.6	11	0.20	119.6	107.4	2
C12-V15A	83.5	15.1	83.1	86.0	96.9	11	0.08	151.1	116.0	2
C12-V16A	105.3	10.9	101.9	106.4	109.2	11	0.13	197.7	108.9	2
C12-G17I	64.4	6.6	60.4	66.6	72.8	11	0.29	131.8	86.9	2
C12-T18A	97.0	18.5	86.7	93.6	105.4	11	0.00	172.6	79.0	2
C12-G6/10I	69.9	9.6	64.5	70.3	76.6	11	0.25	100.6	98.2	2
C12-G6/13I	75.7	18.5	65.3	73.1	83.4	12	0.22	137.0	104.5	2
C12-G6/10/13I	65.2	9.9	61.2	66.3	75.5	11	0.29	122.3	107.0	2
C12-G13/17I	84.6	6.1	84.0	89.3	90.4	11	0.05	114.2	78.5	2
C12-G6/13/17I	62.4	9.3	56.9	65.6	71.2	11	0.30	98.5	99.0	2
C12-G6/10/13/17I	63.2	7.6	59.8	62.8	71.6	11	0.33	154.8	103.1	2
ΔTMD	8.3	6.8	3.8	5.6	10.9	63		1.9	2.2	26
G83A	37.2	14.7	26.8	34.9	44.0	42		106.0	114.0	19
AZ2	82.9	17.3	71.9	73.4	87.0	22		103.7	101.5	13
GpA	100.0	15.3	93.9	100.0	107.0	350		100.0	100.0	45

[a] Most representative TMD of the cluster (wt) or mutated TMD.
[b] The mean of β-Gal activity measured for the TMD in % of GpA.
[c] The standard deviation of β-Gal activity measured for the TMD in % of GpA.
[d] The lower quartile (>25%) of β-Gal activity measured for the TMD in % of GpA.
[e] The median (>50%) of β-Gal activity measured for the TMD in % of GpA.
[f] The upper quartile (>75%) of β-Gal activity measured for the TMD in % of GpA.
[g] The number of measurements of ToxR activity or PD28 growth.
[h] Impact of point mutations calculated from β-Gal activities of wild-type and mutated form of the TMD (2.3.3.2, page 28).
[i] The PD28 integration assay growth in % of GpA calculated from the growth curve with linear regression or logarithmic regression.