nature COMMUNICATIONS

ARTICLE

Received 15 Sep 2014 | Accepted 15 Apr 2015 | Published 21 May 2015

DOI: 10.1038/ncomms8196

Evolutionary-guided *de novo* structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy

Y. Wang[1] & P. Barth[1,2,3]

How specific protein associations regulate the function of membrane receptors remains poorly understood. Conformational flexibility currently hinders the structure determination of several classes of membrane receptors and associated oligomers. Here we develop EFDOCK-TM, a general method to predict self-associated transmembrane protein helical (TMH) structures from sequence guided by co-evolutionary information. We show that accurate intermolecular contacts can be identified using a combination of protein sequence covariation and TMH binding surfaces predicted from sequence. When applied to diverse TMH oligomers, including receptors characterized in multiple conformational and functional states, the method reaches unprecedented near-atomic accuracy for most targets. Blind predictions of structurally uncharacterized receptor tyrosine kinase TMH oligomers provide a plausible hypothesis on the molecular mechanisms of disease-associated point mutations and binding surfaces for the rational design of selective inhibitors. The method sets the stage for uncovering novel determinants of molecular recognition and signalling in single-spanning eukaryotic membrane receptors.

[1] Structural and Computational Biology and Molecular Biophysics Graduate Program, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [2] Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [3] Department of Pharmacology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. Correspondence and requests for materials should be addressed to P.B. (email: patrickb@bcm.edu).

Protein associations regulate the function of a large diversity of membrane proteins, such as tyrosine kinase (RTK), cytokine, immune or G protein-coupled receptors[1–5]. Single spanning receptors such as RTKs can adopt multiple conformations and function by extracellular ligand-induced stabilization of specific receptor homo- or heterodimeric conformations triggering activation of cytoplasmic signalling cascades[6–9]. By changing orientation or oligomerization states, transmembrane (TM) and juxtamembrane (JM) regions play critical roles in regulating receptor associations and in transmitting signals across the membrane[7,8,10]. Numerous point mutations in their TM or TM–JM boundary regions perturb the receptor's conformations and functions, and are associated with severe disease[1,11,12], hence the importance of determining their structure for rational drug design applications.

However, compared with multi-pass membrane proteins, single-pass oligomeric membrane receptors (SPMRs) are highly flexible and remain very difficult to characterize structurally. Several extramembrane (EM) and a few TM domains have been characterized by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy[13–18], respectively, but no high-resolution structure of a full-length SPMR has been solved to date. Nevertheless, current evidence on widely studied receptors such as epidermal growth factor receptor (EGFR) and integrin indicate that TM interactions and structures determined from isolated domains are consistent with those in full-length receptors[8,9,19–21]. Thus, the structural characterization of isolated TM domains can be considered as a valid first approach to identify native TM–TM interactions in full-length receptors. When extensive experimental information is available on TM interactions (for example, mutational, crosslinking, infrared spectroscopy and homologue structures), TM structures can be modelled accurately[22] and full-length receptor structures can be reconstructed by linking EM structures with TM models[19]. However, such experimental information is not available for a large majority of SPMR TMs, which can only be modelled from sequence.

The first characterized TM homodimer structures were of right-handed conformations and stabilized by the frequently occurring GXXXG-binding motif through putative weak CαH–O hydrogen bonds[15]. Corroborating these observations, modelling techniques incorporating a weak CαH–O bond potential allowed for accurately predicting native right-handed TMH homodimer (RH) structures in native TMH docking simulation[23] or grid search from ideal helices[24]. However, a large majority of TMH homo-oligomers does not bear GASright motifs (that is, small-XXX-small residue motif identified at right-handed parallel TMH dimers with small being either Gly, alanine or serine[25]) or are stabilized by a much larger diversity of physical interactions including Van der Waals (VDW), aromatic pi–pi, cation–pi and polar interactions[3,6,26–29]. Accurately predicting TMH oligomeric structures in absence of monomer TMH structures and of specific binding motifs identifiable from the sequence remains a daunting task, because of the large conformational space to be sampled in simultaneously folding and docking TMHs. Approximating TMHs as ideal helices usually cannot recapitulate TM dimer structures with near-atomic accuracy[30]. As demonstrated by several studies[31–34], because protein interactions are very sensitive to atomic details, designing selective inhibitors and predicting functional mechanism or mutational effects require high-resolution models (that is, typically structural divergence to native structures below 1.5 Å and a large fraction of predicted native contacts). A general method that predicts with high accuracy from sequence the structure of TMH oligomers with a wide range of TMH subunits, topologies, conformations and stabilizing interactions would therefore be of great interest but is currently lacking.

