GPCR-I-TASSER: A hybrid approach to G protein-coupled receptor structure modeling and the application to the human genome

Jian Zhang, Jianyi Yang, Richard Jang, Yang Zhang
Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109

SUPPLEMENTARY DATA

Table S1. List of 24 GPCRs that have solved structure in the PDB. This table is supplemental to Table 1 and related to the section entitled "Benchmark Test on 24 Solved GPCRs" in RESULT of the main text.

UniProtID	Length	PDBID	Resolution	Protein Name	Organism
P02699	348	2HPY	2.8	Rhodopsin	Bos taurus
P25024	350	2LNL	NA	C-X-C chemokine receptor type 1	Homo sapiens
P31356	448	2ZIY	3.7	Rhodopsin	Todarodes pacificus
P29274	412	3EML	2.6	Adenosine receptor A2a	Homo sapiens
P61073	352	3ODU	2.5	C-X-C chemokine receptor type 4	Homo sapiens
P35462	400	3PBL	2.9	D(3) dopamine receptor	Homo sapiens
P35367	487	3RZE	3.1	Histamine H1 receptor	Homo Sapiens
P08172	466	3UON	3.0	Muscarinic acetylcholine receptor M2	Homo sapiens
P21453	382	3V2Y	2.8	Sphingosine 1-phosphate receptor 1	Homo sapiens
P25116	425	3VW7	2.2	Proteinase-activated receptor 1	Homo sapiens
P07700	483	4AMJ	2.3	Beta-1 adrenergic receptor	Meleagris gallopavo
P08483	589	4DAJ	3.4	Muscarinic acetylcholine receptor M3	Rattus norvegicus
P41145	380	4DJH	2.9	Kappa-type opioid receptor	Homo sapiens
P42866	398	4DKL	2.8	Mu-type opioid receptor	Mus musculus
P41146	370	4EA3	3.0	Nociceptin receptor	Homo sapiens
P32300	372	4EJ4	3.4	Delta-type opioid receptor	Mus musculus
P07550	413	4GBR	4.0	Beta-2 adrenergic receptor	Homo sapiens
P20789	424	4GRV	2.8	Neurotensin receptor type 1	Rattus norvegicus
P28222	390	4IAR	2.7	5-hydroxytryptamine receptor 1B	Homo sapiens
P41595	481	4IB4	2.7	5-hydroxytryptamine receptor 2B	Homo sapiens
Q99835	787	4JKV	2.5	Smoothened homolog	Homo sapiens
P34998	444	4K5Y	3.0	Corticotropin-releasing factor receptor 1	Homo sapiens
P47871	477	4L6R	3.3	Glucagon receptor	Homo sapiens
P51681	352	4MBS	2.7	C-C chemokine receptor type 5	Homo sapiens

Table S2. Helix and missing residue annotation on the 24 GPCR domains used as test proteins in this study. This table is supplemental to Table 1 and related to the section entitled "Benchmark Test on 24 Solved GPCRs" in RESULT of the main text.

PDBID[a]	L[b]	Helix definition[c]							Structure gaps[d] (*I-I+1:d*)	Dom[e]
		Helix-I	Helix-II	Helix-III	Helix-IV	Helix-V	Helix-VI	Helix-VII		
2hpyB	348	34-64	71-100	107-140	150-172	200-226	246-277	285-309		no
2lnlA	296	10-38	46-73	80-110	121-145	171-200	210-239	248-280		no
2ziyA	370	28-58	65-95	102-135	146-168	192-224	253-283	291-314		no
3emlA1	286	6-30	39-63	75-99	116-135	166-186	209-233	244-267	(46-47:9.8) (198-199:12.9)	yes
3oduA1	304	10-40	46-74	80-113	118-148	167-199	213-243	250-278	(204-205:6.9)	yes
3pblA1	272	3-25	32-60	69-102	116-138	155-185	194-225	234-258	(189-190:10.3)	yes
3rzeA1	268	4-27	34-62	70-103	114-136	154-182	195-224	232-254	(140-141:17.7) (186-187:8.5)	yes
3uonA1	278	4-31	38-67	74-107	118-147	165-194	206-234	241-265	(197-198:9.7)	yes
3v2yA1	295	30-57	64-89	99-130	135-157	178-207	218-245	258-279	(133-134:12.3) (208-209:14.9)	yes
3vw7A1	282	9-43	46-74	82-115	121-144	171-203	209-244	251-279	(118-119:10.9) (205-206:9.9)	yes
4amjB	299	8-37	44-73	80-113	124-148	174-206	223-256	263-286		no
4dajD1	271	2-30	37-66	73-106	117-146	164-191	204-227	234-258	(194-195:15.1) (199-200:8.0)	yes
4djhB1	288	3-32	39-67	74-107	118-142	167-198	216-243	247-274	(206-207:11.0) (245-246:9.7)	yes
4dklA1	282	2-31	38-66	73-106	117-141	165-194	205-234	242-269	(198-199:7.2)	yes
4ea3A	278	2-33	39-67	74-104	111-135	155-186	205-235	242-269	(109-110:13.2)	no
4ej4A1	282	5-36	43-70	78-109	122-146	168-197	212-240	248-275	(203-204:6.6)	yes
4gbrA	286	2-32	39-68	75-104	119-143	169-197	213-242	249-272		no
4grvA1	293	10-36	46-75	84-114	132-152	176-204	219-245	258-285	(41-42:10.5) (212-213:14.8)	yes
4iarA1	273	12-39	46-75	82-113	126-148	163-192	202-228	234-258	(153-154:12.8) (196-197:10.8) (230-231:10.3)	yes
4ib4A1	285	7-34	41-69	78-111	118-146	161-196	204-234	240-266	(150-151:7.9) (197-198:10.0)	yes
4jkvA	346	41-70	74-96	123-152	166-187	206-231	257-283	311-334	(163-164:4.7) (302-303:10.0)	yes
4k5yC	248	2-27	34-60	72-101	109-134	149-174	188-212	221-247	(104-105:7.5)	no
4l6rA1	293	16-42	52-77	85-115	126-153	166-192	209-234	243-267	(79-80:13.4)	yes
4mbsB1	292	9-39	46-73	80-113	124-148	170-203	211-239	251-279	(204-205:10.3)	yes

[a]PDB ID of the GPCR domains

[b]Length of the PDB structure domains

[c]Helix domain definition where target sequences have been re-numbered from 1 to *L* according to the PDB structure

[d]Structure gaps due to the missed residues on the PDB structure, where *I* is the residue order and *d* is the distance of the gap in Angstroms.

