doi:10.1006/jmbi.1999.3489 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 296, 911–919

JMB

The GxxxG Motif: A Framework for Transmembrane Helix-Helix Association

William P. Russ and Donald M. Engelman*

Department of Molecular Biophysics & Biochemistry Yale University, New Haven CT 06520-8114, USA

*Corresponding author

In order to identify strong transmembrane helix packing motifs, we have selected transmembrane domains exhibiting high-affinity homo-oligomerization from a randomized sequence library based on the right-handed dimerization motif of glycophorin A. Sequences were isolated using the TOXCAT system, which measures transmembrane helix-helix association in the *Escherichia coli* inner membrane. Strong selection was applied to a large range of sequences ($\sim 10^7$ possibilities) and resulted in the identification of sequence patterns that mediate high-affinity helix-helix association. The most frequent motif isolated, GxxxG, occurs in over 80% of the isolates. Additional correlations suggest that flanking residues act in concert with the GxxxG motif, and that size complementarity is maintained at the interface, consistent with the idea that the identified sequence patterns represent packing motifs. The convergent identification of similar sequence patterns from an analysis of the transmembrane domains in the SwissProt sequence database suggests that these packing motifs are frequently utilized in naturally occurring helical membrane proteins.

Keywords: membrane protein; protein folding; TOXCAT; glycine; association

Introduction

The association of transmembrane helices in the plane of the membrane plays a crucial role in the folding and oligomerization of integral membrane proteins. Since polypeptides crossing the membrane generally adopt helical secondary structures in order to minimize the free energy associated with the burial of main-chain polar groups (Engelman & Steitz, 1981), the folding of integral membrane proteins can be considered in terms of interactions between the surfaces of these preformed domains (Popot & Engelman, 1990). An understanding of the forces that drive intramembranous helix-helix interactions may allow the prediction of the tertiary and quaternary structures of transmembrane regions.

The types of interactions that can mediate transmembrane domain association are currently under investigation (Arkin *et al.*, 1994; MacKenzie & Engelman, 1998; Russ & Engelman, 1999; Smith *et al.*, 1996; Zhou *et al.*, 2000). In the case of the glycophorin A (GpA) transmembrane domain, sequence-specific homo-dimerization is principally mediated by seven residues (LIxxGVxxGVxxT) that define the interface between a right-handed crossing of α-helices (Lemmon *et al.*, 1994). The combined results of site-directed mutagenesis (Lemmon *et al.*, 1992), computational modeling (Adams *et al.*, 1996) and solution NMR (MacKenzie *et al.*, 1997) have demonstrated that the dimer is stabilized by extensive van der Waals interactions along the length of the interface. The homo-pentameric transmembrane domain of phospholamban, while less well understood, appears to utilize similar interactions (Adams *et al.*, 1998; Arkin *et al.*, 1994; Simmerman & Jones, 1998). Additional studies (Bargmann *et al.*, 1986; Machamer *et al.*, 1993; Manolios *et al.*, 1990; Zhou *et al.*, 2000) have suggested that polar interactions can mediate transmembrane helix-helix association.

The interaction surface of the GpA transmembrane domain is composed of two ridges and their associated groove (MacKenzie *et al.*, 1997). In the dimer, one of the ridges packs into the groove formed in the opposite monomer. Although mutations that change the surface of the interface

Abbreviations used: GpA, glycophorin A; CAM, chloramphenicol; tm, transmembrane; CAT, chloramphenicol acetyltransferase; MBP, maltose-binding protein.

E-mail address of the corresponding author: donald.engelman@yale.edu

0022-2836/00/030911–9 $35.00/0

can destabilize the dimer, it is possible to restore dimerization by making compensatory changes to the opposing monomer (Lemmon *et al*., 1992). This suggests that the wild-type GpA interface is not unique in its ability to dimerize, and that many sequences can generate the complementary surfaces necessary for dimerization. The ability to predict transmembrane interactions will require the study of a larger repertoire of helix-helix packing motifs than is currently available.

In order to isolate previously unidentified transmembrane oligomerization domains, we have applied TOXCAT selection (Russ & Engelman, 1999) to a library of randomized transmembrane sequences to identify sequences with a strong propensity for association. TOXCAT exploits a dimerization-dependent activator of transcription (ToxR) to report the ability of transmembrane sequences to oligomerize in the *Escherichia coli* inner membrane (Figure 1). Transmembrane domain oligomerization results in activation of a gene encoding resistance to chloramphenicol (CAM), allowing a selection of self-associating transmembrane sequences. The library was designed to substitute randomly nine possible amino acids at seven positions, with a periodicity based on the GpA dimerization motif. The resulting library exhibits a range of affinities, assayed by CAM resistance. Analysis of the high-affinity isolates has revealed common themes in sequence motifs that mediate strong helix-helix association.

Results

To sample a large and complete set of dimer interfaces, it is necessary to limit the total number of sequences in the library to a number that can be created readily. To this end, two strategies have been employed: (i) the presentation of variable residues only along one helical face; and (ii) restriction of codon usage. These two restrictions reduce the total number of sequence variants to 4.8×10^6 (Figure 2).

