BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 13 2013, pages 1623–1630
doi:10.1093/bioinformatics/btt247

Sequence analysis

Advance Access publication May 1, 2013

Self-interaction of transmembrane helices representing pre-clusters from the human single-span membrane proteins

Jan Kirrbach[1], Miriam Krugliak[2], Christian L. Ried[1], Philipp Pagel[3,†], Isaiah T. Arkin[2] and Dieter Langosch[1,*]

[1]Lehrstuhl für Chemie der Biopolymere, Technische Universität München, 85354 Freising, and Munich Center for Integrated Protein Science (CIPS[M]), 81377 Munich, Germany, [2]Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel and [3]Lehrstuhl für Genomorientierte Bioinformatik, Technische Universität München, 85354 Freising, Germany

Associate Editor: John Hancock

ABSTRACT

Motivation: Most integral membrane proteins form dimeric or oligomeric complexes. Oligomerization is frequently supported by the non-covalent interaction of transmembrane helices. It is currently not clear how many high-affinity transmembrane domains (TMD) exist in a proteome and how specific their interactions are with respect to preferred contacting faces and their underlying residue motifs.

Results: We first identify a threshold of 55% sequence similarity, which demarcates the border between meaningful alignments of TMDs and chance alignments. Clustering the human single-span membrane proteome using this threshold groups ~40% of the TMDs. The homotypic interaction of the TMDs representing the 33 largest clusters was systematically investigated under standardized conditions. The results reveal a broad distribution of relative affinities. High relative affinity frequently coincides with (i) the existence of a preferred helix–helix interface and (ii) sequence specificity as indicated by reduced affinity after mutating conserved residues.

Contact: langosch@tum.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 20, 2013; revised on April 23, 2013; accepted on April 25, 2013

1 INTRODUCTION

Most integral membrane proteins assemble to non-covalent functional oligomers. This oligomerization is frequently supported by interactions of their α–helical transmembrane domains (TMDs) (Arbely and Arkin, 2004; Cymer *et al.*, 2012; Popot and Engelman, 2000; Rath and Deber, 2008; Senes *et al.*, 2004). On the one hand, their association is favored by the constraints of the lipid bilayer, which concentrates and pre-orients the proteins (Grasberger *et al.*, 1986). On the other hand, TMD–TMD interactions frequently depend on recurrent interfacial amino acid motifs. For example, the TMD helix–helix interface of Glycophorin A (GpA) contains a critical GxxxG motif, which also promotes close association of many other helices (Arbely and Arkin, 2004; MacKenzie *et al.*, 1997; Smith *et al.*, 2001). In some cases, variations of GxxxG, the G/A/SxxxG/A/S motifs (Rath and Deber, 2008; Senes *et al.*, 2004), are crucial, as in ErbB receptors (Cymer and Schneider, 2010; Cymer *et al.*, 2012) and integrins (Luo and Springer, 2006; Schneider and Engelman, 2004). These motifs can cooperate with polar or aromatic amino acids within the same TMD (Herrmann *et al.*, 2010; Unterreitmeier *et al.*, 2007), which makes them dependent on sequence context. Self-interaction of other TMDs is driven by S/T (Dawson *et al.*, 2002), QxxW motifs (Sal-Man *et al.*, 2004), aromatic residues (Johnson *et al.*, 2007; Ridder *et al.*, 2005) or by residues with carboxamide side chains (Laage and Langosch, 1997). TMD–TMD interactions have been studied most intensely with single-span membrane proteins that account for a substantial fraction of membrane proteins that increases from ~15% in bacteria to >40% in humans (Worch *et al.*, 2010). To date, dozens of single-span proteins are known to interact via their TMDs, and about one dozen of high-resolution structures have been published that reveal distinct helix–helix interfaces in structural detail reviewed in (Cymer *et al.*, 2012).

It is currently unclear which fraction of the >3000 predicted single-span membrane proteins self-assemble via TMD–TMD interactions. Further, it is not known whether strongly interacting TMDs generally exhibit distinct interfaces. It is likely that some interacting helices exhibit multiple interfaces depending on the functional state of the protein. Alternative interfaces may interchange by rotation (Seubert *et al.*, 2003) or by interaction of N-terminal versus C-terminal parts (Arkhipov *et al.*, 2013; Escher *et al.*, 2009). In addition, many TM helices exhibit unilateral residue conservation (Ried *et al.*, 2012; Zviling *et al.*, 2007). To which extent these interactions are based on conserved amino acid motifs, such as G/A/SxxxG/A/S, has also been debated (Li *et al.*, 2012).

Here, we approached these issues in a systematic way by identifying TMDs that represent clusters of homologous sequences and by experimentally characterizing their self-interaction. To this end, we clustered the human single-span membrane proteome based on pairwise alignments of TMDs, using a meaningful sequence similarity threshold derived in this study. Analyzing the self-interaction of the TMDs representing each major cluster revealed a broad distribution of relative affinities. In a number of high-affinity TMDs, mutational analyses confirmed the sequence specificity of the self-interaction. Although the majority of

*To whom correspondence should be addressed.
†Present address: LipoFIT Analytic GmbH, Regensburg, Germany.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at Hebrew University of Jerusalem on July 24, 2013

clusters reflect the divergent evolution of duplicated and thus functionally related proteins, other clusters may have arisen by convergent evolution of interacting TMDs within structurally and functionally otherwise unrelated proteins.