[e]If additional domains have been fused to assist the determination of the GPCR structure.

Table S3. GPCR-I-TASSER modeling results on 24 test GPCRs where all homologous templates with sequence identity >30% or detectable by PSI-BLAST with E-value <0.05 were excluded[a]. This table is supplemental to Table 1 and related to the section entitled "Benchmark Test on 24 Solved GPCRs" in RESULT of the main text.

Target	L[b]	T_ID[c]	id%[d]	Cov[e]	Template by LOMETS[f]		MODELLER[g]		GPCR-I-TASSER model[h]	
					RMSD	TM-score	RMSD	TM-score	R/R_ali	TM-score
2hpyB	348	3oduA	0.27	0.813	7.66(2.28)	0.655(0.864)	13.63(2.25)	0.681(0.877)	5.25/3.34(1.35/1.34)	0.828(0.941)
2lnlA	296	3v2yA	0.28	0.949	7.87(5.82)	0.507(0.529)	8.11(6.23)	0.537(0.539)	8.39/8.05(6.13/5.94)	0.514(0.526)
2ziyA	370	3vw7A	0.22	0.876	4.71(2.56)	0.632(0.836)	26.05(2.55)	0.658(0.843)	7.36/3.70(1.33/1.30)	0.797(0.951)
3emlA1	286	4grvA	0.29	0.913	4.55(2.18)	0.748(0.855)	5.78(2.10)	0.763(0.868)	3.39/3.42(1.70/1.69)	0.847(0.902)
3oduA1	304	3uonA	0.27	0.908	6.40(3.10)	0.684(0.809)	10.09(3.41)	0.708(0.820)	6.90/4.91(2.28/2.08)	0.814(0.887)
3pblA1	272	4jkvA	0.25	0.941	5.55(4.41)	0.659(0.709)	5.54(4.33)	0.679(0.719)	4.20/3.94(2.41/2.41)	0.793(0.848)
3rzeA1	268	3oduA	0.28	0.978	5.93(2.25)	0.780(0.880)	5.02(2.18)	0.799(0.888)	2.52/2.52(1.29/1.29)	0.917(0.948)
3uonA1	278	4grvA	0.28	0.96	5.08(3.62)	0.744(0.797)	5.73(3.59)	0.764(0.802)	2.90/2.93(2.16/2.16)	0.884(0.907)
3v2yA1	295	4grvA	0.24	0.854	5.11(2.86)	0.675(0.824)	10.84(2.96)	0.700(0.818)	8.26/3.47(1.61/1.61)	0.809(0.929)
3vw7A1	282	4grvA	0.27	0.979	3.69(2.55)	0.785(0.839)	3.66(2.67)	0.805(0.846)	3.59/3.53(2.67/2.60)	0.814(0.843)
4amjB	299	4jkvA	0.21	0.963	6.56(4.48)	0.611(0.665)	6.62(4.58)	0.638(0.681)	6.26/6.22(2.62/2.63)	0.772(0.854)
4dajD1	271	4grvA	0.28	0.956	5.23(3.24)	0.747(0.828)	7.38(3.24)	0.765(0.834)	3.80/3.82(2.49/2.49)	0.872(0.908)
4djhB1	288	4jkvA	0.24	0.931	5.56(4.10)	0.654(0.697)	5.70(4.29)	0.682(0.720)	4.08/4.06(2.69/2.68)	0.788(0.834)
4dklA1	282	4jkvA	0.23	0.936	5.21(4.04)	0.671(0.711)	5.35(3.98)	0.689(0.728)	5.28/5.34(2.64/2.60)	0.815(0.846)
4ea3A	278	3v2yA	0.26	0.942	4.93(2.53)	0.767(0.825)	4.47(2.76)	0.795(0.842)	2.48/2.42(1.60/1.52)	0.901(0.929)
4ej4A1	282	4jkvA	0.24	0.947	6.11(4.66)	0.645(0.696)	5.83(4.30)	0.665(0.714)	2.98/3.00(1.45/1.44)	0.894(0.940)
4gbrA	286	4jkvA	0.23	0.927	6.26(4.69)	0.625(0.654)	6.43(4.71)	0.639(0.663)	6.22/5.82(2.54/2.54)	0.776(0.855)
4grvA1	293	4mbsA	0.27	0.901	4.77(2.89)	0.735(0.807)	5.44(3.19)	0.733(0.802)	4.52/4.38(2.51/2.42)	0.785(0.848)
4iarA1	273	4jkvA	0.22	0.963	5.89(4.67)	0.644(0.697)	6.42(5.00)	0.635(0.687)	2.82/2.67(1.70/1.70)	0.884(0.918)
4ib4A1	285	4jkvA	0.26	0.975	5.78(4.52)	0.656(0.688)	5.78(4.60)	0.675(0.699)	6.04/5.09(2.60/2.52)	0.807(0.866)
4jkvA	346	4l6rA	0.19	0.902	7.49(3.84)	0.537(0.683)	15.15(5.51)	0.586(0.704)	10.41/7.41(3.51/3.15)	0.635(0.787)
4k5yC	248	4jkvA	0.27	0.984	7.17(6.62)	0.658(0.681)	7.83(7.18)	0.635(0.664)	5.61/4.66(4.45/3.58)	0.731(0.778)
4l6rA1	293	4grvA	0.27	0.942	6.02(4.12)	0.635(0.729)	10.78(3.96)	0.660(0.746)	5.89/3.84(2.21/2.23)	0.796(0.882)
4mbsB1	292	4grvA	0.28	0.914	4.12(2.79)	0.753(0.821)	5.95(2.85)	0.764(0.826)	3.03/2.82(1.80/1.78)	0.876(0.915)
Average	**292**		**0.25**	**0.931**	**5.74(3.70)**	**0.675(0.755)**	**8.07(3.85)**	**0.694(0.764)**	**5.09/4.22(2.40/2.32)**	**0.806(0.868)**