We used a transmembrane domain of approximately 19 amino acid residues in length, with seven variable positions spaced as in the GpA dimerization motif. Synthetic DNA oligonucleotides encoding the transmembrane region were designed with degenerate codons at positions corresponding to the interfacial residues in a right-handed crossing of α-helices. Each of the seven positions (1, 2, 5, 6, 9, 10 and 13) was simultaneously allowed to vary over nine possible amino acids: glycine, alanine, valine, leucine, isoleucine, serine, threonine, proline and arginine. Since this set contains seven of the eight most frequently occurring amino acids in transmembrane domains (Senes *et al*., 2000), our library represents an optimal approximation of natural sequences. Intervening positions (on the opposite helical face) were fixed as either alanine (ALALIB), or leucine (LEULIB) in separate libraries. Transmembrane sequences were cloned into TOXCAT, selected for oligomerization using CAM, selected for membrane insertion using the malE complementation assay, and sequenced. Since there is no possibility of forming a charge pair between side-chains in this library, isolates containing arginine were suspected to be cytoplasmic and therefore excluded from the analysis.

Distribution of affinities

As an initial test of the feasibility of selecting oligomeric transmembrane sequences using TOXCAT, each of the populations of plasmid encoding each library was isolated and transformed into *E. coli* NT326, a *malE*$^-$ strain. Each library was subjected to a range of selective pressure by plating on varying concentrations of CAM, and screened for transmembrane insertion using the *malE* comple-

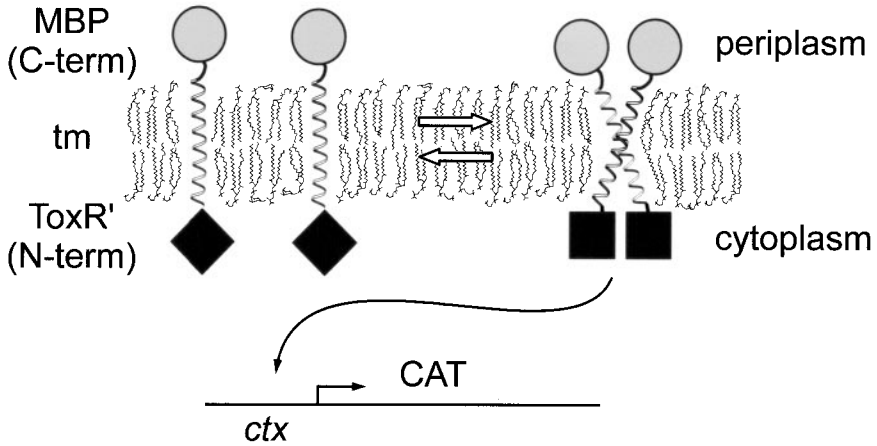

Figure 1. The TOXCAT assay for transmembrane domain oligomerization. Transmembrane (tm) helix-helix association mediates the activation of chloramphenicol acetyltransferase (CAT) from the *ctx* promoter by dimerized ToxR cytoplasmic domains (ToxR′, squares). The C-terminal, periplasmic maltose-binding protein (MBP, circles) domain anchors the chimera in the *E. coli* inner membrane.

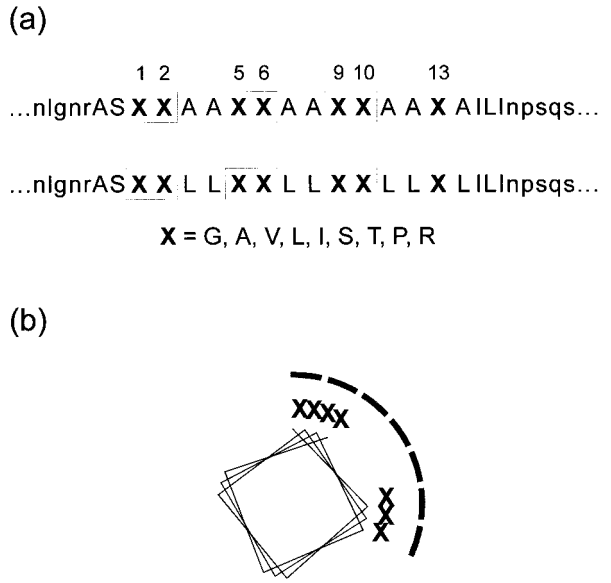

Figure 2. Design of transmembrane domain libraries. (a) One face of the helix was allowed to vary, at positions (X) that define the interface between right-handed crossing α-helices. The remaining positions were held constant as either alanine (ALALIB) or leucine (LEULIB). Positions marked X were allowed to vary over the following nine amino acids: glycine, alanine, valine, leucine, serine, threonine, proline and arginine. The probable transmembrane residues are shown in uppercase, and flanking residues are shown in lowercase. The variable region is numbered beginning with the first variable residue. (b) Helical wheel projection showing the distribution of variable residues at the interface between a right-handed crossing of α-helices.