Rapid expansion of high-throughput sequencing and statistical methods distinguishing direct couplings from indirect correlations in residue sequence covariation patterns have led to high-precision residue contact prediction in protein structures[35–41]. Applying these predicted contacts as distance constraints in folding simulations considerably restrict the conformational space sampled and allowed for the reliable prediction of large polypeptide chain structures[42,43]. Co-evolutionary-based protein modelling approaches have recently been extended towards characterizing protein conformational diversity including the structure of transient or hidden functional states[44]. Residue contacts controlling important functional protein–protein interactions can also be identified in sequence co-evolution patterns of strongly interacting proteins[45,46]. When combined with protein surface chemical complementarities, such contacts can guide the prediction of both stable and transient protein–protein-associated structures[47]. However, applying this approach to homo-oligomers remains a challenge, because it relies on the ability to discriminate between intra- and inter-monomer contacts.

To address this problem, we here develop and implement in RosettaMembrane[23,34,48] EFDOCK-TM (Evolutionary-guided Fold and Dock of TransMembrane proteins), a protocol to accurately predict self-associated TM protein structures guided by co-evolutionary signals enriched in true inter-chain contacts using predicted TMH binding surfaces from sequence. We show that a very small number (less than three on average) of these selected contacts are necessary to accurately predict single-spanning TMH homo-oligomer structures with a wide range of size, subunit number, conformations and binding interactions. We apply EFDOCK-TM to blindly predict uncharacterized members of the KIT and fibroblast growth factor receptor (FGFR) family of RTK receptors and propose molecular interpretations of disease-occurring point mutations.

Results

Inter-chain-contacting residues co-evolve strongly. The structural interpretation of co-evolutionary signals in protein sequences folding into homo-oligomers has been a challenge, because they can in principle reflect both intra-chain and inter-chain residue contacts. When sequences fold into parallel helical homo-oligomers, discriminating between intra- and inter-monomer constraints becomes even more difficult because both encode short-range contacts (that is, between residues close in sequence) (Fig. 1a). We define a short-range intra-chain contact, an interaction between residues i and j close in sequence and on the same monomer chain A (i_A—$j_A \leq 8$). We define a short-range inter-chain contact, any interaction between residues i and j close in sequence but belonging to distinct chains A and A′ in a homodimer formed by two copies of the same monomer sequence (i_A—$j_{A'} \leq 8$) (Fig. 1a). Since most residues in helices are involved in short-range intra-chain contacts, we reasoned that the fraction of residue pairs forming additional inter-chain short-range interactions should be enriched in strongly co-evolving residues at the binding interface (Fig. 1b). To test this hypothesis, we first analysed the strength of residue covariations along the binding interface of all experimentally characterized TMH homodimer structures. Residue covariations were calculated using the widely used and benchmarked method EVfold[42]. As shown in Fig. 2a,b and Supplementary Table 1, the average direct interaction (DI) score (measuring the strength of co-evolution) and the fraction of strongly co-evolving residues (high DI score) calculated by EVfold were significantly higher for residue pairs

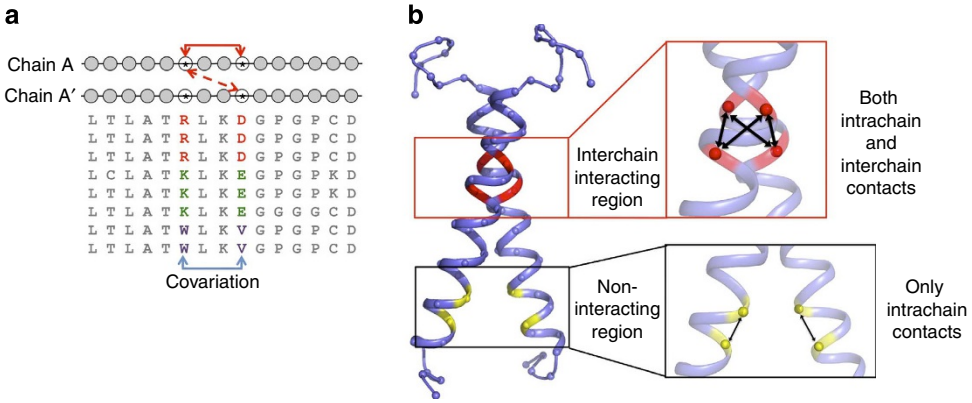


Figure 1 | Covariation in protein sequences folding as homo-oligomers reflect both intra- and inter-molecular contacts. (**a**) Covariation patterns in protein sequences forming homo-oligomeric structures can reflect both intra- (solid red arrow) and inter- (dash red arrow) monomer evolutionary constraints. (**b**) The strength of co-evolution signals between residues forming both intra- and inter-chain contacts (red) is expected to be stronger (thicker arrows) than pairs forming only intra-chain contacts (yellow, thinner arrows at the bottom).