[a]Numbers in parenthesis are RMSD and TM-score in the transmembrane regions.

[b]Length of the target sequence

[c]PDB ID of the best template detected by LOMETS

[d]Sequence identity of template to the target

[e]Coverage of the threading alignment defined as the number of the aligned residues divided by the target length

[f]RMSD and TM-score of the threading template

[g]RMSD and TM-score of the MODELLER model built based on the best template

[h]Result of the first GPCR-I-TASSER model, where R and R_ali are RMSD of the entire chain and the aligned regions, respectively.

Table S4. Modeling results without using any homologous or membrane proteins as templates.[a] This table is supplemental to Table 1 and related to the section entitled "Benchmark Test on 24 Solved GPCRs" in RESULT of the main text.

Target	Models by *Ab initio* folding		Models by GPCR-I-TASSER	
	R_1[b]/R_B[c]	TM_1[d]/TM_B[e]	R_1[b]/R_B[c]	TM_1[d]/TM_B[e]
2hpyB	11.95/11.95(8.13/8.13)	0.394/0.398(0.409/0.417)	11.36/10.77(6.80/6.73)	0.482/0.487(0.501/0.512)
2lnlA	11.09/10.82(8.52/8.14)	0.380/0.410(0.379/0.410)	9.60/9.25(7.16/6.95)	0.460/0.488(0.443/0.498)
2ziyA	16.18/16.18(10.34/9.28)	0.347/0.392(0.396/0.430)	12.63/10.90(6.20/6.20)	0.487/0.487(0.548/0.554)
3emlA1	12.44/12.44(5.84/5.84)	0.449/0.449(0.496/0.496)	7.88/7.88(4.98/4.98)	0.569/0.569(0.581/0.581)
3oduA1	11.69/11.69(8.45/8.45)	0.402/0.402(0.423/0.423)	10.35/10.05(6.18/6.18)	0.517/0.517(0.535/0.535)
3pblA1	9.70/9.60(8.88/8.39)	0.411/0.413(0.368/0.388)	7.87/7.87(6.50/6.50)	0.518/0.518(0.508/0.508)
3rzeA1	9.85/9.85(8.96/7.38)	0.379/0.442(0.354/0.462)	7.80/7.53(6.33/6.33)	0.514/0.514(0.498/0.500)
3uonA1	12.19/9.65(9.27/8.40)	0.384/0.427(0.412/0.426)	7.63/7.61(6.40/6.14)	0.522/0.535(0.518/0.530)
3v2yA1	12.17/11.80(9.40/9.40)	0.369/0.369(0.334/0.352)	8.16/8.16(6.19/6.09)	0.527/0.528(0.538/0.542)
3vw7A1	11.29/11.20(9.29/9.29)	0.370/0.376(0.382/0.386)	7.65/7.65(6.53/6.53)	0.502/0.502(0.493/0.493)
4amjB	10.82/10.82(9.00/9.00)	0.417/0.417(0.417/0.417)	8.22/8.04(6.15/6.15)	0.523/0.526(0.523/0.531)
4dajD1	9.51/9.51(8.84/8.84)	0.410/0.410(0.395/0.405)	7.27/7.27(6.02/6.02)	0.555/0.555(0.539/0.539)
4djhB1	11.12/10.53(8.14/8.04)	0.440/0.445(0.473/0.473)	8.33/8.33(6.71/6.71)	0.521/0.521(0.527/0.527)
4dklA1	10.21/9.94(9.49/8.49)	0.386/0.393(0.359/0.401)	7.68/6.93(6.65/5.74)	0.520/0.558(0.508/0.554)
4ea3A	10.35/10.31(9.05/8.70)	0.401/0.401(0.378/0.378)	6.77/6.77(6.00/5.82)	0.556/0.556(0.535/0.535)
4ej4A1	10.52/10.18(9.02/8.36)	0.403/0.413(0.385/0.423)	7.43/7.43(6.31/6.31)	0.527/0.527(0.509/0.509)
4gbrA	11.43/10.62(9.88/9.32)	0.360/0.368(0.352/0.368)	7.46/7.46(5.99/5.99)	0.540/0.540(0.525/0.525)
4grvA1	10.06/10.06(8.48/7.43)	0.403/0.443(0.392/0.463)	7.51/7.51(6.12/6.12)	0.537/0.537(0.532/0.532)
4iarA1	11.62/10.58(10.85/7.97)	0.344/0.406(0.317/0.403)	7.54/7.54(6.51/6.51)	0.517/0.517(0.491/0.491)
4ib4A1	10.61/10.61(8.99/7.95)	0.395/0.455(0.407/0.453)	8.12/8.12(6.66/6.58)	0.518/0.523(0.505/0.512)
4jkvA	11.01/11.01(7.91/7.72)	0.396/0.405(0.408/0.417)	11.58/11.50(5.79/5.67)	0.497/0.507(0.535/0.552)
4k5yC	9.94/9.09(9.40/8.09)	0.385/0.431(0.363/0.425)	7.78/7.78(7.04/6.97)	0.532/0.532(0.523/0.530)
4l6rA1	14.10/9.61(9.90/7.81)	0.334/0.442(0.338/0.440)	11.67/10.63(7.60/6.88)	0.431/0.488(0.444/0.499)
4mbsB1	13.49/11.41(8.93/8.93)	0.365/0.392(0.394/0.402)	7.45/7.45(6.01/6.01)	0.543/0.543(0.536/0.536)
Average	**11.39/10.81(8.96/8.31)**	**0.389/0.412(0.389/0.419)**	**8.57/8.35(6.37/6.25)**	**0.517/0.524(0.517/0.526)**