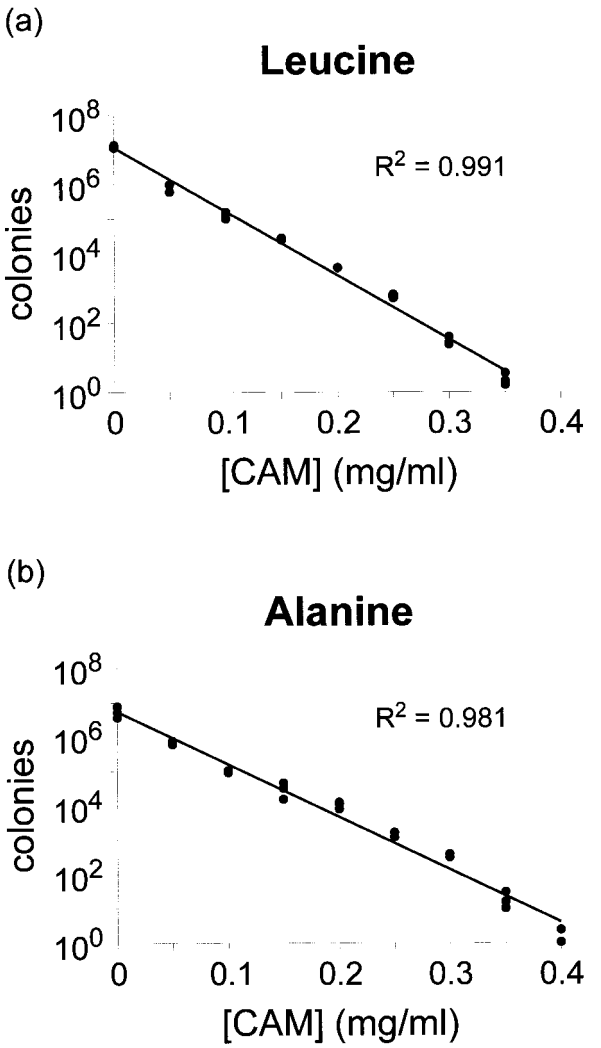

Figure 3. Oligomerization affinity of the population of transmembrane sequences. The sequence library, cloned into the TOXCAT construct, was transformed into bacteria and assayed for oligomerization. The number of colonies surviving as a function of CAM concentration is shown. As selection pressure increases, fewer sequences oligomerize with sufficient affinity to promote growth. Libraries containing (a) leucine and (b) alanine at constant positions in the variable region were assayed separately.

mentation assay. The observed logarithmic decay in the number of surviving colonies with increasing CAM concentration (Figure 3) suggests that the selection method is highly effective. First, it is evident that only a small subset of sequences encode strongly dimerizing transmembrane domains. Second, the observation that colony number decreases as a continuous function of CAM concentration suggests that a broad and continuous range of helix-helix affinities exists in this population of transmembrane sequences. Finally, a logarithmic relationship is seen in both libraries, suggesting that the trend in CAM resistance is independent of the context of non-interfacial residues. These results indicate that we have explored a broad range of helix-helix interactions, and a small subset of sequences exhibiting strong association can be isolated.

Selection and sequencing of high-affinity isolates

Helix-helix interactions with the highest affinities were analyzed. Colonies were picked for sequencing from plates containing sufficiently high concentration of chloramphenicol (250-350 μg/ml) to restrict the total number of surviving colonies to roughly 10^{-5} of the total possible sequences (~50-100 colonies). Surviving colonies were assayed for membrane insertion using *malE* complementation assays in order to ensure that only transmembrane sequences were analyzed. Several isolates were directly tested for oligomerization by measuring CAM resistance using disk diffusion assays (Russ & Engelman, 1999), and were found to give signals comparable to wild-type GpA constructs (not shown). The transmembrane sequences isolated are shown in Table 1. It should be noted that not every sequence with high oligomerization potential has

been isolated in this screen. Although a substantial percentage of the total population of sequences has been examined, the library screen was not complete. Evidence of this comes from the failure of this assay to isolate the exact GpA dimerization motif (LIxxGVxxGVxxT), although it should promote survival at the levels of CAM selection used (Russ & Engelman, 1999).

Sequence analysis

The normalized frequency of occurrence for each amino acid at the degenerate positions is shown in Figure 4. The frequencies shown have been corrected against a database of sequences that have been selected for membrane insertion, but not for association. There is a general trend of alternating small and large residues at positions 5 and 6, and positions 9 and 10 in both libraries. The most striking observation is the frequent occurrence of glycine at positions 5 and 9 of the LEULIB, and at positions 6 and 10 of the ALALIB. Given the background frequency of glycine in a set of sequences selected only for membrane insertion (not selected for association), the probability that such a high number of glycine residues would occur at random in the selected sequences (statistical significance, p) is $<10^{-18}$ for both libraries. Generally, these glycine residues occur as pairs with a separation of four residues (GxxxG): 96 % of the LEULIB ($p < 10^{-42}$) isolates and 79 % of the ALALIB isolates ($p < 10^{-45}$) have glycine at both positions. When glycine is not present at these positions, it is generally replaced by serine, a substitution often seen in the transmembrane domains of homologous proteins (Jones *et al*., 1994). If SxxxG and SxxxS pairs at these same positions are included, 100 % of the LEULIB and 93 % of the ALALIB sequences are accounted for. Additionally, glycine occurs with relatively high frequency at position 2 of the ALALIB (in 42 % of sequences), suggesting that tandem GxxxG motifs may further contribute to oligomerization.

Additional significant deviations from random include the apparent preference for bulky aliphatic residues (V, L and I) at positions 5, 9 and 13 of ALALIB. The occurrence of these residues follows the rank I > V > L, suggesting a selection for β-branched side-chains. The apparent requirement for β-branched side-chains at position 13 in ALALIB is mirrored by the frequent occurrence of Thr at position 13 in LEULIB ($p < 10^{-42}$). In the NMR structure of the GpA transmembrane dimerization domain (MacKenzie *et al*., 1997), packing of opposing Thr 87 γ-methyl groups against one another contributes to the stability of the dimer. MacKenzie & Engelman (1998) have provided a rationale to explain the use of β-branched side-chains at helix-helix interfaces. Since these side-chains exhibit limited torsional flexibility in an α-helix (Creamer & Rose, 1992), helical faces that include residues such as valine, isoleucine and threonine provide largely preformed surfaces for helix-helix packing, reducing the entropic cost of interaction. The association of interfaces containing β-branched residues is therefore stabilized, consistent with our observation that valine, isoleucine and threonine occur frequently in strongly associating transmembrane domains.

