Supplementary Information

Title: A topological and conformational stability alphabet for multi-pass membrane proteins

Authors: Feng, X.[1] & Barth, P.[1,2,3] ¶

¶ Correspondences should be addressed to: P.B. (patrickb@bcm.edu)

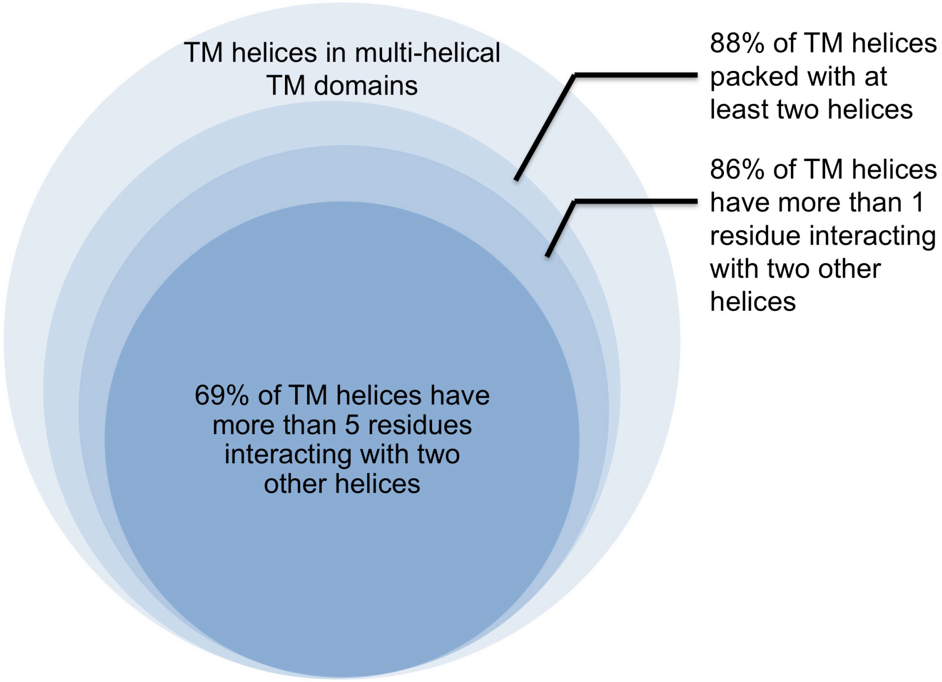
[1] Department of Pharmacology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

[2] Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

[3] Structural and Computational Biology and Molecular Biophysics Graduate Program, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