[a]Numbers in parenthesis are RMSD and TM-score in the transmembrane regions.

[b]R_1: RMSD of the first model

[c]R_B: RMSD of the best in top five models

[d]TM_1: TM-score of the first model

[e]TM_B: TM-score of the best in top five models

Table S5. Top 10 groups in the GPCRDock 2010 experiment based on total Z-score of receptor and ligand models. Data are taken from http://ablab.ucsd.edu/GPCRDock2010/. This table is supplemental to Figure 3 and related to the section entitled "Blind Test in the GPCRDock Experiment" in RESULT of the main text.

Group ID	Total Z-score		Z-score of receptor		Z-score of ligand	
	First model	Best model	First model	Best model	First model	Best model
UMich-Zhang/0460	5.23	5.92	3.12	3.22	2.11	3.30
COH-Vaidehi/2560	3.40	5.60	-1.65	-1.61	5.05	6.11
PharmaDesign/0400	2.73	2.94	3.08	3.08	-0.35	0.16
UNM/7334	2.41	2.41	1.13	1.15	1.28	1.28
CDD-CMBI/8004	2.20	2.59	-0.20	-0.07	2.40	2.79
Monash-Hall/3801	1.74	1.86	0.93	1.04	0.81	0.99
Helsinki-Xhaard/5508	1.72	1.72	1.26	1.26	0.46	0.46
UMich-Pogozheva/7425	1.62	1.62	2.45	2.45	-0.83	-0.83
QUB/3682	1.22	1.74	0.21	0.28	1.01	1.52
Baylor-Barth/7533	1.17	2.42	2.05	3.54	-0.88	-0.36

Table S6. Top twenty GPCR families which have the highest number of GPCRs with high C-score predictions. This table is supplemental to Figure 5 and related to the section entitled "Structure Modeling of 1026 GPCRs in the Human Genome" in RESULT of the main text.

#	**Counts**[a]	**Family**
1	419	Odorant/olfactory and gustatory
2	29*	Chemokines and chemotactic factors
3	25	Family T2R (taste receptors)
4	17*	Adenosine and adenine nucleotide
5	14	Lysolipids
6	9	Opsins
7	7	Serotonin
8	7*	Adrenergic
9	7	Trace amine
10	7	Prostanoids
11	6	Somatostatin and urotensin
12	5	Pheromone
13	5	Neuropeptide Y
14	5	Releasing hormones
15	5*	Dopamine
16	5	Melanocortins
17	4	Vasopressin / oxytocin
18	4*	Histamine
19	4	Opioid peptides
20	3	Acetylcholine (muscarinic)

[a]Number of GPCRs that have a C-score >-1.5 in each family. '*' indicates the families which have at least one member with experimentally solved structures.

Table S7. The values of the van der Waals radius parameter r_i (Å) and solvation parameter σ_i (cal/mol/Å^2) for the TM transfer energy in Eq. S3. This table is supplemental to the EXPERIMENTAL PROCEDURES of the main text.

Atoms	C-sp3	C-sp2	N-sp3	N-sp2	O	S	Water
r_i	1.87	1.76	1.50	1.65	1.40	1.85	1.40
σ_i	22.6	19.0	-53.0	-53.0	-57.0	-10.0	

SUPPLEMENTAL EXPERIMENTAL PROCEDURE

Generation of transmembrane helix framework

The initial TM-helix bundle is constructed by either threading or *ab initio* TM folding depending on the quality of templates as measured by the significance score (Z-score) of the threading alignments.

Template identification by threading. The query GPCR sequence is threaded through the PDB library by LOMETS (Wu and Zhang, 2007) to identify appropriate structure templates. LOMETS is a multi-threading approach consisting of nine complementary threading algorithms from FFAS (sequence profile-profile match) (Rychlewski et al., 2000), HHsearch (hidden Markov model to hidden Markov model alignment) (Soding, 2005), MUSTER (multiple resource profile-profile alignment) (Wu and Zhang, 2008), PRC (hidden Markov model match) (Madera, 2008), PROSPECT2 (contact-assisted profile-profile alignment) (Xu et al., 1999), dPPAS (depth assisted profile-profile match) (Yan et al., 2013), SAM-T02 (sequence to hidden Markov model alignment) (Karplus et al., 1998), SPARKS (profile alignment assisted with single-body potential) (Zhou and Zhou, 2004), SP3 (profile alignment assisted with fragment depth) (Zhou and Zhou, 2005).

The threading algorithms are designed for generic proteins. To enhance the accuracy of threading alignments for GPCRs, we exploit three transmembrane helix prediction programs: HMMTOP (Tusnady and Simon, 1998), MEMSAT (Jones et al., 1994), and TMHMM (Krogh et al., 2001), to predict the location of the TM helices along the sequence. An additional term accounting for TM residue matches was integrated in all the threading alignments in LOMETS, i.e.,

$$\text{Score}(i,j) = \sum_{p=1}^{3} \delta\left(s_q(i,p), s_t(j)\right) \tag{S1}$$

where $s_q(i, p)$ indicates the secondary structure type (SS, TM helix or loops) for the ith residue by the pth TM-helix prediction program, and $s_t(j)$ labels if the jth residue is located on the TM helix of the template structure according to the DSSP assignment. This term was designed to guide the alignment algorithms to match the TM- and non-TM-regions correctly.