Frequently occurring sequence patterns often indicate favorable packing motifs. The motifs identified in this study are shown in Table 2. In addition to the GxxxG motif, there is an apparent correlation between positions 6, 10 and 13 in leulib, suggesting that these positions contribute coordinately to the interface. Residues occurring at position 13 are always either threonine, or a small

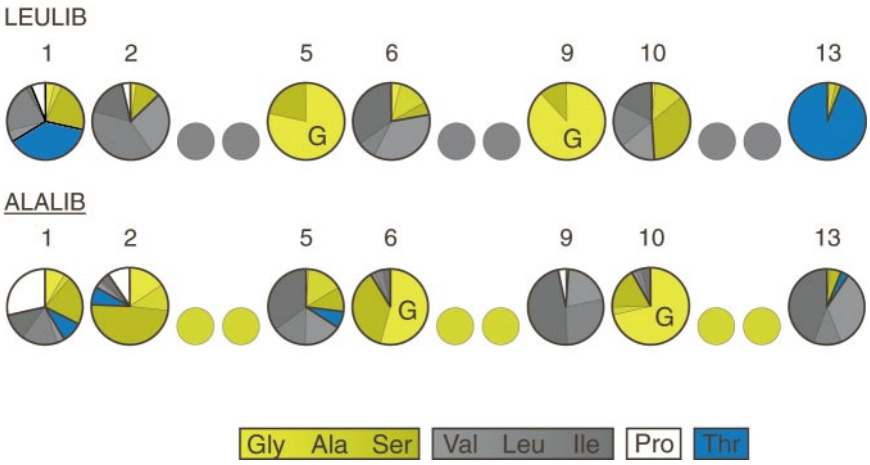

Figure 4. Distribution of amino acids occurring in high-affinity isolates. Pie charts indicate the normalized frequency of occurrence of amino acids at each of the seven variable positions in CAM-resistant transmembrane isolates for LEULIB and ALALIB. Small residues (glycine, alanine, serine) are colored shades of yellow and green; large residues (valine, leucine, isoleucine) are colored shades of gray; threonine is blue; proline is black. The order of residues in each pie is Gly, Ala, Ser, Thr, Val, Leu, Ile, Pro. Positions containing the glycine residues of the GxxxG motif are labeled G.

Table 1. Sequences of library isolates

LEULIB

GVLLGVLLGLLLGL	GVLLGLLLGVLLTL	GPLLGGLLGGLLAL	SLLLGVLLGLLLAL
AGLLGALLGSLLTL	VLLLGVLLGVLLTL	VGLLGVLLGILLAL	VLLLGILLGVLLSL
LLLLGVLLGVLLAL	LVLLGILLGLLLAL	AVLLGVLLGSLLTL	GVLLGVLLGSLLTL
LILLGALLGGLLTL	ISLLSSLLSSLLTL	VLLLGGLLGALLTL	LVLLGVLLGLLLAL
LVLLGVLLGVLLTL	SVLLGVLLGVLLTL	LLLLGALLGALLTL	LLLLGVLLGALLTL
LSLLSGLLGSLLTL	SVLLGLLLGALLTL	TILLGVLLGSLLTL	LLLLGGLLGALLTL
SILLGILLGILLTL	VLLLGVLLGVLLAL	GVLLGVLLGSLLTL	LVLLGVLLGALLTL
SLLLGVLLGLLLAL	LVLLGVLLGLLLAL	PLLLGVLLGILLTL	VLLLGILLGVLLSL
PLLLGLLLGLLLGL	ALLLGVLLGVLLAL	PGLLGLLLGALLGL	TLLLGALLGVLLTL
TVLLGVLLGLLLTL	LVLLGVLLGVLLSL	GILLGILLGILLTL	LVLLGALLGILLTL
GLLLGILLGLLLGL	SLLLGILLGLLLGL	LVLLGALLGSLLTL	LLLLGGLLGALLTL
SLLLGVLLGVLLTL	GVLLGILLGVLLTL	LLLLGVLLGLLLGL	LVLLGALLGALLTL
			LVLLGVLLGLLLGL

ALALIB

ISAAAGAALGAAIA	IGAALGAAVGAAIA	ISAAVGAALGAAVA	SGAASGAAIGAALA
PAAAIGAAIGAAVA	PSAAAGAAIGAALA	GSAAIGAAIGAAVA	AGAAAGAAIGAALA
TSAAISAAVSAAVA	GGAAVGAALGAAIA	VAAAAGAAVGAALA	SSAAAGAALGAAVA
LSAAVGAALGAAAA	PGAAVSAALGAAIA	TGAAIGAALGAAIA	SSAAIGAAVGAAIA
LGAAAGAAIGAAVA	PGAALGAAVGAAIA	GGAALGAALGAAVA	PVAAALAAGIAAGA
SSAASGAAVAAAIA	GPAAVGAALGAAVA	GSAALGAAIGAAVA	LGAALGAAVGAAVA
SSAAIGAALGAAVA	LSAAIGAAVGAAAA	IGAAAGAAIGAAVA	GSAAVGAALGAAVA
GAAAIGAALGAAVA	LGAAIGAAVGAAVA	SSAAVGAAIGAAVA	VGAALGAAIGAAIA
PGAALGAALGAAVA	LSAALGAAIGAAIA	LPAALGAAIGAALA	LVAAISAAVGAAVA
LGAAVGAAVGAAVA	LAAAVGAAIGAAIA	IGAALGAAIGAAVA	PTAAIGAAVGAAIA
SSAALGAAIGAAVA	LSAASGAAIGAAIA	VGAAVGAAIGAATA	PGAAVSAALGAAIA
LSAALGAALGAAVA	IGAAVGAAIGAAAA	ISAALGAALGAAVA	GGAAGIAAVSAALA
SSAAVGAALGAAVA	GGAAVGAALGAAIA	GTAAVGAALGAAIA	GGAATSAAIGAAIA
GAAAIGAAIGAAVA	AGAAIGAAVGAAVA	VVAAISAAVSAAVA	PGAAIGAAVGAAVA
LLAAGVAAGVAAGA	GAAALGAAVGAAIA	SSAAISAALGAAIA	TSAAISAAVSAAVA
LGAAAGAAIGAAVA	GGAAIGAAIGAAVA	SGAATGAALGAALA	LGAAIGAAVGAALA
GSAATGAALGAAIA	TTAASLAAPLAAIA	LPAAAGAAVGAALA	GGAAVGAAVGAAVA
PGAAIGAALGAAIA	GAAAVGAAIGAAVA	GIAASSAAIGAAVA	SIAAVSAALGAALA