2 METHODS

A database of human single-span membrane proteins was created from the UniProtKB database (UniProt Consortium, 2008) (release 57.9, October 2009), by selecting all human proteins annotated as 'single-pass membrane protein'. The TMDs and signal peptides were predicted using Phobius (Käll *et al.*, 2004) (version 1.01), but only proteins containing one TMD were selected for further analysis. From the resulting dataset of 3534 single-span membrane proteins, the TMD sequences were extracted. Only unique TMDs were retained, yielding a database of 2205 distinct TMD sequences.

The 'water' application from the EMBOSS package (Rice *et al.*, 2000) performed pairwise gapless alignments of TMDs and calculated pairwise Smith–Waterman bit scores (Smith and Waterman, 1981) between TMD sequences using the PHAT substitution matrix (Ng *et al.*, 2000) generated for TMDs. The following parameters were used: gapopen = 100.00, gapextend = 10.00, datafile = PHAT7573, aformat = score. By repeating the procedure for each sequence in the dataset, a matrix of all-against-all pairwise bit scores was obtained. The entire score matrix was normalized for TMD length by dividing the scores by their corresponding self-scores according to Equation (1).

$$ssr_{s1,s2} = \frac{S_{s1,s2} + S_{s2,s1}}{S_{s1} + S_{s2}} \times 100\% \quad (1)$$

$S_{s1,s2}$ and $S_{s2,s1}$ are the bidirectional bit scores of TMD 1 against TMD 2 and vice versa. S_{s1} and S_{s2} are the bit scores of TMDs 1 and 2 against themselves. $ssr_{s1,s2}$ represents the score/selfscore ratio (ssr) of TMD 1 against TMD 2 in percentage. To identify a meaningful similarity threshold for clustering, we compared the similarities of the TMD sequences in the database ($ssr_{natural}$) with the similarities of their randomized counterparts (ssr_{random}). By comparing the distributions of values of ssr_{random} and $ssr_{natural}$, a homology threshold of ssr = 55% was identified.

Clusters of TMDs that share ≥55% homology were built by searching the ssr matrix, whose order follows the appearance of proteins in the UniProtKB. The TMD with the largest number of hits was retrieved from our database together with its homologs and corresponds to the 'most representative' TMD sequence of the first cluster. The procedure was repeated on the reduced database until no further similarities ≥55% could be found.

The ToxR system (Langosch *et al.*, 1996) was used to measure self-interaction of TMDs. TMD sequences, which had a fixed length of 20 amino acids for better comparability, were introduced into the ToxR chimeric protein by ligation of respective oligonucleotide cassettes between the *Nhe*I and *Bam*HI sites of the pToxRV plasmid. All TMDs were introduced in four helical registers to determine the orientation dependence of the signal. Point mutations were introduced using the QuikChange® Site-Directed Mutagenesis Kit (Agilent Technologies, Waldbronn, Germany). TMD self-interaction was determined as described previously using 0.0025% (w/v) L-arabinose (Sigma-Aldrich, Steinheim, Germany) to induce expression and adding 0.4 mM Isoprenyl-1-thio-β-D-galactopyranoside (AppliChem, Darmstadt, Germany) (Gurezka *et al.*, 1999). A minimum of three replicates done in quadruplicate was performed for each construct.

The ToxR protein expression and efficiency of membrane integration was tested by complementing the MalE deficiency of *Escherichia coli* PD28 cells as described previously (Brosig and Langosch, 1998). Transformed PD28 cells were grown in minimal medium including 0.4% maltose (AppliChem, Darmstadt, Germany) as sole carbon source. OD_{600} was measured after 20–24 h and compared with the construct, which contains the TMD of GpA. 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 human GpA.

3 RESULTS

In the initial part of this study, our goal was to identify paradigmatic TMDs that would represent clusters of homologs and thus cover a significant part of the single-span membrane proteome. These representative TMDs were subsequently investigated for self-interaction.

3.1 TMD-based clustering of human single-span membrane proteins

As the level of sequence homology at which TMDs could be clustered was not known, we first identified a meaningful homology threshold. By performing all-against-all pairwise TMD alignments, we calculated score/selfscore ratios (ssr), which reflect the similarity between any two human TMD sequences (see Section 2 for details). The distribution of these $ssr_{natural}$ values was compared with the distribution of ssr_{random} values obtained by aligning the same TMD pairs after sequence randomization (Fig. 1). The obtained ssr_{random} reflect chance homology, which peaks ~17%. Both distributions significantly differ at high ssr values where alignments of natural TMDs are more frequent than chance alignments. Above a similarity threshold of ssr = 55%, alignments of natural TMD pairs are ≥20 times more abundant than those of randomized TMD pairs. In other words, the probability of aligning the average natural TMD pair at random above the 55% homology threshold is ≤5%.

Clustering the entire TMD database based on the similarity threshold of 55% groups 40.5% of the 2205 predicted human single-span protein TMDs into 278 clusters. 33 'top' clusters (C1–C3, C5–C31 and C33–C35) include ≥5 TMDs, each, and cover 13.5% of the human single-span proteome (Table 1)