For each threading program, a Z-score (defined as the difference between the alignment score and the mean in units of standard deviation) is assigned to assess the significance of the alignments where a set of program-specific cutoffs are calculated from a training set of membrane proteins for scaling the Z-score thresholds, i.e. Z^{cut}=18.0, 9.7, 6.0, 21.0, 3.2, 12.0, 15.0, 7.0, 7.0, for FFAS, HHsearch, MUSTER, PRC, PROSPECT2, dPPAS, SAM-T02, SPARKS, SP3, respectively. This training protein set is non-homologous to the testing GPCRs used in this study. In general, if none of the threading alignments have a Z-score $>Z^{cut}$, the GPCR is a classified as a "hard" target and the template alignments generally have low quality.

Ab initio folding of TM domains. For the hard targets, a new *ab initio* folding approach is developed to construct the TM framework from scratch. Following the TM-helix predictions, an initial TM-helix bundle conformation is built by laying seven ideal helices sequentially along the perimeter of a circle of radius 8 Å, in which all helices are initially perpendicular to the two membrane planes (Figure 9A). Replica-exchange Monte Carlo (REMC) simulation is then implemented to reassemble the TM-helix topology. Two sets of MC movements are used for conformation updates: the global TM movements including translation, rotation, and tilting of the helices; the local TM movements containing sequence shifts along the helix, addition/deletion of residues, and helix kinking (Figure 9B). After each helix movement, a set of inter-TM loops is rebuilt by the CCD algorithm (Dunbrack and Canutescu, 2003) to connect the helix bundle into full-length structures, which also confines the global TM movements.

The REMC simulations were guided by a simple force field consisting of two atomic energy terms. The first is a knowledge-based, distance-specific contact potential, RW (Zhang and Zhang, 2010b), which was derived from the statistics of PDB structures normalized by a sample of random-walked chains, i.e.,

$$\bar{u}(\alpha,\beta,R) = -kT \ln \frac{N_{obs}(\alpha,\beta,R)}{(R/R_0)^2 N_{obs}(\alpha,\beta,R_0) \frac{\sum_{n=1}^{N} \frac{\exp(-3R^2/2n\lambda)}{n^{3/2}}}{\sum_{n=1}^{N} \frac{\exp(-3R_0^2/2n\lambda)}{n^{3/2}}}} \quad \text{(S2)}$$

where R is the distance between atoms of the atom types α and β; $N_{obs}(\alpha,\beta,R)$ is the observed number of atom pairs (α, β) in a set of 1,383 high-resolution PDB structures within a distance shell R to $R+\Delta R$; N is the protein length; λ=460, R_0=15.5 Å; k is the Boltzmann constant with T=298K.

The second term counts for the free energy change of GPCR and water/lipid interactions using the form of Lomize *et al*. (Lomize et al., 2006):

$$\Delta G = \sum_i S(r_i) \frac{\sigma_i}{1+e^{d_i/0.9}} \quad \text{(S3)}$$

where $S(r_i)$ is the accessible surface area of the ith atom of the structure which is calculated using the Shrake-Rupley algorithm (Shrake and Rupley, 1973) with r_i being the van der Waals radius. σ_i is the solvation parameter that measures the free-energy transfer of the ith atom from solvent to the membrane interior in cal/mol/Å^2. d_i is the distance from the ith atom to the closest membrane plane as defined in Figure 9B. ΔG is calculated only for the atoms on the surfaces of the TM domain, which varies when the conformation and the relative location of the TM domain in the membrane change. The parameters for the different atom types re-optimized for GPCR modeling are listed in Table S7. A combination of the two energy terms from Eqs. S2 and S3 with equal weight was found to work best in our training, so this was used in our simulations.

For each target, 40 replicas are simulated in parallel, with each replica having 10 million MC movement attempts, which are accepted/rejected according to the Metropolis criterion. The structure decoys in the low-temperature replicas are clustered by SPICKER (Zhang and Skolnick, 2004c) and the structures with the highest cluster density are selected as the *ab initio* TM-helix models for the fragment assembly simulations.

Template-based fragment assembly simulations

Starting with the threading templates or *ab initio* models, full-length GPCR models were constructed by REMC simulations following the I-TASSER protocol (Roy et al., 2010; Wu et al., 2007; Zhang and Skolnick, 2004a). The GPCR sequences are split into two types of regions in the simulations which are reassembled and refined on an "on-and-off" lattice system. First, continuous structure fragments from the TM regions are excised from the threading alignments or *ab initio* models. These continuous fragments are kept semi-rigid and represented off-lattice. Second, the loop/tail regions are represented on-lattice and rebuilt from scratch. Accordingly, two types of conformational updates are implemented: the movements of TM helices involve rigid translation and rotation of the fragments by the 3 Euler angles; lattice confined residues are subject to 2-6 bond movements and multi-bond sequence shifts (Zhang et al., 2003). To account for the local kinks in the TM helices, we introduced a small kinking deformation of the TM helices using a strong penalty term of $E \sim \Delta RMSD^4$, where $\Delta RMSD$ denotes the RMSD between the excised template substructure and the deformed substructure in the simulation.

The force field of GPCR-I-TASSER consists of three components. The first component is a generic knowledge-based potential extended from I-TASSER (Roy et al., 2010; Wu et al., 2007), which includes statistical Cα and side-chain contact potentials derived from the PDB, backbone-orientation specific hydrogen-bond, solvation from neural network prediction, and predicted secondary structure propensities. While most of the energy terms were extended from I-TASSER, the parameters in the contact potentials and

the solvation term were re-trained on a set of non-redundant membrane proteins in the PDB.