High-affinity transmembrane isolates from LEULIB and ALALIB. The sequence of the variable region is shown, with mutable positions in bold.

residue (Gly, Ala, Ser). In 18 out of 20 cases, sequences having a small residue at position 13 have large residues (Val, Leu, Ile) at both positions 6 and 10 (G[Lg]xxG[Lg]xx[Sm]). Conversely, sequences with small residues at positions 6 and 10 generally have threonine at position 13 (eight out of nine instances, G[Sm]xxG[Sm]xxT). It seems that size complementarity is maintained at the helix-helix interface: residues at positions 6 and 10 generally have the same size. However, small residues at positions 6 and 10 provide little packing surface to stabilize the interaction. Additional stability may be gained from the packing of larger side-chains, in this case Thr. When small residues occur at position 13, the interaction must be stabilized at other positions. The inter-helical packing of large side-chains at positions 6 and 10 against the glycine residues at positions 5 and 9 may be the source of oligomeric stability.

In ALALIB, a different implementation of the GxxxG motif is evident. Although the spacing between glycine residues is preserved, the motif is shifted one residue towards the C-terminal end of the helix. As a result, the additional mutable residues available to contribute to the interface are at positions immediately before the glycine residues. These positions are most frequently occupied by large aliphatic residues, with a preference for the β-branched side-chains of Ile (positions 5 and 9) or Ile and Val (position 13).

The possibility exists that the interfaces of ALALIB isolates use both the glycine residues at 6 and 10, and the immutable alanine residues at positions 7 and 11. In this case, mutable positions 5 and 9 (and

Table 2. Comparison of sequence motifs with sequences from the SwissProt database

Pattern from library[a]	Matches in homology cleared SwissProt[b]	Related patterns in SwissProt[c]
LEULIB		
GxxxG	1641	GxxxG
GxxxGxxxT	68	GxxxGxxxT
G[Sm]xxG[Sm]xxT	5	[GAS]xxx[GAS] G[GAS]xxG GxxxG[GAS]
G[Lg]xxG[Lg]xx[Sm]	78	[VLI]xxx[VLI] G[VLI]xxG GxxxG[VLI]
ALALIB		
GxxxG	1641	GxxxG
[Lg]Gxx[Lg]Gxx[VI]	80	[VLI]Gxx[VLI] [VLI]xxx[VLI]G [VLI]GxxxG Gxx[VLI]G

[a] [Sm] indicates residues with small side-chains (Gly, Ala, Ser). [Lg] indicates residues with large side-chains (Val, Leu, Ile). Positions involved in the pattern are shown as capitals, intervening positions are labeled x.

[b] The number of times each motif appears in a homology-cleared database of 13,606 sequences derived from SwissProt (Senes *et al*., 2000).

[c] Related two- and three-residue sequence patterns identified by Senes *et al*. (2000) to occur in transmembrane domains from the SwissProt database much more frequently than expected.

possibly 13) would be away from the helix-helix interface. In the alanine library, the frequent occurrence of valine, leucine and isoleucine at these three positions may simply reflect a requirement for hydrophobic amino acid residues to compensate for the low hydrophobicity of the immutable alanine residues; however, this fails to explain the bias toward β-branched amino acids at all three positions. Also, a right-handed interface involving alanine residues 7 and 11 must use the immutable alanine residues at positions 3 and 14, resulting in a relatively featureless interface composed of four alanine and two glycine residues. Since relatively conservative mutations can disrupt the GpA dimer by removing contacts between helices (Lemmon *et al*., 1992; MacKenzie & Engelman, 1998), it seems unlikely that flat surfaces can oligomerize with substantial affinity.

Discussion

In this study, we have considered these sequences to engage in parallel, homo-oligomeric interactions formed by a right-handed crossing of straight α-helices. It should be noted that any of these sequences could be forming left-handed associations, and that the interaction surface could sometimes involve different faces of the helices (in effect, a hetero-oligomeric interface). However, initial results indicate that distinctly different sequences may predominate in the high-affinity isolates from a library designed with a left-handed interaction surface (data not shown).

It is possible that some transmembrane sequences with substantial homo-oligomerization affinity were not isolated, since poorly expressed sequences would exhibit low apparent signals. Also, little structural information is available about the ToxR cytoplasmic domain. It is possible that its functional dimerization requires a very strict geometry, and that some transmembrane interaction crossing angles promote more functional ToxR dimerization than do others. This seems unlikely, however, in the light of several transcriptionally active ToxR chimerae that have been constructed (Brosig & Langosch, 1998; Kolmar *et al*., 1994, 1995; Ottemann & Mekalanos, 1995).