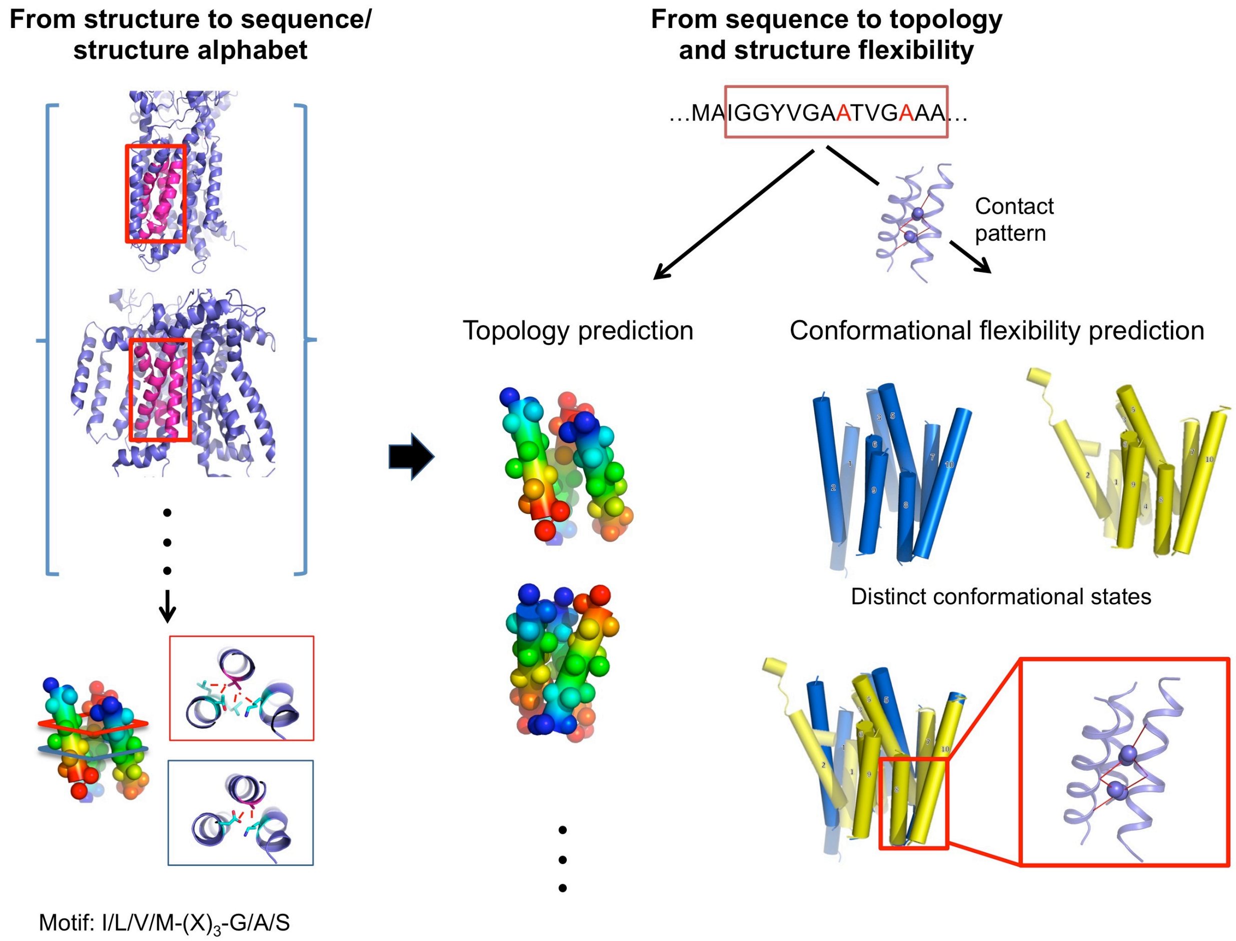
Supplementary Results

Supplementary Figure 1. A large majority of TMHs in multi-pass membrane proteins form binding interfaces with two other helices. Venn Diagram describing the fractions of TMHs in multi-pass membrane proteins: 1) interacting with two other helices, 2) interacting with two other helices with more than one residue forming contacts with the 2 helices simultaneously, 3) interacting with two other helices with more than five residues forming contacts with the 2 helices simultaneously. Contacts are defined by residues pairs with Cα distance less than 9Å.

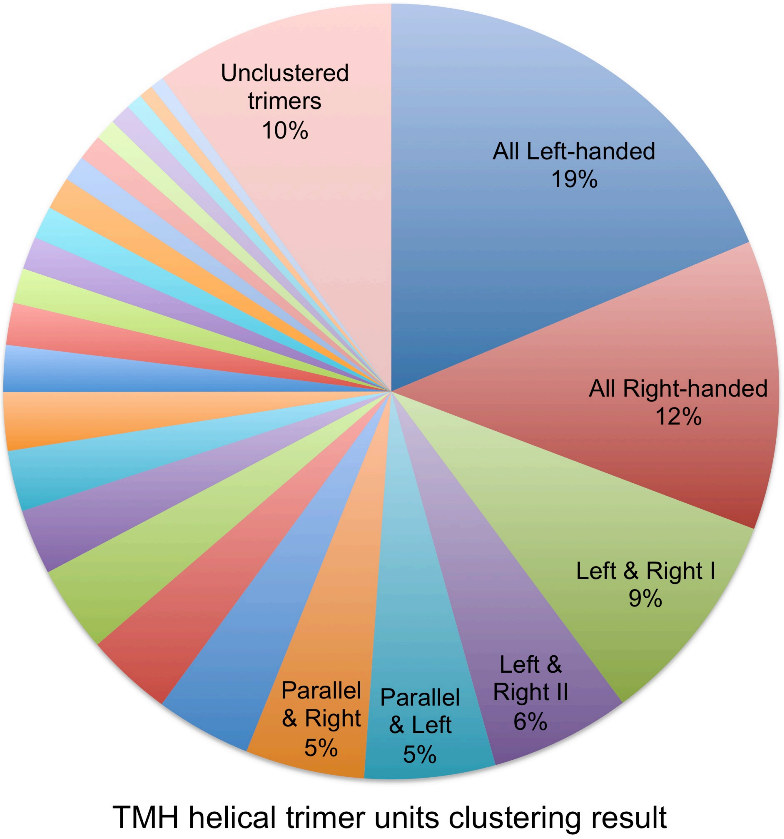

TM helices in multi-helical TM domains
88% of TM helices packed with at least two helices
86% of TM helices have more than 1 residue interacting with two other helices
69% of TM helices have more than 5 residues interacting with two other helices

Helical packing in multi-helical TM domains

Supplementary Figure 2. Uncovering universal sequence/structure principles governing multi-pass membrane protein topology and structure flexibility. a. The panel describes the process of identifying consensus sequence/structure motifs from the protein structure database. Elementary interacting TMH trimer units are extracted from unrelated protein structures, clustered into structurally similar families. Combinations of residues enriched within each trimer family creating consensus networks of stabilizing interhelical contacts are identified. **b.** The panel describes the stringent validation of the sequence/structure determinants of TMH trimer packing. If the sequence motifs are strong determinants of trimer conformations, prediction of trimer topology from sequence should be feasible. If motifs and associated interhelical contacts are strong determinants of trimer stability, they should guide the prediction of local conformational flexibility in TM proteins.