The second component is the spatial restraints derived from LOMETS templates and/or *ab initio* TM-helix models, which consists of Cα distance maps and Cα and side-chain contacts, i.e., $E_{restr} = -\sum_{i<j} 1/|\, d_{ij} - d_{ij}^0 \,| + \sum_{i<j} w_{ij}\delta(d_{ij} > d_{cut})$, where d_{ij} and d_{ij}^0 denote respectively the distance in decoys and the predicted distance between ith and jth atoms; w_{ij} is the confidence of the predicted contact and d_{cut} is an amino acid-specific distance cutoff for defining residue contacts. The first term is to encourage the satisfaction of Cα distance maps (with a cutoff on the denominator to avoid singularity) while the second penalizes the violations of the predicted Cα and side-chain contacts. For the "hard" GPCRs which do not have strong threading alignments, we re-ranked all the threading template alignments based on TM-score to the *ab initio* TM-helix models before deriving the restraints from the templates. This re-ranking helps pick up the templates of correct TM topology (Zhang, 2014).

The third component of the GPCR-I-TASSER potential is GPCR- and/or transmembrane-specific and consists of the following six energy terms.

1. Membrane repulsive energy. TM-helices in GPCRs are all embedded between two membrane surfaces and packed within a narrow cylinder that is approximately perpendicular to the membrane surface due to the repulsive force of the lipid bilayer. Following this topology, we define two parallel transmembrane planes which cross the center of mass of the terminal Cα atoms of the 7-TM helices and run perpendicular to the sum of the first two principal component vectors of the top and bottom ending Cα atoms (Figure 9A). A membrane repulsive energy is defined as

$$E_{mrep} = \sum_i E_{mrep}^i, \quad \text{where } E_{mrep}^i = \begin{cases} d_i, & \text{if } d_i > 0 \,\&\, r_i > 6.0 \\ 0, & \text{otherwise} \end{cases} \tag{S4}$$

where r_i is the distance of the ith atom to the axis of the TM-helix bundle and d_i is the distance of the atom to the closest membrane plane as defined in Eq. S3 (the direction of d_i is defined so that $d_i > 0$ if the atom is inside membrane and $d_i < 0$ otherwise, Figure 9B). A cutoff of r_i=6 Å is used which is equal to the average radius of TM-helix bundles of GPCRs in the PDB. Defining the membrane repulsive energy in this manner penalizes TM helices from moving away from each other to outside the experimentally established ranges and therefore helps guide TM-helix packing in the membrane. The membrane repulsive energy in Eq. S4 also protects the loop and tail residues outside the membrane from moving into the TM regions inappropriately. Figure 4B shows a modeling example from the *Homo sapiens* olfactory receptor 1S2 (UniProt ID: Q8NGQ3), where the N-terminal tail entered into the TM region when the membrane repulsive energy was not used, but was successfully moved out of the membrane when the energy term was introduced.

2. Extra/intra-cellular hydrophilic interactions. Hydrophilic interactions for GPCR residues inside and outside the membrane follow different regularities, i.e., the hydrophobic residues in the TM region tend to be exposed to the membrane due to the exclusion of water in lipids while the residues outside the membrane tend to be buried from contact with the solvent. We introduce a solvation energy term for the extra/intra-cellular domain:

$$E_{EIH} = \sum_{i=1}^{N_{EI}} \left(\frac{x_i^2}{x_0^2} + \frac{y_i^2}{y_0^2} + \frac{z_i^2}{z_0^2} - 2.5 \right) P(i) \tag{S5}$$

where (x_i, y_i, z_i) is the coordinate of the ith atom in the ellipsoid of a given GPCR conformation and (x_0, y_0, z_0) gives the lengths of the principal axes of the ellipsoid. The value 2.5 is a fitting parameter to tune the average depth of the exposed residues. $P(i)$ is the exposure index of the residue as predicted by a neural network. The potential is similar to that used in I-TASSER, but the sum i only goes through N_{EI} residues in the extracellular and intracellular loop regions.

3. Hydrophobic moment energy. This term is introduced to adjust TM-helix orientations:

$$E_{hm} = \sum_{k=1}^{7} \mathbf{V}_k \cdot \mathbf{H}_k \tag{S6}$$

where $\mathbf{V}_k$ is a unit vector directed from the center of the kth TM helix to the center of mass of the TM helix bundle; $\mathbf{H}_k$ is the hydrophobic moment of kth TM helix calculated as the mean vector sum of the hydrophobicities of all side chains along the helix (Eisenberg et al., 1982). The purpose of Eq. S6 is to orient the hydrophilic residues towards the interior of the TM-helix bundle, which accommodates the relative packing of the seven helices.

4. Aromatic interactions. The enhanced interactions between aromatic residues (Phe, Tyr, Trp, His) are incorporated by

$$E_{aa} = \sum_{i>j}^{N} c_{ij} q_{ij} \tag{S7}$$

where q_{ij} is the quasi-chemical pair-wise contact potential between the ith and the jth residues (Zhang et al., 2003), which is negative. By trial and error, we set c_{ij}=3.0 for (i, j)=(Phe, Tyr, Trp, His), and c_{ij}=1 for other residue pairs. The parameter c_{ij} accounts for the enhanced stability from clusters of aromatic residues, which are widely observed in GPCRs and are important for TM-helix packing (Burley and Petsko, 1985).