The packing motifs identified in this study are shown in Table 2. These sequence patterns were used to search a database of 13,606 non-homologous transmembrane domains derived from SwissProt (Senes *et al*., 2000). The number of sequences matching these motifs is shown. The large number of matches to the motifs suggests that natural proteins frequently utilize these packing surfaces. It is interesting to note, however, that the majority of the SwissProt transmembrane domains occur in polytopic membrane proteins, and therefore more frequently engage in anti-parallel interactions with non-identical helices. A more rigorous analysis of sequence patterns in transmembrane domains (see the accompanying paper, Senes *et al*., 2000) has identified similar motifs as occurring with much greater than expected frequency. In the database analysis, GxxxG is the most highly biased sequence motif in naturally occurring transmembrane domains. Also, the GxxxG pair occurs frequently in conjunction with large (Val, Leu, Ile) residues at positions neighboring the glycine residues. A similar relationship exists between GxxxG and small residues (Gly, Ala, Ser). Since the SwissProt database analysis was limited to two and three amino acid correlations, related sequence patterns identified by Senes *et al*. (2000) are shown in Table 2. The convergence of these independent experimental approaches supports the hypothesis that the helix-helix interfaces isolated from our library represent naturally occurring packing motifs.

The frequent appearance of GxxxG in the library isolates suggests that this may represent a general sequence motif involved in high-affinity association of transmembrane helices. Analysis of sequence databases has revealed that pairs of glycine residues are most frequently separated by three residues in transmembrane domains (Arkin & Brunger, 1998; Senes *et al*., 2000). Furthermore, this motif occurs in hundreds of non-homologous transmembrane domains (Table 2) and in thousands of homologous sequences (not shown). Our results suggest that these transmembrane domains may often be involved in packing interactions with other transmembrane α-helices. The observation by Brosig and Langosch that in a polymethionine background, GxxxG is sufficient to mediate dimerization (Brosig & Langosch, 1998) supports the idea that sequences containing this motif are likely to

have substantial oligomerization potential. However, the possibility that GxxxG is sufficient for oligomerization would suggest that these hundreds of transmembrane sequences (many in the same membrane) can interact with one another. Such promiscuous transmembrane domain association seems implausible, and it is probable that additional specificity elements must be employed.

The specific use of glycine in this sequence motif indicates that the fundamental properties of this amino acid are being exploited. An investigation of the glycine residues in the GpA transmembrane domain NMR structure (MacKenzie *et al.*, 1997; and see Figure 5) suggests that these glycine residues effect dimerization by (i) providing a surface for packing, (ii) permitting helix proximity and (iii) by entropic effects. The lack of side-chains at positions G79 and G83 produces a flat surface against which the side-chains of other interfacial residues pack. This surface is clearly critical, since mutation of either of these residues generally results in a strong destabilization of dimerization (Bu & Engelman, 1999; Langosch *et al.*, 1996; Lemmon *et al.*, 1992; Russ & Engelman, 1999). In addition, glycine residues allow the two helices to come into close proximity, thereby facilitating intimate contact between the other interfacial side-chains. Finally, a consequence of dimerization is that, although stabilized by van der Waals packing, the loss of accessible rotamer states for the interacting side-chains destabilizes the dimer. This decrease in side-chain entropy is thought to play a pivotal role in the disruption of GpA dimerization by certain mutations, partially mitigating the stabilizing effects of favorable van der Waals contacts (MacKenzie & Engelman, 1998). The lack of side-chain atoms in glycine residues results in no loss of side-chain entropy upon dimerization at these positions, a less destabilizing effect relative to most other amino acids.

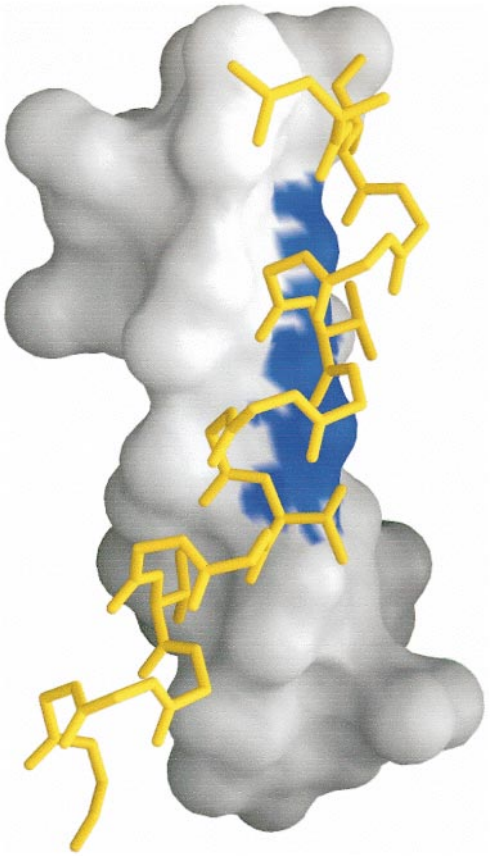

Figure 5. The dimerization interface of the GpA transmembrane peptide exhibits a GxxxG motif. The glycine residues 79 and 83 are colored blue on the surface representation of one monomer. The backbone of the second monomer is shown, with side-chains drawn for residues involved in the dimerization motif. Residues 75-87 of PDB entry 1afo are shown. This Figure was made using GRASP (Nicholls *et al.*, 1991).