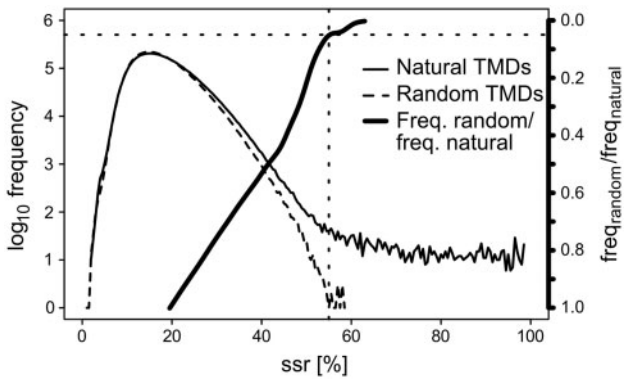

Fig. 1. Establishing a TMD homology threshold for cluster building. Frequency distributions of ssr values characterizing pairwise alignments of natural TMDs from human single-span proteins ($ssr_{natural}$, solid curve) and their randomized counterparts (ssr_{random}, dashed curve). The ssr_{random} values reflect chance homology. At high ssr values, alignments of natural TMDs are more abundant than chance alignments. The ratio of frequencies of $ssr_{random}/ssr_{natural}$ (bold curve) reveals that the probability of aligning two natural TMDs by chance is ≤5% above a threshold of ssr = 55%

Downloaded from http://bioinformatics.oxfordjournals.org/ at Hebrew University of Jerusalem on July 24, 2013

Table 1. The 33 top clusters of human single-span TMDs

Cluster	Representative protein[a]	Members[b]	Most prevalent functional annotation[c]	Functional diversity [%][d]	Representative TMD sequence[e]
C1	Q9UN71	29	Protocadherin	0	lqf**Y**Lvv**A**lali**S**vlflvam
C2	P01892	22	HLA class I α chain	9	ipiv**G**iiA**G**Lvlfgavitga
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-associated protein-like	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	vvcal**g**LsvG**l**v**G**ivv**G**tvl
C13	Q9H1U4	8	Syntaxin	63	niiiltviiivvvllmgfvg
C14	Q8IYS5	8	Leukocyte Ig	25	gnLvRl**g**lA**g**LvLisLgalv
C15	Q9Y286	8	Sialic-acid-binding Ig	13	vllgavgGa**G**at**A**lvflsfc
C16	P56199	8	Integrin α	13	vpl**W**vill**s**afa**G**llllmll
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	rkMLsGia**a**F**l**L**G**Li**f**llvG
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	0	HvLIGTSVVkipFtillFfL
C26	P23763	5	VAMP/Synaptobrevin	20	nckmmImL**G**aIC**A**iivvviv
C27	Q6PJG9	5	Fibronectin domain containing	0	GGTltvavGGvlVAalLVFt
C28	Q7L4S7	5	Armadillo-repeat containing	0	revGwmaA**G**lmi**g**AGacYcv
C29	Q14126	5	Desmoglein	20	glgPaaialmilafllLlLv
C30	Q8IZU9	5	Kin of IRRE-like protein 3	60	mavii**G**vav**G**aGvaflVlma
C31	Q8IUN9	5	No prevalent annotation	80	pchlllslGlglllviicv
C33	P32856	5	GRAM domain containing	20	rklmfiiicvivlLviLgii
C34	Q6UXC1	5	Lysosome associated protein	60	sv**P**avv**g**salllmllVLlg
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 Fig. 3). Uppercase amino acids are at least 90% conserved within the alignment of the cluster's members. Bold amino acids were selected for mutation analysis. Underlined G/A/SxxxG/A/S motifs are present in ≥60% of the members.

(We note that C4 and C32 members were re-annotated in UniProtKB as soluble in the course of this study and thus excluded here). Each cluster contains one most representative TMD, which is similar (ssr ≥55%) to all other members. Most (20/33) top clusters (C1–C3, C9, C10, C12, C14–C16, C18, C20–C22, C24–C29 and C33) contain mainly proteins of similar biological function according to the respective annotation in UniProtKB (UniProt Consortium, 2008) and are thus designated 'functionally homogeneous'. By contrast, >40% of the members of the top clusters C5–C8, C11, C13, C17, C19, C23, C30, C31, C34 and C35 have functions that deviate from the most prevalent function and are thus designated 'functionally heterogeneous' (Table 1).

To compare our TMD-based clustering to the more traditional approach of using complete sequences, we extended the clusters by including alignments between the complete sequences of the representative proteins and previously not clustered single-span proteins using a homology threshold of 25%. A sequence conservation level of 25% signifies structural homology of soluble proteins as shown by analyses of X-ray structures (Rost, 1999). This procedure increased the fraction of clustered proteins from 40.5 to 51.9%. This small increase in coverage provides support to the efficiency of our TMD-based clustering.

3.2 Homology and functional diversity

To explore the relationship between sequence homology and functional diversity within our clusters, we calculated pairwise ssr values at the level of TMD (ssr_{TMD}) and complete sequences ($ssr_{complete}$) between cluster members and their respective most representative sequences. The distribution of the $ssr_{TMD}/ssr_{complete}$ ratios shows roughly two major populations

Downloaded from http://bioinformatics.oxfordjournals.org/ at Hebrew University of Jerusalem on July 24, 2013