From structure to sequence/ structure alphabet
From sequence to topology and structure flexibility
...MAIGGYVGAATVGAAA...
Contact pattern
Topology prediction
Conformational flexibility prediction
Distinct conformational states
Motif: I/L/V/M-$(X)_3$-G/A/S

Supplementary Figure 3. Pi chart describing the clustering of TMH trimers into structurally similar families based on Cα RMSD. 56% of the trimer unit library can be classified into only 6 well-defined structural classes. The topology and percentage of each trimer type is labeled in the chart.



Supplementary Table 1. The geometry of the 6 most populated trimer structure clusters. The table describes the geometrical features of the 3 helical pairs constituting each trimer type. Helices are numbered (second column) according to a reference trimer structure in each class to assign helix pairs with specific topology. The same helix numbers are used to assign enriched motifs to specific helices in main text Figure 1 and Supplementary Figure 2.

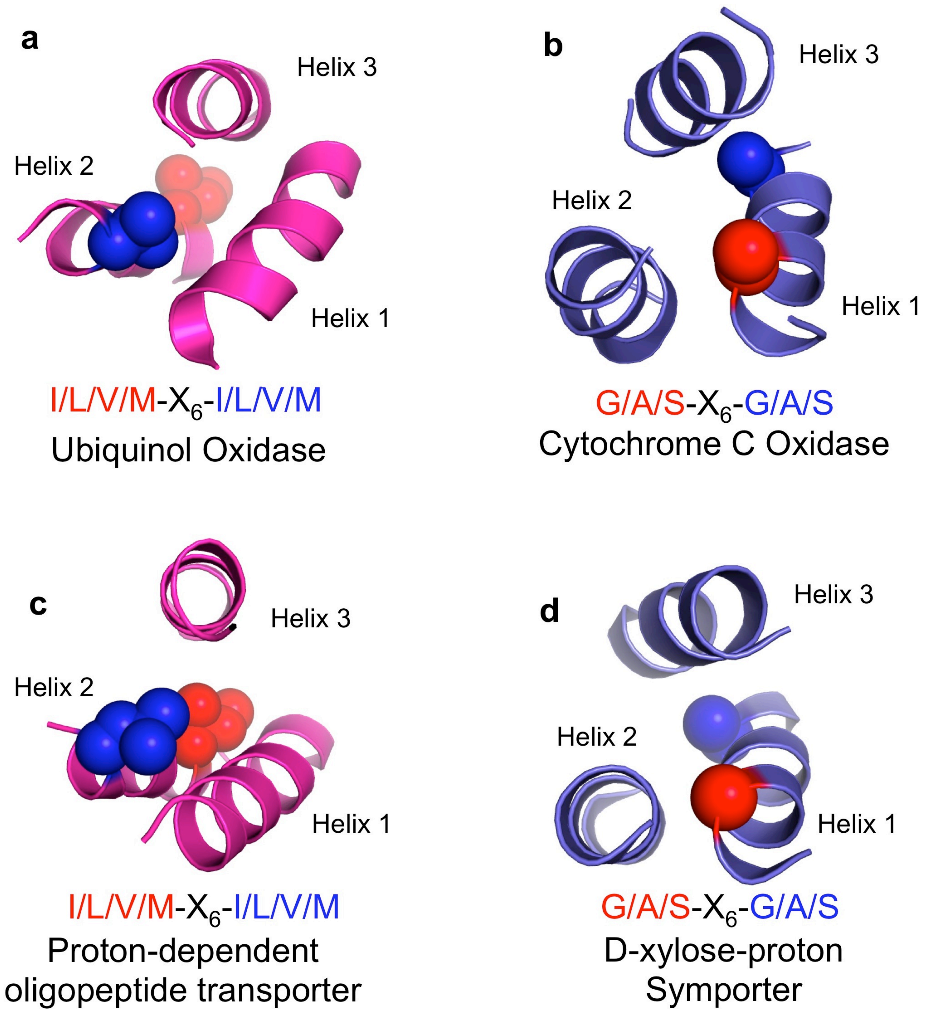
			Inter-helical angle			Inter-helical distance	
Trimer type Designation	pairs	handedness	Average angle (degree)	standard deviation	Topology: parallel/anti-parallel	Average distance (Å)	standard deviation
All-left	1-2	left	155.4	14.0	Anti-parallel	10.0	1.7
	2-3	left	29.1	14.8	Parallel	11.1	1.6
	1-3	left	154.0	10.7	Anti-parallel	10.5	2.0
All-right	1-2	right	-37.1	10.1	Parallel	8.8	1.8
	2-3	right	-145.2	10.8	Anti-parallel	9.5	1.5
	1-3	right	-144.6	11.3	Anti-parallel	9.4	2.0
Left & Right I	2-3	right	-41.0	11.7	Parallel	8.0	1.8
	1-3	left	155.8	10.6	Anti-parallel	9.6	1.8
Left & Right II	2-3	left	155.9	10.4	Anti-parallel	8.9	1.2
	1-3	right	-152.1	10.6	Anti-parallel	9.4	1.3
Parallel & left	1-2	left	157.8	8.1	Anti-parallel	9.0	1.4
	2-3	left	161.3	10.0	Anti-parallel	8.9	1.8
	1-3	small inter-angle	13.9	7.8	Parallel	10.9	0.9
Parallel & right	1-2	right	-147.6	10.0	Anti-parallel	9.9	1.6
	1-3	right	-150.1	9.6	Anti-parallel	9.6	1.1
	2-3	small inter-angle	-18.5	11.1	Parallel	11.6	1.2