While a pair of glycine residues separated by three amino acid residues is an almost invariant feature of the transmembrane domains isolated in this study, residues at the other interfacial positions varied substantially. In the GpA transmembrane domain, mutation of these non-glycine interfacial residues results in a range of phenotypes, caused by either loss of side-chain rotamer entropy, steric clash, or loss of favorable van der Waals contacts (MacKenzie & Engelman, 1998). Residues Leu75, Ile76, Val80, Val84 and Thr87 present two complementary ridges that pack against one another in the dimer (MacKenzie *et al.*, 1997). The GxxxG-containing sequence variants isolated in the library screen most likely present alternative ridge topologies that promote specific associations by defining the geometry of the dimer interface.

Considered in this context, the GxxxG sequence motif can be regarded as a framework for the dimerization of transmembrane α-helices. The flat surface provided by the glycine residues can serve as a template for many possible combinations of side-chains at other interfacial positions that act as determinants of specificity. As a result, the wide variety of specific interactions that can form around GxxxG motifs should give rise to a broad range of interaction energies, as well as enabling many such domains to co-exist in the same membrane and still engage in specific interactions. An additional element of specificity may arise from the positioning of the GxxxG motif within the transmembrane domain, relative to the depth of the bilayer. Transmembrane helices with GxxxG motifs located near the surface of the bilayer would be unlikely to interact with other helices presenting GxxxG motifs near the center. Transmembrane helix orientation may also factor into specificity, since it is possible that some dimerization domains describing parallel interaction interfaces may not promote anti-parallel interactions as well. As such, two proteins featuring the same transmembrane sequence could promote distinct specific interactions by presenting their association motifs in opposite orientations. While it is almost certain that other right-handed oligomerization

motifs exist, the infrequent appearance of other transmembrane sequences in libraries selected at high CAM concentration suggests that the GxxxG-mediated interaction is a dominant theme.

Materials and Methods

Library cloning

Oligonucleotides were synthesized encoding the leucine (acacaccgcaggctagcVBtVBtctcttaVBtVBtttgcttVBtVBtctattaVBtctgatcgccctaacggatatc) and alanine (aacacaccgcaggctagcVBtVBtgctgcaVBtVBtgctgcaVBtVBtgctgcaVBtgcgatcgccctaacggatatc) libraries, where V and B refer to equimolar mixtures of (a, g, c) and (g, c, t), respectively, coupled at a given position. The *Nhe*I and *Dpn*II sites used for cloning are underlined. The resulting codons specify the amino acids Gly, Ala, Val, Leu, Ile, Ser, Thr, Pro and Arg. Double-stranded library inserts were generated by Klenow-catalyzed extension of a short primer (AS1) annealed to the library oligonucleotide. The resulting double-stranded product was digested with *Nhe*I and *Dpn*II, dephosphorylated using alkaline phosphatase (NEB), and gel-purified. pccKAN (Russ & Engelman, 1999) was digested with *Nsi*I, dephosphorylated with alkaline phosphatase, digested with *Nhe*I and *Bam*HI, and isolated by gel purification. The prepared vector and insert were ligated at a molar ratio of 8:1 (vector to insert), at a concentration of 50 ng vector/µl using 20 units of phage T4 DNA ligase/µl at 16 °C for ~20 hours, followed by precipitation in ethanol and transformation of *E. coli* by electroporation. A control ligation containing no insert was used to gauge the efficiency of ligation. Dilutions of the transformation were plated to estimate the total number of transformants. The remaining transformation culture was plated on 28 LB/AMP library plates (150 × 15 mm style, Falcon 1058). Following an overnight incubation at 37 °C, the near-confluent lawn of bacteria was resuspended from the library plates in LB, brought to 15 % (v/v) glycerol, and frozen in 200 µl aliquots at −80 °C. To determine the titer of transformed cells in the library glycerol stock, an aliquot was thawed on ice and dilutions were plated on LB/AMP.

Library selection

The titered library glycerol stock was diluted between 100 and 500-fold (to deliver twice the number of sequences expected in the library, 10^7) and plated on LB/AMP plates with varying concentrations of CAM. At low concentrations of CAM, colonies were counted using further dilutions of the stock. The most resistant colonies, representing ~0.001 % of the total sequences, were used to inoculate PCR reactions for sequencing. Isolates were screened for membrane insertion using a *malE* complementation assay (Kolmar *et al*., 1995). Alternatively, the library glycerol stocks were plated on M9-maltose containing Amp and CAM to simultaneously select for sequences that were both transmembrane and oligomeric. Disk diffusion assays were performed as described (Russ & Engelman, 1999). Immunoblotting was carried out using standard methods and detected using the ECL kit (Amersham). Anti-MBP antibodies (NEB) were diluted 1:10,000 to 1:20,000; HRP conjugated goat anti-rabbit (Pierce) were diluted 1:20,000.

Strains

E. coli strains DH5α, or strains MM39 (*araD lac*Δ*U1269, malE*Δ*444, str*^{R}; kindly supplied by J. Beckwith, Harvard University Medical School) and NT326 (*F*−(*argF-lac*)*U169*, *rpsL150*, *relA1*, *rbsR*, *flbB5301*, *ptsF25*, *thi-1*, *deoC1*, Δ*malE444*, *recA*, *srlA*$^+$; kindly supplied by H. Shuman, Columbia University) were used for CAM selection.

Library sequencing

The transmembrane region was PCR amplified, purified using the QIAQUICK PCR purification kit (QIAGEN), and sequenced by the W.M. Keck facility (New Haven, CT).