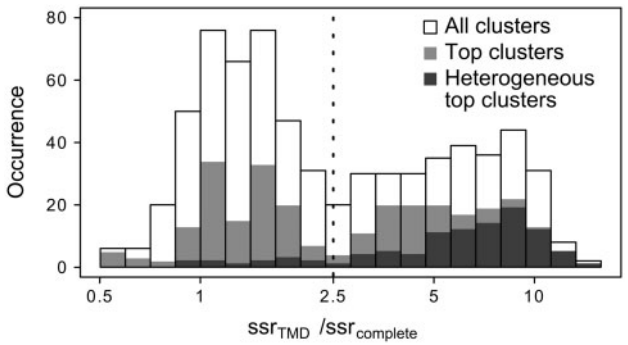

Fig. 2. Comparing TMD similarity to complete sequence similarity. Pairwise alignments of TMDs from cluster members to their respective most representative TMD sequences were used to calculate TMD similarities for each cluster (ssr_{TMD}). Similarly, the corresponding complete protein similarities ($ssr_{complete}$) were calculated. If a TMD is more similar to the representative TMD than is the complete protein sequence, the $ssr_{TMD}/ssr_{complete}$ ratio is >1. The distributions of these ratios were compared for all 298 clusters, the 33 top clusters and the subset of functionally heterogeneous top clusters. Bars are superimposed, i.e. their heights are non-cumulative. 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

of proteins (Fig. 2). Proteins where the homologies 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 ($ssr_{TMD}/ssr_{complete} > 2.5$). These distributions are compared for all clusters, our top clusters and for functionally heterogeneous top clusters. Interestingly, most members of functionally heterogeneous clusters are more homologous at the level of the TMDs than at the level of the complete sequence. Further, members of functionally heterogeneous clusters are less similar at the level of complete sequence ($ssr_{complete} = 13.6\%$) than functionally homogeneous clusters ($ssr_{complete} = 38.8\%$), which is below or above the 25% threshold, respectively, which signifies structural homology (Rost, 1999). As the complete sequences are mostly extramembranous, the latter tend to fold into 3D structures that are similar between the members of functionally homogeneous clusters but dissimilar within functionally homogeneous clusters. By contrast, the TMDs are rather similar ($ssr_{TMD} = 62.2\%$ for heterogeneous and $ssr_{TMD} = 71.3\%$ for homogeneous clusters).

3.3 Abundance of potential TMD–TMD interaction motifs

Clustered TMDs were searched for amino acid patterns that are implicated in TMD–TMD interaction (Table 1). Indeed, single or multiple GxxxG or GxxxA motifs were found twice as often in clustered TMDs (22.6%) compared with non-clustered TMDs (11.6%). In contrast, the abundance of G/A/SxxxG/A/S motifs is nearly independent of clustering (64.1% in clusters, 56.7% in non-clustered sequences). 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. The functionally heterogeneous top clusters exhibit a ~2-fold enrichment of GxxxG, AxxxG and SxxxG motifs compared with functionally homogeneous top clusters (Table 1). Taken together, certain G/A/SxxxG/A/S motifs are enriched within the clustered TMDs and in particular in those of functionally heterogeneous clusters.

3.4 Homotypic interaction of the most representative TMDs

The representative TMDs from the 33 top clusters were now tested for self-interaction in a biological membrane using the ToxR assay. Self-interaction of a ToxR/TMD/MalE hybrid protein at the level of its TMD in a bacterial membrane drives transcription of the *lacZ* gene that is under control of the *ctx* promoter (Langosch *et al.*, 1996). Affinities were determined relative to the high-affinity GpA TMD and its weakly interacting mutant G83A that is thought to produce a non-specific background signal (Langosch *et al.*, 1996) and to the medium-affinity AZ2 leucine zipper (Gurezka *et al.*, 1999). We initially determined the optimal orientation of the potentially interacting faces of the TMDs relative to the DNA-binding ToxR domain. 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 interfaces by up to 3×100, i.e. almost a full helix turn, relative to the ToxR domain. The differences between the β-Gal activities elicited by different constructs (Supplementary Table S1) indicate the extent to which self-interaction requires a specific interface, which is given by the 'orientation-dependence'. Six TMDs (C5, C11, C12, C15, C26 and C28) exhibit a clear preference for a particular TMD orientation (Fig. 3A, 'strongly orientation-dependent', indicated by dark shading). Another nine TMDs (C6–C8, C10, C19 and C21–C24) show little dependence on orientation ('weakly orientation-dependent', indicated by light shading). The other tested TMDs are nearly independent of orientation ('not orientation-dependent', without shading). Interestingly, four/six strongly orientation-dependent TMDs (C5, C11, C12 and C28) contain a GxxxG motif conserved in at least 60% of their homologs. In contrast, only 6/26 weakly or not orientation-dependent TMDs (C2, C7, C17, C20, C30 and C35) share this pattern. Therefore, orientation-dependent TMDs tend to contain conserved GxxxG motifs.

Comparing the relative affinities of TMDs in their optimal orientation reveals a broad range (Fig. 3B). Notably, high affinity TMDs tend to be orientation-dependent (Spearman's correlation test, $\rho = 0.48$, $P = 0.005$). To examine whether the relative affinities of TMDs are conserved within clusters, we selected eight clusters (C3, C7–C9, C12, C15, C30 and C31) of different interaction strength, functional diversity and GxxxG content (Supplementary Table S2). From each of these clusters, the self-interaction of 2–5 TMDs was determined. In six/eight chosen clusters (C3, C7–C9, C12 and C31), the relative affinity was comparable (±20%) to the most representative sequences (Supplementary Fig. S1), suggesting that the affinities tend to be conserved within the clusters. In the remaining clusters, the homology might mainly result from non-interfacial residue positions.