Supplementary Table 2: Table of enriched sequence motifs at TMH trimer interfaces. For each trimer structure class (i.e. cluster), the p-value of enrichment of a motif on a specific helix is calculated using TMSTAT (see methods) and reported for two homology reduction thresholds: 60% sequence identity (60SI) removing close homologs for the structural analysis and one protein per superfamily (1/superfamily), a stringent homology reduction for training and testing the topology predictor (see methods). Motifs are classified as “major” if they form consensus contacts at trimer interfaces and are found in at least 3 protein families and superfamilies (see methods). P-values highlighted in red correspond to motifs that loose significant enrichment (p-value ≥ 0.05) in the dataset composed of only one protein per superfamily.

		60SI	1/superfamily
Cluster 1	**All Left-handed**	P-value	P-value
assigned helix	Major		
1	I/L/V/M-X_3-G/A/S	1.60E-03	4.80E-02
1	I/L/V/M-X_6-I/L/V/M	1.40E-03	1.10E-01
2	I/L/V/M-X_6-I/L/V/M	4.00E-04	4.60E-02
1	G/A/S-X_6-G/A/S	9.25E-05	4.00E-04
3	I/L/V/M-X_6-F/W/Y	2.60E-02	1.09E-01
	Minor		
1	I/L/V/M-X_4-I/L/V/M	2.60E-03	1.36E-02
2	I/L/V/M-X_4-I/L/V/M	1.00E-04	3.80E-03
3	G/A/S-X_3-I/L/V/M	1.40E-02	2.50E-02
1	P-X_6-F/W/Y	2.80E-03	1.20E-02
3	I/L/V/M-X_2-I/L/V/M	4.75E-05	8.00E-02
2	T/S-X_3-I/L/V/M	6.00E-04	1.12E-01
3	T/S-X_6-I/L/V/M	1.87E-02	1.90E-01
Cluster 2	**All Right-handed**		
	Major		
1	G/A/S-X_3-G/A/S	3.78E-08	1.46E-05
2	G/A/S-X_3-G/A/S	3.56E-09	7.52E-05
3	G/A/S-X_3-G/A/S	1.42E-09	8.53E-08
2	G/A/S-X_7-G/A/S	3.00E-03	4.10E-03
	Minor		
3	T/S-X_3-T/S	2.00E-02	8.60E-02

		60SI	1/superfamily
Cluster 3	**Left & Right handed**	P-value	P-value
assigned helix	Major		
1	F/W/Y(H)-X_3-G/A/S	1.20E-03	1.40E-02
3	F/W/Y-X_6-G/A/S	5.80E-03	4.50E-03
	Minor		
2	F/W/Y-F/W/Y	3.10E-02	4.40E-02
2	F/W/Y-X_6-T/S	5.90E-03	5.10E-02
Cluster 4	**Left & Right handed II**		
	Major		
1	F/W/Y-X_6-G/A/S	2.90E-03	5.00E-02
3	F/W/Y-X_6-G/A/S	4.69E-06	4.70E-02
Cluster 5	**Parallel & left handed**		
	Major		
1	G/A/S-X_2-F/W/Y(M)	8.00E-03	3.50E-02
1	I/L/V/M-X_6-I/L/V/M	5.00E-02	2.00E-01
	Minor		
2	D/E/N/Q-X_3-F/W/Y	5.20E-03	8.00E-03
1	F/W/Y-F/W/Y	3.00E-02	3.50E-02
Cluster 6	**Parallel & Right handed**		
	Major		
3	I/L/V/M-X_3-F/W/Y	8.20E-03	1.67E-02
3	F/W/Y-X_2-F/W/Y	3.50E-03	5.60E-03
	Minor		
3	F/W/Y-X_3-F/W/Y	2.50E-03	1.00E-01
1	T/S-X_3-I/L/V/M	1.60E-02	1.70E-01