Sequence analysis

Statistical significance (p) was calculated using the binomial distribution, and represents the probability that the observed number of events could have occurred by chance. To account for an observed bias in the composition of the degenerate oligonucleotides, the expected frequency of occurrence for individual amino acids in the entire library was calculated from a database of sequences selected for membrane insertion (using *malE* complementation), but not for oligomerization (0 mg/ml CAM). Isolates that contained arginine were suspected to be non-transmembrane, and were discarded from the analysis. The expected probability of occurrence for the GxxxG pair in the absence of bias was taken as the product of the individual probability for G residues in the entire library. The normalized frequencies reported in Figure 4 were also corrected to account for oligonucleotide synthesis bias and membrane insertion.

Acknowledgements

We thank Drs J. Beckwith and H. Shuman for generous donation of *E. coli* strains, A. Senes, L. Hanakahi, K. Sonoda and K. MacKenzie for intellectual contribution, the Engelman laboratory for critical review of this manuscript, and the NIH, NSF and National Foundation for Cancer Research for funding.

References

Adams, P. D., Engelman, D. M. & Brunger, A. T. (1996). Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. (Published erratum appears in *Proteins: Struct. Funct. Genet.* 1997, **27**, 132). *Proteins: Struct. Funct. Genet.* **26**, 257-261.

Adams, P. D., Lee, A. S., Brunger, A. T. & Engelman, D. M. (1998). Models for the transmembrane region of the phospholamban pentamer: which is correct? *Ann. NY Acad. Sci.* **853**, 178-185.

Arkin, I. T. & Brunger, A. T. (1998). Statistical analysis of predicted transmembrane alpha-helices. *Biochim. Biophys. Acta,* **1429**, 113-128.

Arkin, I. T., Adams, P. D., MacKenzie, K. R., Lemmon, M. A., Brunger, A. T. & Engelman, D. M. (1994). Structural organization of the pentameric transmembrane alpha-helices of phospholamban, a cardiac ion channel. *EMBO J.* **13**, 4757-4764.

Bargmann, C. I., Hung, M. C. & Weinberg, R. A. (1986). Multiple independent activations of the neu oncogene by a point mutation altering the transmembrane domain of p185. *Cell,* **45**, 649-657.

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

Bu, Z. & Engelman, D. M. (1999). A method for determining transmembrane helix association and orientation in detergent micelles using small angle X-ray scattering. *Biophys. J.* 1064-1073.

Creamer, T. P. & Rose, G. D. (1992). Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc. Natl Acad. Sci. USA,* **89**, 5937-5941.

Engelman, D. M. & Steitz, T. A. (1981). The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. *Cell,* **23**, 411-422.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994). A mutation data matrix for transmembrane proteins. *FEBS Letters,* **339**, 269-275.

Kolmar, H., Frisch, C., Kleemann, G., Gotze, K., Stevens, F. J. & Fritz, H. J. (1994). 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**, 61-70.

Kolmar, H., Hennecke, F., Gotze, K., Janzer, B., Vogt, B., Mayer, F. & Fritz, H. J. (1995). Membrane insertion of the bacterial signal transduction protein ToxR and requirements of transcription activation studied by modular replacement of different protein substructures. *EMBO J.* **14**, 3895-3904.

Langosch, D., Brosig, B., Kolmar, H. & Fritz, H. J. (1996). Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. *J. Mol. Biol.* **263**, 525-530.

Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J. & Engelman, D. M. (1992). Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry,* **31**, 12719-12725.

Lemmon, M. A., Treutlein, H. R., Adams, P. D., Brunger, A. T. & Engelman, D. M. (1994). A dimerization motif for transmembrane alpha-helices. *Nature Struct. Biol.* **1**, 157-163.

Machamer, C. E., Grim, M. G., Esquela, A., Chung, S. W., Rolls, M., Ryan, K. & Swift, A. M. (1993). Retention of a cis Golgi protein requires polar residues on one face of a predicted alpha-helix in the transmembrane domain. *Mol. Biol. Cell,* **4**, 695-704.

MacKenzie, K. R. & Engelman, D. M. (1998). Structure-based prediction of the stability of transmembrane helix- helix interactions: the sequence dependence of glycophorin A dimerization. *Proc. Natl Acad. Sci. USA,* **95**, 3583-3590.

MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science,* **276**, 131-133.

Manolios, N., Bonifacino, J. S. & Klausner, R. D. (1990). Transmembrane helical interactions and the assembly of the T cell receptor complex. *Science,* **249**, 274-277.

Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281-296.

Ottemann, K. M. & Mekalanos, J. J. (1995). Analysis of *Vibrio cholerae* ToxR function by construction of novel fusion proteins. *Mol. Microbiol.* **15**, 719-731.

Popot, J. L. & Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry,* **29**, 4031-4037.

Russ, W. P. & Engelman, D. M. (1999). TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc. Natl Acad. Sci. USA,* **96**, 863-868.

Senes, A., Gerstein, M. & Engelman, D. M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with β-branched residues at neighboring positions. *J. Mol. Biol.* **296**, 921-936.

Simmerman, H. K. & Jones, L. R. (1998). Phospholamban: protein structure, mechanism of action, and role in cardiac function. *Physiol. Rev.* **78**, 921-947.

Smith, S. O., Smith, C. S. & Bormann, B. J. (1996). Strong hydrogen bonding interactions involving a buried glutamic acid in the transmembrane sequence of the neu/erbB-2 receptor. *Nature Struct. Biol.* **3**, 252-258.

Zhou, F. X., Cocco, M. J., Russ, W. P., Brunger, A. T. & Engelman, D. M. (2000). Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nature Struct. Biol.* In the press.

Edited by G. von Heijne

(Received 10 November 1999; received in revised form 29 December 1999; accepted 29 December 1999)