Finally, 12 representative TMDs were mutated to assess the sequence-specificity of self-interaction. The mutations target mainly G/A/SxxxG/A/S motifs and other highly conserved amino acids (Table 1, bold type). Depending on the TMD and

Downloaded from http://bioinformatics.oxfordjournals.org/ at Hebrew University of Jerusalem on July 24, 2013

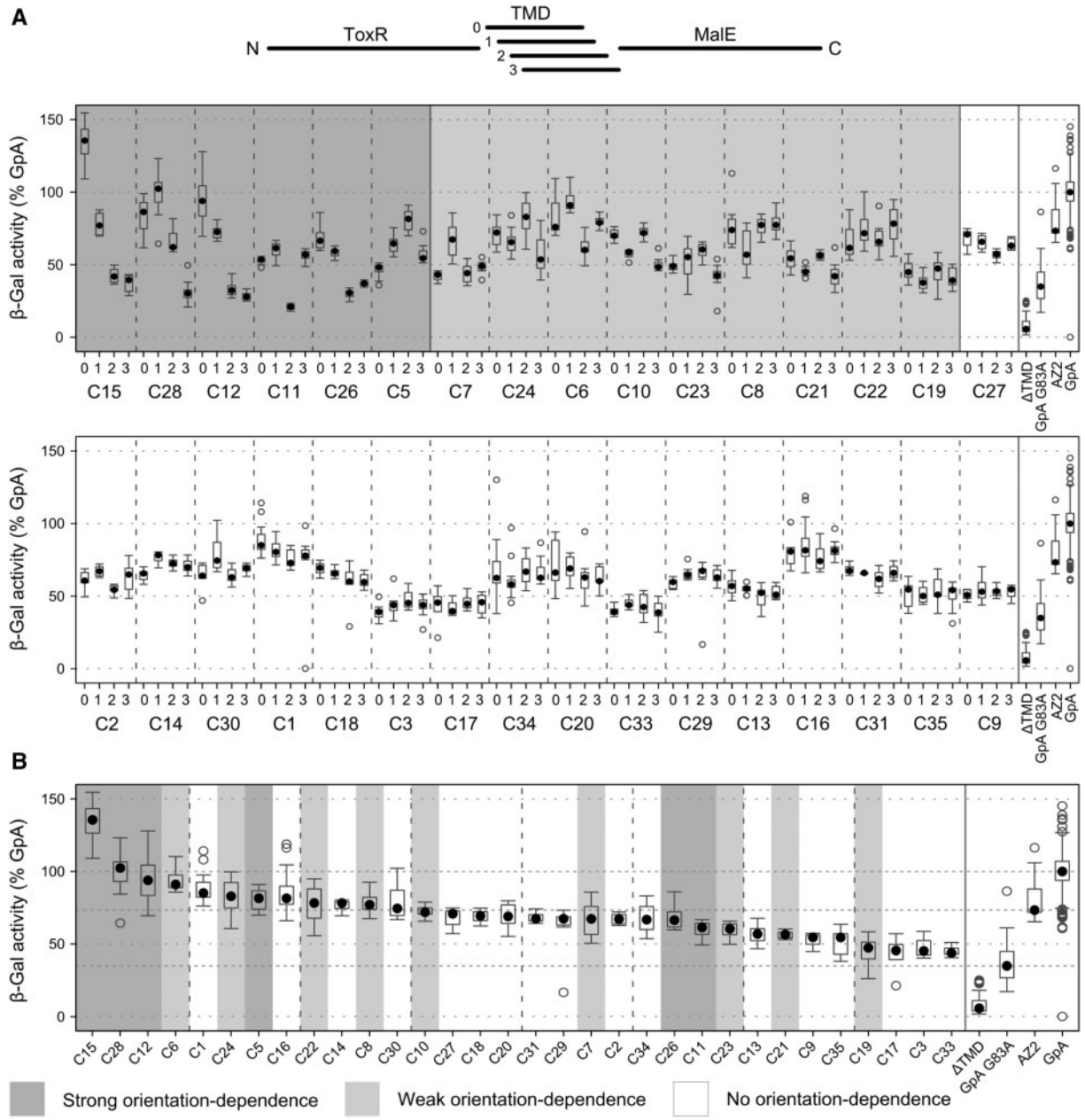

Fig. 3. Self-interaction of each top cluster's most representative TMD. Data represent relative β-Gal activities (GpA = 100%) as measured with the ToxR system [dot: median, box: interquartile range (IQR), whiskers: upper/lower quartile with max. 1.5 × IQR]. (**A**) Dependence of the β-Gal signal on the orientation of the TMD relative to the ToxR domain. The scheme at the top shows the stepwise insertion of three additional residues at a TMDs' N-terminus concurrent with the stepwise deletion of three residues at its C-terminus, which rotates the potential TMD–TMD interface by up to 3 × 100. The results are sorted according to the orientation dependence of the signal to structure the results. The different orientations of six TMDs show >40% difference in relative β-Gal activity, which is considered as strong orientation-dependence (dark shading), nine TMDs show weak orientation dependence with 20–40% difference in relative β-Gal activity (light shading), whereas the signal elicited by 17 TMDs is unaffected by orientation (no shading). (**B**) Self-interaction of TMDs in their optimal orientation, as identified in part A, ordered by decreasing affinity. The results of the PD28 assay that controls for membrane insertion are shown in Supplementary Figures S2 and S3

the type of targeted residue, mutation reduced the relative affinity by up to 79% of the wild-type TMD (Fig. 4). For 6/12 TMDs (C1, C8, C12, C15, C20 and C28), the relative affinity dropped by ≥30% (denoted 'mutation-sensitive'); the other mutated TMDs are termed 'mutation-insensitive'. In general, orientation-dependent TMDs tend to contain more mutation-sensitive amino acids than orientation-independent TMDs. Although mutating glycines had strong effects in 5/12 cases (C8, C12, C15, C20 and C28), GxxxG or other G/A/SxxxG/A/S motifs in several other TMDs are insensitive to mutation. We