5. Cation-π interactions. Cation-π interactions play an important role to the stability of the helix protein packing (Gromiha, 2003). The enhanced Cation-π interactions for specific non-covalent binding propensities between TM helices of GPCRs are incorporated by

$$E_{cp} = \sum_{i>j}^{N} f_{ij} q_{ij} \tag{S8}$$

where we set by trial and error f_{ij}=3.0 for i=(Arg, Lys) and j=(Phe, Tyr, Trp, His); f_{ij}=1 for other residue pairs.

6. GPCR-RD experimental restraints. Two types of spatial restraints are derived from the site-directed mutagenesis and affinity labeling experiments collected from the GPCR-RD database (Zhang and Zhang, 2010a). First, contact restraints are implemented for the experimentally identified disulfide bridges and the functionally important residues that bind to a particular ligand, i.e.,

$$E_{contact}(i,j) = \begin{cases} -1 & d_{ij} < 10 \\ 0 & d_{ij} \geq 10 \end{cases} \tag{S9}$$

where d_{ij} is the distance between side-chain centers of residues i and j. A distance cutoff 10 Å is used since this is close to the average distance of ligand-binding residues in receptor proteins in the PDB (Yang et al., 2013). Although the contact restraint cutoff is relatively large, many of the initial conformations from threading and *ab initio* TM domain assembly were found not to satisfy the restraints, in particular for the hard targets where the introduction of the GPCR-RD restraints helped adjust the topology of the helical arrangements.

Second, most of the functionally related point mutations (binding pocket or helix-helix interfaces) are on the residues that face the inside of the TM-helix bundle (Schushan et al., 2010; Shacham et al., 2004). To reflect this observation, we introduce an orientation restraint for point mutation residues from the GPCR-RD, i.e.,

$$E_{orientation}(i) = \begin{cases} -1 & \boldsymbol{V}_i \cdot \boldsymbol{e}(i) > 0 \\ 0 & \boldsymbol{V}_i \cdot \boldsymbol{e}(i) \leq 0 \end{cases} \tag{S10}$$

where $\boldsymbol{V}_i$ is a unit vector as defined in Eq. S6, and $\boldsymbol{e}(i)$ is a unit vector directed from the TM-helix axis to the

Cα atom of the *i*th mutated residue (Figure 9A). Since the structural interpretation of mutagenesis data is often not univocal and with false positives, the restraints were implemented as energy terms rather than as hard constraints.

The energy terms from different components were combined in a linear regression with the weighting factors optimized using a similar protocol used previously (Zhang et al., 2003), by maximizing the correlation between TM-score and the total energy based on structure decoys from 50 non-redundant membrane proteins that are non-redundant with the testing GPCRs.

Here we note that most of the above energy terms in the third component can be applicable to general transmembrane proteins. But there are still many features in GPCR-I-TASSER that are specifically designed for GPCR structure modeling. For instance, the *ab initio* modeling starts from idealized seven-helix bundle structures that limits the *ab initio* folding procedure applicable only to GPCRs. The current threading library for the GPCR-I-TASSER server only contains GPCR structures to enhance the GPCR alignment accuracy. Several parameters, including the water/lipid interaction potentials in Table S7 and the distance cutoff parameters in the membrane repulsive energy term (Eq. S4), were optimized based on the statistics of GPCR structures. In particular, spatial restraints, including residue contacts and helix orientations, are taken from GPCR mutagenesis experiments. These features have made the GPCR-I-TASSER pipeline highly specific for GPCR structure modeling.

Model selection and fragment-guided structure refinement

Following the GPCR-I-TASSER simulations, structure decoys generated in low-temperature replicas are submitted to SPICKER (Zhang and Skolnick, 2004c) for structure clustering. The decoys with the highest number of structural neighbors are selected which correspond to the states of the lowest free energy in the REMC simulations. Full-atomic models are finally constructed from the selected decoys, which are refined by FG-MD, the fragment-guided molecule dynamic simulations using AMBER99 force field assisted with distance map restraints, explicit hydrogen binding and an experience repulsive potential (Zhang et al., 2011). Furthermore, the SPICKER centroid model is used as a probe to identify analog fragments from the PDB by TM-align, which provides additional spatial restraints to improve the energy landscape funnel in atomic-level structure refinements. Since the MD simulations have been strongly constrained by the initial models and TM-align templates, no membrane was implemented for accelerating the simulations.

Multiple-domain assembly

Several dozen GPCR sequences in the human genome are associated with a long loop/tails, which often fold as independent domains. For these GPCRs, we first use ThreaDom (Xue et al., 2013) to identify the domain boundary and then use GPCR-I-TASSER and I-TASSER to fold the receptor and globular domains separately. The full-length models are built by docking the domain models using the whole-chain model as a reference template, where the reference template was selected from the whole-chain GPCR-I-TASSER model that has the highest TM-score to the individual domain models. Once the full-chain template is selected, the domains are docked onto the template through a quick Metropolis Monte Carlo simulation, where the simulation energy is defined as the RMSD of the domain models to the whole-chain model template plus the reciprocal of the number of inter-domain steric clashes (Zhang, 2014). An example of the domain parsing and assembly procedure from the *Homo sapiens* gene Q6ZMI9 is presented in Figure 4.

Estimation of residue-specific local structure quality

To estimate the residue-level quality of local structures, we conducted I-TASSER based structure modeling for 1,270 non-redundant single-domain proteins from the PDB, which are randomly split into two sets of training and test proteins. Support vector regressions (SVRs) was used to train residue-specific distance error and B-factor of the predicted models on the following five features.

(1) ***Structural variation of assembly simulations.*** The structural variation of the *j*th residue in the REMC simulations is defined by the average and standard deviations:

$$\begin{cases} \mu_j = \dfrac{1}{N}\sum_{i=1}^{N} d_{ij} \\ \nu_j = \sqrt{\dfrac{1}{N}\sum_{i=1}^{N}\left(d_{ij} - \mu_j\right)^2} \end{cases} \quad \text{(S11)}$$

where N is the number of decoys in the SPICKER cluster; d_{ij} is the distance for the *j*th residue between the *i*th decoy and the centroid structure model after TM-score superposition (Zhang and Skolnick, 2004b). In general, residues with a higher variation have a larger error relative to the native, and vice versa.