Supplementary Figure 4. Proteins from the same superfamily display different sequence/structure determinants of the same trimer topology. Four examples of trimers from the all-left handed trimer structure class. **a,b.** Trimers are from two proteins of the Cytochrome c oxidase subunit I-like superfamily. These proteins share more than 30% sequence identity but bear distinct enriched sequence motif at the same trimer interface. **c,d.** Trimers are from two proteins of the MFS general substrate transporter superfamily but bear distinct enriched sequence motif at the same trimer interface. Trimers across protein superfamilies (**a**,**c** and **b**,**d**) share the same motifs.



I/L/V/M-X$_6$-I/L/V/M
Ubiquinol Oxidase

G/A/S-X$_6$-G/A/S
Cytochrome C Oxidase

I/L/V/M-X$_6$-I/L/V/M
Proton-dependent oligopeptide transporter

G/A/S-X$_6$-G/A/S
D-xylose-proton Symporter

Supplementary Table 3: Summary of consensus contact map and number of interacting helices for each trimer motif. The table describes the number of consensus contacts and the number of helices interacting with each position of the motif. A contact between one position of a motif and a residue on adjacent helices is defined as consensus if it is found in more than 50% of the trimers from the same structure class sharing the same sequence motif. For comparison, the average number of contacts established by the same position of the motif in all trimers sharing the same sequence motif is provided in parentheses. * Threonine is also observed at both positions. + Methionine is also observed in the second position, # Histidine is also observed in the first position.

	1st Position			2nd position		
	Number of consensus contacts	**Average number of contacts**	**Number of interacting helix(ces)**	**Number of consensus contacts**	**Average number of contacts**	**Number of interacting helix(ces)**
All Left Handed						
G/A/S-X_6-G/A/S	4	3.5	2	4	3.7	2
I/L/V/M-X_3-G/A/S	5	4.8	2	3	3.7	2
I/L/V/M-X_6-I/L/V/M	5	4.9	2	5	5.6	2
I/L/V/M-X_6-F/W/Y	5	5.7	2	6	7.8	2
All Right Handed						
G/A/S-X_3-G/A/S	4	4.4	2	4	4.2	2
G/A/S-X_7-G/A/S*	4	4.5	2	4	3.5	2
Right & Left 1						
F/W/Y-X_3-G/A/S#	4	4.1	2	2	2.5	1
F/W/Y-X_6-G/A/S	5	5.6	2	3	4.0	2
Right & Left 2						
F/W/Y-X_6-G/A/S	4	4.8	2	2	3.8	1 or 2
Parallel & Left						
G/A/S-X_2-F/W/Y+	3	2.8	1	8	8.3	2
I/L/V/M-X_6-I/L/V/M	4	5.1	2	4	5.7	2
Parallel & Right						
I/L/V/M-X_3-F/W/Y	3	3.1	2	3	3.7	2
F/W/Y-X_2-F/W/Y	3	5.7	2	3	5.1	2

Supplementary Table 4. Support Vector Classification (SVC) of trimer sequences into structures. The table reports the accuracy in correctly assigning trimer sequences into one of two possible trimer structure classes. The accuracy of classification is calculated using 5-fold cross validation (see method). Results are given for all possible 15 pairs of trimer structure classes. Random assignment would correspond to 50% accuracy. Below the table, the average accuracy for two-classes assignment is reported. * Average accuracy for two classes assignment is significantly higher than random assignment with P-value < 0.0001 using Welch's t-test. The accuracy in correctly assigning trimer sequences into one of six possible trimer structure classes (6-classes assignment) is reported below the table. Random assignment would correspond to 16.7% accuracy.