Downloaded from http://bioinformatics.oxfordjournals.org/ at Hebrew University of Jerusalem on July 24, 2013

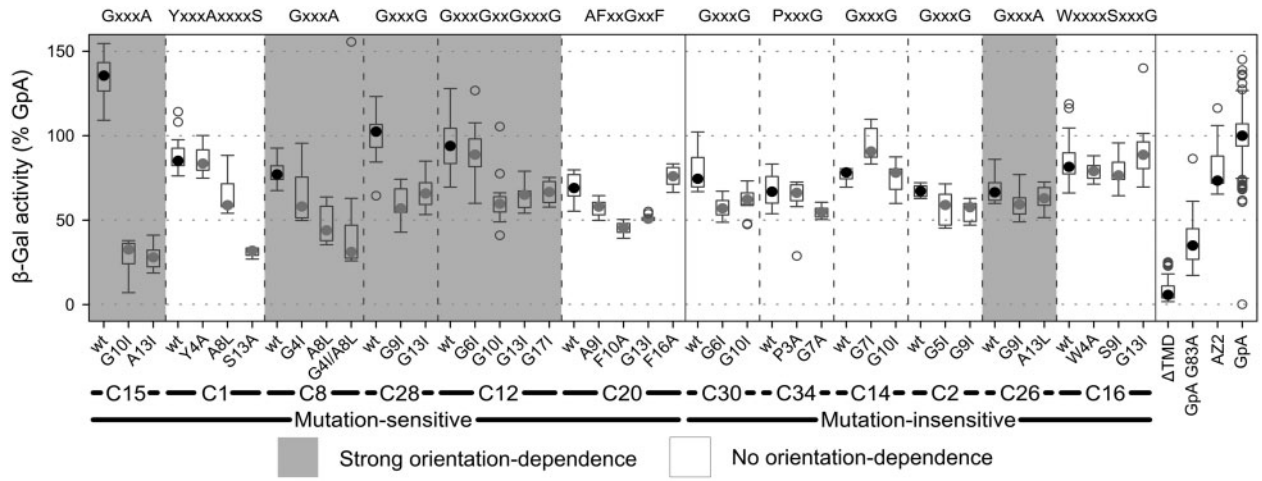

Fig. 4. Sequence specificity of self-interaction. Twelve exemplary sequences were mutated by exchanging the most conserved residues within the respective alignments. Data represent relative β-Gal activities (GpA = 100%) as measured with the ToxR system [dot: median, box: interquartile range (IQR), whiskers: upper/lower quartile with max. 1.5 × 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 are classified as mutation insensitive if a mutation reduced the β-gal signal by <30%. TMDs showing orientation-dependent self-interaction (see Fig. 3A) 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 Supplementary Figure S4

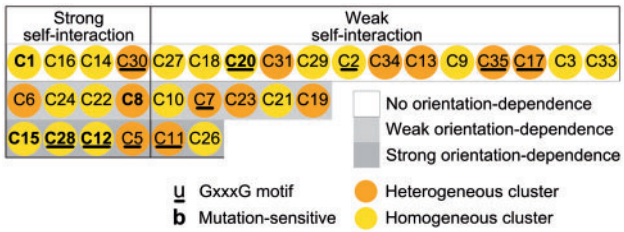

Fig. 5. The relationship of relative affinity, orientation-dependence (Fig. 3A), maximal impact of point mutations (Fig. 4), presence of conserved GxxxG motifs (Table 1) and functional homogeneity of the respective top clusters (Table 1) among most representative TMDs (Fig. 3B). Strong self-interaction (>AZ2 reference TMD) is often accompanied by orientation dependence and mutation sensitivity. Functionally heterogeneous clusters are enriched in conserved GxxxG motifs. Not all representative TMDs have been investigated to mutational analysis

note that mutation sensitivity is technically more difficult to establish for low-affinity TMDs, which exhibit β-Gal signals close to the low-affinity GpA G83A.

For control, each construct was tested for its proper insertion into the inner bacterial membrane by determining its ability to complement the MalE deficiency of *E.coli* PD28 cells; this strain lacks endogenous MalE and its growth in minimal medium with maltose as the only carbon source depends on correctly inserted ToxR/TMD/MalE hybrid proteins (see Section 2 and Supplementary Figs S2–S5). Different constructs elicit slightly different levels of complementation, which, however, do not correlate with β-Gal activity (Supplementary Fig. S6). Thus, normalizing β-Gal activities for slightly varying levels of complementation is not expected to improve the data. By contrast, strongly reduced membrane integration, as in the ToxR ΔTM negative control, prohibits β-Gal activity, as expected. Cluster C25 was removed from our dataset because of insufficient membrane integration (Supplementary Fig. S2).

4 DISCUSSION

The objective of this study was to identify TMDs representing a significant fraction of the human single-span membrane proteome and to systematically characterize their homotypic interactions. Our results have several implications.

First, our comparative analysis of homotypic affinity, which was done at the same lipid composition and protein density of the host membrane, shows that the representative TMDs of top clusters exhibit a broad range of relative affinities. The interaction of most high-affinity TMDs depends on preferential helix–helix interfaces and on conserved residues. In general, therefore, high relative affinity, existence of a preferred interface and mutation sensitivity characterize efficient and specific interaction (Fig. 5). On the other hand, there are exceptions to this rule. Two high-affinity TMDs (C1 and C8) may have multiple, yet sequence-specific interfaces. We cannot exclude, however, that terminal regions of these TMDs unwind to allow re-orientation of ToxR domains into an orientation where they can activate transcription. In C1, C8, C12, C15 and C28, mutating various G/A/SxxxG/A/S motifs or other conserved residues strongly reduced affinity, whereas no significant effect was seen in other cases (C14 and C16). Thus, some TMD–TMD interfaces appear to be rather robust, which may be a property that has been optimized through evolution. The mere presence of GxxxG or related motifs does not predict self-interaction (Li *et al.*, 2012).