(2) ***Consistency of model and sequence-based feature predictions***. Secondary structure (SS) and solvent accessibility (SA) of the target sequence are predicted by PSSpred (http://zhanglab.ccmb.med.umich.edu/PSSpred/) and SOLVE (Zhang et al, unpublished) programs, which are compared with the actual SS and SA of the 3D structural models that are assigned by STRIDE (Frishman and Argos, 1995). The residues with inconsistent SS and SA usually have larger errors.

(3) ***Threading alignment coverage***. The alignment coverage of a residue is defined as the number of threading templates that have the query residue aligned divided by the total number of templates by LOMETS. The residues with a higher threading coverage indicate more constraints on them and presumably have a higher modeling accuracy.

(4) ***Template-based B-factor assignment.*** B-factor values of the top threading templates are extracted from the PDB entries, which are used to assign the B-factor profiles of the query sequence by:

$$b_q(j) = \frac{1}{n_j}\sum_{i=1}^{n_j} b_t(i,j) \quad \text{(S12)}$$

where n_j is the number of the templates that have a residue aligned to the query residue j, and $b_t(i, j)$ is the normalized B-factor of the residue from the *i*th template that is aligned to j by LOMETS.

(5) ***Sequence profiles***. The target sequence is searched by PSI-BLAST through NCBI's non-redundant sequence database to retrieve homologous sequences, which are represented in the form of a position-specific scoring matrix (PSSM). For each residue, a sliding window with size 15 residues is used to extract profile features from the PSSM after converting its elements x in the range of (0, 1) by 1/[1+exp(-x)].

SUPPLEMENTAL REFERENCES

Burley, S.K., and Petsko, G.A. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. Science *229*, 23-28.

Dunbrack, R.L., and Canutescu, A.A. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci *12*, 963-972.

Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. (1982). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature *299*, 371-374.

Frishman, D., and Argos, P. (1995). Knowledge-based protein secondary structure assignment. Proteins *23*, 566-579.

Gromiha, M.M. (2003). Influence of cation-pi interactions in different folding types of membrane proteins. Biophys Chem *103*, 251-258.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry-Us *33*, 3038-3049.

Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics *14*, 846-856.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology *305*, 567-580.

Lomize, A.L., Pogozheva, I.D., Lomize, M.A., and Mosberg, H.I. (2006). Positioning of proteins in membranes: a computational approach. Protein science : a publication of the Protein Society *15*, 1318-1333.

Madera, M. (2008). Profile Comparer: a program for scoring and aligning profile hidden Markov models. Bioinformatics *24*, 2630-2631.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc *5*, 725-738.

Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein science : a publication of the Protein Society *9*, 232-241.

Schushan, M., Barkan, Y., Haliloglu, T., and Ben-Tal, N. (2010). C(alpha)-trace model of the transmembrane domain of human copper transporter 1, motion and functional implications. Proceedings of the National Academy of Sciences of the United States of America *107*, 10908-10913.

Shacham, S., Marantz, Y., Bar-Haim, S., Kalid, O., Warshaviak, D., Avisar, N., Inbal, B., Heifetz, A., Fichman, M., Topf, M.*, et al.* (2004). PREDICT modeling and in-silico screening for G-protein coupled receptors. Proteins *57*, 51-86.

Shrake, A., and Rupley, J.A. (1973). Environment and Exposure to Solvent of Protein Atoms - Lysozyme and Insulin. Journal of Molecular Biology *79*, 351-371.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics *21*, 951-960.

Tusnady, G.E., and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. Journal of molecular biology *283*, 489-506.

Wu, S., Skolnick, J., and Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol *5*, 17.

Wu, S., and Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. Nucl. Acids. Res. *35*, 3375-3382.

Wu, S., and Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins *72*, 547-556.

Xu, Y., Xu, D., Crawford, O.H., Einstein, Larimer, F., Uberbacher, E., Unseren, M.A., and Zhang, G. (1999). Protein threading by PROSPECT: a prediction experiment in CASP3. Protein Eng *12*, 899-907.

Xue, Z., Xu, D., Wang, Y., and Zhang, Y. (2013). ThreaDom: extracting protein domain boundary information from multiple threading alignments. Bioinformatics *29*, i247-i256.

Yan, R., Xu, D., Yang, J., Walker, S., and Zhang, Y. (2013). A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep *3*, 2619.

Yang, J., Roy, A., and Zhang, Y. (2013). BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic acids research *41*, D1096-1103.

Zhang, J., Liang, Y., and Zhang, Y. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure *19*, 1784-1795.

Zhang, J., and Zhang, Y. (2010a). GPCRRD: G protein-coupled receptor spatial restraint database for 3D structure modeling and function annotation. Bioinformatics *26*, 3004-3005.

Zhang, J., and Zhang, Y. (2010b). A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. Plos One *5*.

Zhang, Y. (2014). Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins *82 Suppl 2*, 175-187.

Zhang, Y., Kolinski, A., and Skolnick, J. (2003). TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophys. J. *85*, 1145-1164.

Zhang, Y., and Skolnick, J. (2004a). Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. USA *101*, 7594-7599.

Zhang, Y., and Skolnick, J. (2004b). Scoring function for automated assessment of protein structure template quality. Proteins *57*, 702-710.

Zhang, Y., and Skolnick, J. (2004c). SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem *25*, 865-871.

Zhou, H., and Zhou, Y. (2004). Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins *55*, 1005-1013.

Zhou, H., and Zhou, Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins *58*, 321-328.