	All Right-handed	Parallel & Left	Parallel & Right	Right & Left I	Right & Left II
All Left-handed	78.5%	83.5%	79.2%	65.9%	74.2%
All Right-handed		81.7%	75.0%	60.5%	76.1%
Parallel & Left			70.8%	76.9%	70.4%
Parallel & Right				71.6%	68.5%
Right & Left I					67.1%
Accuracy for six classes assignment: 41.2%					
Average accuracy for two classes assignment: 73.0% *					

Supplementary Table 5. Trimers bearing sequence/contact motifs are less flexible. The left part of the Supplementary Table displays the name of the proteins and the PDB codes of different protein conformations (states) that differ by at least 0.5Å in Cα RMSD. The right part of the table indicates the extent of conformational changes of each trimer unit in the protein measured by Cα RMSD (Å). The trimer unit is defined by the trimer type (i.e. topology) and motif type. The topology is encoded as follows: AL: All left-handed Trimer; AR, All Right-handed Trimer; PL: Parallel & Left handed Trimer; PR: Parallel & Right handed Trimer; LR1: Left & Right Type 1 Trimer; LR2: Left & Right Type 2 Trimer. Motif types are encoded as follows: NA: no sequence/structure motif found in the trimer; (AL, M1): I/L/V/M-X_3-G/A/S; (AL, M2): I/L/V/M-X_6-I/L/V/M; (AL, M3): G/A/S-X_6-G/A/S; (AL, M4): I/L/V/M-X_6-F/W/Y; (AR, M1): G/A/S-X_3-G/A/S; (AR, M2): G/A/S-X_7-G/A/S; (PR, M1): I/L/V/M-X_3-F/W/Y; (PR, M2): F/W/Y-X_2-F/W/Y; (LR1, M1): F/W/Y-X_3-G/A/S; (LR1, M2): F/W/Y-X_6-G/A/S; (LR2, M1): F/W/Y-X_6-G/A/S.

Proteins	State1 description/PDB code		State2 description/PDB code		Entire TM domain Cα RMSD (Å)	Trimer Cα RMSD(Å) (Trimer Type, Motif)			
β2 adrenergic receptor	inactive	2RH1	active	3P0G	1.9	0.4 (AL,M1)	1.2(AL, NA)	0.6(LR2,M1)	
Adenosine A2A receptor	inactive	3EML	active	2YDO	1.7	1.5 (AL, NA)	1.4 (AL, M1)	0.9(LR2, M1)	
MalFGK2-MBP Maltos uptake transporter	pre-translocan	4KHZ	out-ward facing	4KI0	1.2	0.3 (AR, M1, M2)	0.3(AR, NA)	0.4(LR1)	
ABCB10 Mitochondrial ABC transporter	Rod Form B	4AYX	Plate form	4AYW	1.1	0.4(LR1, M1)			
ABCB10 Mitochondrial ABC transporter	Rod Form A	4AYT	Plate form	4AYW	1.2	0.4(LR1, M1)			
ABCB10 Mitochondrial ABC transporter	Rod Form A	4AYT	Rod Form B	4AYX	0.8	0.2(LR1, M1)			
YetJ pH-sensitive calcium-leak channel,	open form	4PGS	close	4PGR	0.5	3.2 (PL, NA)			
Leucine Transporter	inward open	3TT1	3TT2	3TT3	2.6	0.8(AL, M2)	1.4(AL, NA)		
bile acid transporter	inward open	4N7W	Outward open	4N7X	3.7	2.7(AL, NA)	0.4(AR, M1)	2.6(AL, M3)	2.3 (LR1, NA)
AdiC Arginine: Agmatine Antiporter E coli	outward-facing	3L1L	Open-to-out conformation	3OB6	2.1	0.8 (AR, M1,M2) 0.7(AR, NA)	2.7(PL, NA) 0.5 (PR, M1)	1.0(LR1,NA)	0.9(AL, M3)
Calcium ATPase	E1 state	1SU4	E2 state	1WPG	6.1	2.1 (PR, NA) 4.3 (LR2, NA)	1.4 (PR, NA)	0.9 (AR, M1)	0.7(AR, M1)

AcrB bacterial multi-drug efflux transporter	MonomerA/B(T)	2GIF	monomerC(L)	2DHH	2.0	1.1 (AR, M1, M2)
						1.6 (AR, NA)
							0.9(AR, M1, M2)
							1.4(PR, M2)
								0.9 (LR1, M1)
								1.3(LR2, M1)
									1.8(PR, NA)
									1.4(LR2, NA)