Although homotypic TMD–TMD interactions have been demonstrated previously for members of some high-affinity clusters, e.g. for protocadherins (C1) (Chen *et al.*, 2007) and integrin α chains (C5 and C16) (Li *et al.*, 2001; Li *et al.*, 2004), a second outcome of this study is that it uncovers novel self-interacting

Downloaded from http://bioinformatics.oxfordjournals.org/ at Hebrew University of Jerusalem on July 24, 2013

TMDs. These include TMDs of sialic-acid-binding Ig (C15), armadillo repeat-containing proteins (C28), HLA class II α chains (C12) and others (Supplementary Table S1). The affinity of 12 TMDs is close to that of the structurally well-characterized high-affinity GpA dimer (Fig. 5). These TMDs represent a total of 124 TMDs homologous to them which equals 5.6% of sequence space. As they are likely to represent a similar percentage of structure space, they may be regarded as paradigmatic targets for future structure analysis. In some cases, self-interaction may parallel functionally relevant heterotypic interaction. For example, HLA class II α (C12) and β (C20) chains are known for heterotypic interaction via extramembraneous domains (Germain, 1995; Schafer *et al.*, 1995), which could be supported by the TMDs. The TMDs of integrin α (C5 and C16) and β (C7) chains also support heterodimerization (Berger *et al.*, 2010; Lau *et al.*, 2009), whereas the functional relevance of their homotypic interactions (Berger *et al.*, 2010; Li *et al.*, 2001; Li *et al.*, 2003; Li *et al.*, 2004; Schneider and Engelman, 2004) is unclear (Wang *et al.*, 2011). The homotypic interactions of several tested TMDs are rather inefficient, although some of them may be functionally important, as in cadherins (C3) (Huber *et al.*, 1999), integrin β chains (C7) (Li *et al.*, 2001; Li *et al.*, 2003), syntaxin (C13) (Hofmann *et al.*, 2006; Laage *et al.*, 2000) and synaptobrevins (C26) (Tong *et al.*, 2009). It should be borne in mind that low affinity detected under our standardized conditions does not exclude efficient TMD-based self-interaction of proteins that are present at high concentration. In addition, the lipid composition of the relevant host membrane may affect affinity.

Third, 90/265 pairwise TMD alignments within the top clusters suggest relationships between TMDs that belong to proteins being apparently unrelated in function. Part of these alignments might result from random homology. In other cases, TMD-based clustering of functionally unrelated proteins could reflect convergent evolution of their TMDs, despite dissimilar soluble domains. Interestingly, these TMDs are enriched in GxxxG, AxxxG and SxxxG motifs relative to the TMDs of functionally homogeneous clusters. Further, GxxxG motifs are conserved in TMDs of 6/13 functionally heterogeneous clusters but only in the TMDs of 4/20 functionally homogeneous clusters (Fig. 5). This might indicate that TMDs of functionally unrelated proteins tend to converge toward G/A/SxxxG/A/S-based interaction motifs. The evolution of such short interaction motifs requires fewer mutations than that of more complex helix–helix interfaces, and they could therefore develop rather rapidly by convergent evolution. In addition to homotypic TMD–TMD interaction, the conservation of TMD sequences within the functionally heterogeneous clusters could reflect conserved heterotypic interaction, interaction with lipids and/or binding of co-factors.

Finally, the homology threshold of 55% may be useful when aligning membrane proteins for homology-based structure modeling, especially for proteins that mainly consist of TMDs. It must be borne in mind, however, that the threshold may be different for multi-span proteins.

ACKNOWLEDGEMENTS

The authors would like to thank Barbara Rauscher for help with ToxR experiments, Manuel Mohr and Felix Behr for contributing ToxR data of cluster 12, and Christina Scharnagl for valuable comments on the manuscript.

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (grant La699/13-1 to D.L. and to I.A.), the Center for Integrative Protein Science Munich, the Deutsche Akademische Austauschdienst, and the TUM Graduate School.

Conflict of Interest: none declared.

REFERENCES

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

Arkhipov,A. *et al.* (2013) Architecture and membrane interactions of the EGF receptor. *Cell*, **152**, 557–569.

Berger,B.W. *et al.* (2010) Consensus motif for integrin transmembrane helix association. *Proc. Natl Acad. Sci. USA*, **107**, 703–708.

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

Chen,X. *et al.* (2007) Structural elements necessary for oligomerization, trafficking, and cell sorting function of paraxial protocadherin. *J. Biol. Chem.*, **282**, 32128–32137.

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

Cymer,F. *et al.* (2012) Transmembrane helix-helix interactions are modulated by the sequence context and by lipid bilayer properties. *Biochim. Biophys. Acta*, **1818**, 963–973.

Dawson,J.P. *et al.* (2002) Motifs of serine and threonine can drive association of transmembrane helices. *J. Mol. Biol.*, **316**, 799–805.

Escher,C. *et al.* (2009) Two GxxxG-like motifs facilitate promiscuous interactions of the human ErbB transmembrane domains. *J. Mol. Biol.*, **389**, 10–16.