LacY Lactose Permease	pH 6.5, no substrate	2CFQ	Occluded, partially open to periplasmic side	4OAA	4.8	0.5 (AL, M2) 0.7 (LR1, M2)	0.6 (AL, M1)	1.7(LR1, NA)	2.2 (LR1,NA)
vSGLT Sodium Galactose Transporter	inward-occluded	3DH4	inward-open	2XQ2	1.2	1.3 (AL, NA)	0.6(AL, M2)	1.0(LR1, NA)	0.3 (AR, M1, M2)
Mhp1 Benzyl-hydantoin transporter	outward-facing	2JLN	inward-facing	2X79	3.2	2.8(LR1,NA)	0.5 (AR, NA)	0.4(AR, M1, M2)	0.9(AR, M1)
BetP glycine betaine transporter	inward intermedia	2WIT	outward-facing	4DOJ	2.2	2.2 (AL, NA)	1.1 (AL, NA)	2.8 (LR1, M2)	
XylE proton:xylose symporter	normal	4GBY	inward-facing open	4QIQ	3.0	0.7(AR, M1) 2.4 (LR1, NA)	0.9(AL, M2) 1.4 (AL, NA)	1.0(LR1, NA)	0.5 (LR1, NA)
Glutamate Transporter	inward-facing	4P19	outward-facing	1XFH	9.4	1.0 (AR, M1)	1.2(AR,M1, M2)	7.5 (LR1, NA)	2.3 (AR, NA)
GlpG Rhomboid protease	conformation 1	2XOV	conformation 2	2NRF	2.7	0.2 (AR, M1, M2)			

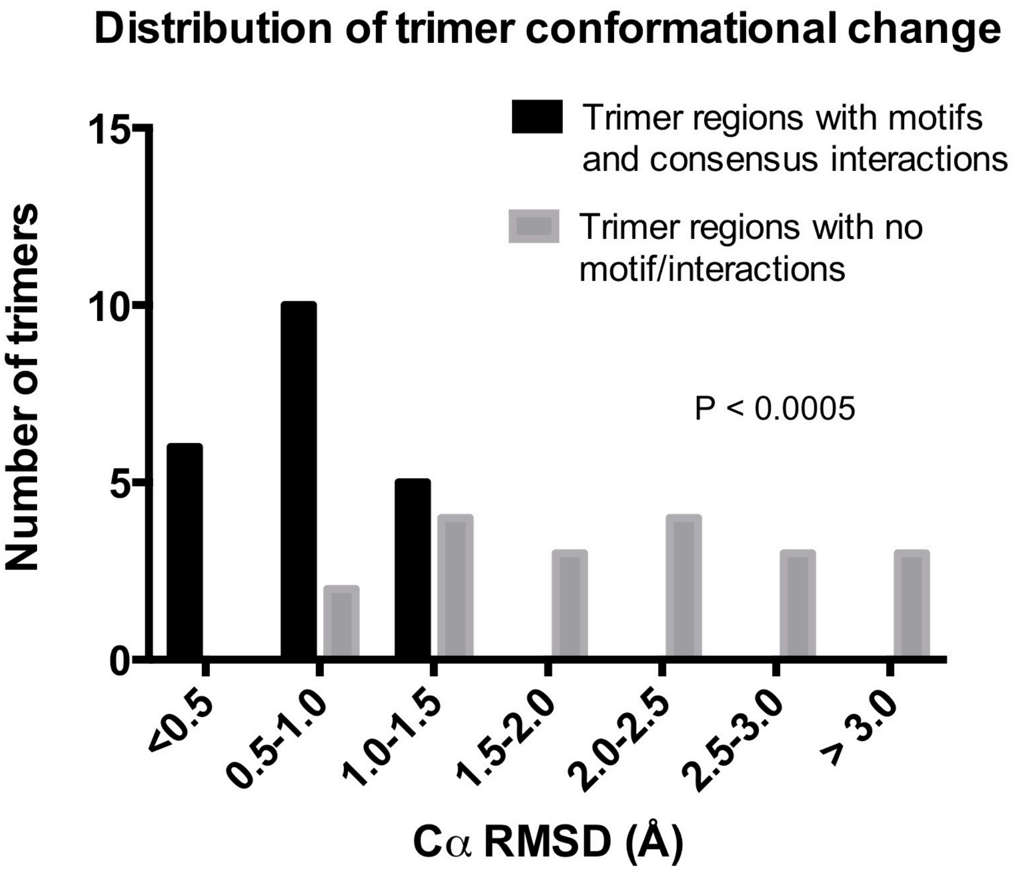
Supplementary Table 6. Trimers extracted only from distinct protein superfamilies bearing sequence/contact motifs are less flexible. Compared to Main Text Figure 6 and Supplementary Table 5, only one protein per superfamily was selected leading to 10 protein structures and 40 trimer units. The left part of the Supplementary Table displays the name of the proteins and the PDB codes of different protein conformations (states) that differ by at least 0.5Å in Cα RMSD. The right part of the table indicates the extent of conformational changes of each trimer unit in the protein measured by Cα RMSD (Å). The trimer unit is defined by the trimer type (i.e. topology) and motif type. The topology is encoded as follows: AL: All left-handed Trimer; AR, All Right-handed Trimer; PL: Parallel & Left handed Trimer; PR: Parallel & Right handed Trimer; LR1: Left & Right Type 1 Trimer; LR2: Left & Right Type 2 Trimer. Motif types are encoded as follows: NA: no sequence/structure motif found in the trimer; (AL, M1): I/L/V/M-X_3-G/A/S; (AL, M2): I/L/V/M-X_6-I/L/V/M; (AL, M3): G/A/S-X_6-G/A/S; (AR, M1): G/A/S-X_3-G/A/S; (AR, M2): G/A/S-X_7-G/A/S; (PR, M1): I/L/V/M-X_3-F/W/Y; (PR, M2): F/W/Y-X_2-F/W/Y; (LR1, M1): F/W/Y-X_3-G/A/S; (LR1, M2): F/W/Y-X_6-G/A/S; (LR2, M1): F/W/Y-X_6-G/A/S.