Germain,R.N. (1995) Binding domain regulation of MHC class II molecule assembly, trafficking, fate, and function. *Semin. Immunol.*, **7**, 361–372.

Grasberger,B. *et al.* (1986) Interaction between proteins localized in membranes. *Proc. Natl Acad. Sci. USA*, **83**, 6258–6262.

Gurezka,R. *et al.* (1999) A heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments. *J. Biol. Chem.*, **274**, 9265–9270.

Herrmann,J.R. *et al.* (2010) Ionic interactions promote transmembrane helix-helix association depending on sequence context. *J. Mol. Biol.*, **396**, 452–461.

Hofmann,M.W. *et al.* (2006) Self-interaction of a SNARE transmembrane domain promotes the hemifusion-to-fusion transition. *J. Mol. Biol.*, **364**, 1048–1060.

Huber,O. *et al.* (1999) Mutations affecting transmembrane segment interactions impair adhesiveness of E-cadherin. *J. Cell Sci.*, **112**, 4415–4423.

Johnson,R.M. *et al.* (2007) Aromatic and cation-pi interactions enhance helix-helix association in a membrane environment. *Biochemistry*, **46**, 9208–9214.

Käll,L. *et al.* (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.

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

Laage,R. *et al.* (2000) A conserved membrane-spanning amino acid motif drives homomeric and supports heteromeric assembly of presynaptic SNARE proteins. *J. Biol. Chem.*, **275**, 17481–17487.

Langosch,D. *et al.* (1996) Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. *J. Mol. Biol.*, **263**, 525–530.

Lau,T.L. *et al.* (2009) The structure of the integrin alphaIIbbeta3 transmembrane complex explains integrin transmembrane signalling. *EMBO J.*, **28**, 1351–1361.

Li,E. *et al.* (2012) Transmembrane helix dimerization: beyond the search for sequence motifs. *Biochim. Biophys. Acta*, **1818**, 183–193.

Li,R. *et al.* (2001) Oligomerization of the integrin alphaIIbbeta3: roles of the transmembrane and cytoplasmic domains. *Proc. Natl Acad. Sci. USA*, **98**, 12462–12467.

Li,R. *et al.* (2003) Activation of integrin alphaIIbbeta3 by modulation of transmembrane helix associations. *Science*, **300**, 795–798.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Hebrew University of Jerusalem on July 24, 2013

Li,R. *et al.* (2004) Dimerization of the transmembrane domain of Integrin alphaIIb subunit in cell membranes. *J. Biol. Chem.*, **279**, 26666–26673.
Luo,B.H. and Springer,T.A. (2006) Integrin structures and conformational signaling. *Curr. Opin. Cell Biol.*, **18**, 579–586.
MacKenzie,K.R. *et al.* (1997) A transmembrane helix dimer: structure and implications. *Science*, **276**, 131–133.
Ng,P.C. *et al.* (2000) PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*, **16**, 760–766.
Popot,J.L. and Engelman,D.M. (2000) Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.*, **69**, 881–922.
Rath,A. and Deber,C.M. (2008) Surface recognition elements of membrane protein oligomerization. *Proteins*, **70**, 786–793.
Rice,P. *et al.* (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.
Ridder,A. *et al.* (2005) Tryptophan supports interaction of transmembrane helices. *J. Mol. Biol.*, **354**, 894–902.
Ried,C.L. *et al.* (2012) Homotypic interaction and amino acid distribution of unilaterally conserved transmembrane helices. *J. Mol. Biol.*, **420**, 251–257.
Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
Sal-Man,N. *et al.* (2004) The composition rather than position of polar residues (QxxS) drives aspartate receptor transmembrane domain dimerization *in vivo*. *Biochemistry*, **43**, 2309–2313.
Schafer,P.H. *et al.* (1995) The structure of MHC class II: a role for dimer of dimers. *Semin. Immunol.*, **7**, 389–398.
Schneider,D. and Engelman,D.M. (2004) Involvement of transmembrane domain interactions in signal transduction by alpha/beta integrins. *J. Biol. Chem.*, **279**, 9840–9846.
Senes,A. *et al.* (2004) Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr. Opin. Struct. Biol.*, **14**, 465–479.
Seubert,N. *et al.* (2003) Active and inactive orientations of the transmembrane and cytosolic domains of the erythropoietin receptor dimer, *Mol. cell*, **12**, 1239–1250.
Smith,S.O. *et al.* (2001) Structure of the transmembrane dimer interface of glycophorin A in membrane bilayers. *Biochemistry*, **40**, 6553–6558.
Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
Tong,J. *et al.* (2009) A scissors mechanism for stimulation of SNARE-mediated lipid mixing by cholesterol. *Proc. Natl Acad. Sci. USA*, **106**, 5141–5146.
UniProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
Unterreitmeier,S. *et al.* (2007) Phenylalanine promotes interaction of transmembrane domains via GxxxG motifs. *J. Mol. Biol.*, **374**, 705–718.
Wang,W. *et al.* (2011) Tests of integrin transmembrane domain homo-oligomerization during integrin ligand binding and signaling. *J. Biol. Chem.*, **286**, 1860–1867.
Worch,R. *et al.* (2010) Focus on composition and interaction potential of single-pass transmembrane domains. *Proteomics*, **10**, 4196–4208.
Zviling,M. *et al.* (2007) How important are transmembrane helices of bitopic membrane proteins? *Biochim. Biophys. Acta*, **1768**, 387–392.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Hebrew University of Jerusalem on July 24, 2013