Proteins	State1 description/PDB code		State2 description/PDB code		Entire TM domain Cα RMSD (Å)	Trimer Cα RMSD(Å) (Trimer Type, Motif)			
β2 adrenergic receptor	inactive	2RH1	active	3P0G	1.9	0.4 (AL,M1)	1.2(AL, NA)	0.6(LR2,M1)	
ABCB10 Mitochondrial ABC transporter	Rod Form B	4AYX	Plate form	4AYW	1.1	0.4(LR1, M1)			
ABCB10 Mitochondrial ABC transporter	Rod Form A	4AYT	Plate form	4AYW	1.2	0.4(LR1, M1)			
ABCB10 Mitochondrial ABC transporter	Rod Form A	4AYT	Rod Form B	4AYX	0.8	0.2(LR1, M1)			
YetJ pH-sensitive calcium-leak channel,	open form	4PGS	close	4PGR	0.5	3.2 (PL, NA)			
bile acid transporter	inward open	4N7W	Outward open	4N7X	3.7	2.7(AL, NA)	0.4(AR, M1)	2.6(AL, NA)	2.3 (LR1, NA)
AdiC Arginine: Agmatine Antiporter E coli	outward-facing	3L1L	Open-to-out conformation	3OB6	2.1	0.8 (AR, M1,M2) 0.7(AR, NA)	2.7(PL, NA) 0.5 (PR, M1)	1.0(LR1,NA)	0.9(AL, NA)
Calcium ATPase	E1 state	1SU4	E2 state	1WPG	6.1	2.1 (PR, NA) 4.3 (LR2, NA)	1.4 (PR, NA)	0.9 (AR, M1)	0.7(AR, M1)

AcrB bacterial multi-drug efflux transporter	MonomerA/B(T)	2GIF	monomerC(L)	2DHH	2.0	1.1 (AR, M1, M2)
						1.6 (AR, NA)
							0.9(AR, M1, M2)
							1.4(PR, M2)
								0.9 (LR1, M1)
								1.3(LR2, M1)
									1.8(PR, NA)
									1.4(LR2, NA)

LacY Lactose Permease	pH 6.5, no substrate	2CFQ	Occluded, partially open to periplasmic side	4OAA	4.8	0.5 (AL, M2) 0.7 (LR1, M2)	0.6 (AL, M1)	1.7(LR1, NA)	2.2 (LR1,NA)
Glutamate Transporter	inward-facing	4P19	outward-facing	1XFH	9.4	1.0 (AR, M1)	1.2(AR,M1, M2)	7.5 (LR1, NA)	2.3 (AR, NA)
GlpG Rhomboid protease	conformation 1	2XOV	conformation 2	2NRF	2.7	0.2 (AR, M1, M2)			

Supplementary Figure 5. Sequence/3D contact motifs are strong predictors of local conformational stability. Distribution of trimer unit structural changes (measured by Cα rmsd in **Å**) in multi-pass membrane proteins crystallized in distinct conformations. Only one protein per superfamily was selected leading to 10 protein structures and 40 trimer units. Data for trimers containing enriched sequence/contact motifs are in black, others are in grey. Trimers with sequence motifs and corresponding interaction patterns had substantially smaller Cα rmsd (P < 0.0005, Welch's t-test) between distinct protein conformations and were therefore significantly more rigid than the trimers without such sequence/3D contact features.


Distribution of trimer conformational change
Trimer regions with motifs and consensus interactions
Trimer regions with no motif/interactions
Number of trimers
P < 0.0005
<0.5
> 3.0
Cα RMSD (Å)