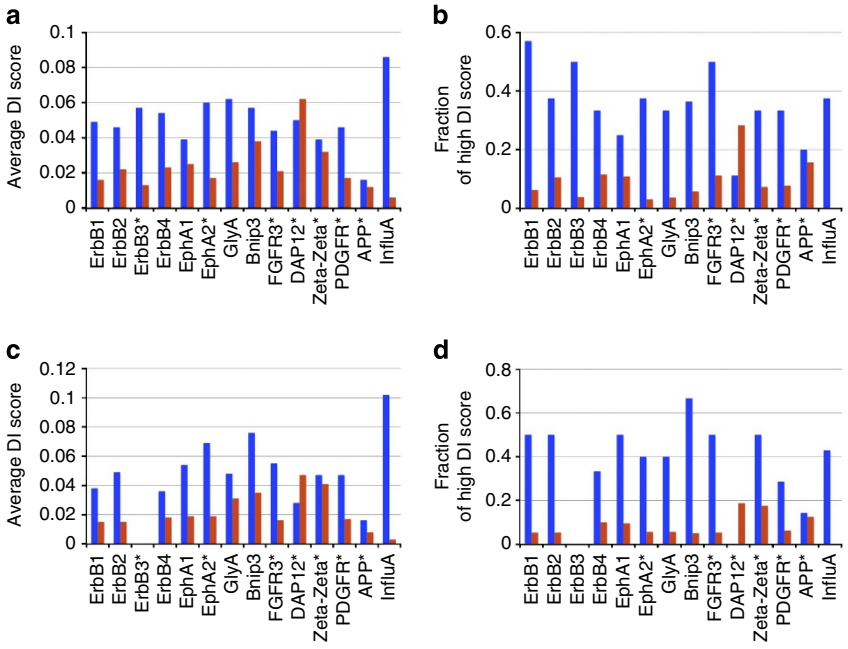


Figure 2 | Residue pairs making inter-chain contacts have stronger co-evolutionary signals. Comparison between direct interaction score (DI score) calculated by EVfold[42] for residue pairs interacting (blue) or not interacting (red) across the chains of the 14 TM homo-oligomers indicated on the *x* axis. Only short-range residue pairs (that is, separated by up to eight amino acids) that belong to either the true binding interface (**a**,**b**) or the predicted binding helical surface by LIPS[49] (**c**,**d**) are considered in the analysis (Methods). The average DI score for all residue pairs (**a**,**c**) or the percentage of high DI score pairs (**b**,**d**) in each interacting category is reported (Methods). DI score differences between the two populations of residue pairs are statistically significant (paired two-sample *T*-test *P* values $<4\times E-03$, see Supplementary Fig. 2).

involved in inter-chain contacts than for other residues present at the binding interface. Similar results were obtained when the same analysis was performed on a large data set of coiled-coil homodimers selected from diverse protein families (Methods; Supplementary Table 1). These results validate our hypothesis and indicate that co-evolutionary signals at homodimer helical binding interfaces are often stronger for residue pairs involved in inter-chain contacts than for pairs forming intra-chain contacts only.

Most single-spanning self-associating receptors, however, are largely uncharacterized with no information on their binding interfaces. To address this problem, we assessed whether TMH-interacting surfaces could be predicted from sequence by the method LIPid-facing Surface (LIPS)[49]. On average, 85% of the residues predicted by LIPS to be the least exposed to lipids were located at the experimentally characterized binding interface (Supplementary Fig. 1). As shown in Fig. 2c,d, when combined with the method EVfold, the method LIPS[49] was able to predict from sequence TMH surfaces that bear a large fraction of strongly co-evolving contacts. Both average DI score and fraction of strongly co-evolving contacts were found to be significantly higher for residues predicted to interact across monomers than for other residues along the predicted binding interface (Supplementary Table 1). Our results indicate that the prediction of interacting TMH surfaces by LIPS is accurate enough to identify a large majority of the strongly co-evolving residue pairs involved in inter-chain contacts.

Enrichment of intermolecular contacts in blind predictions. To take advantage of such signals in blind prediction of TMH

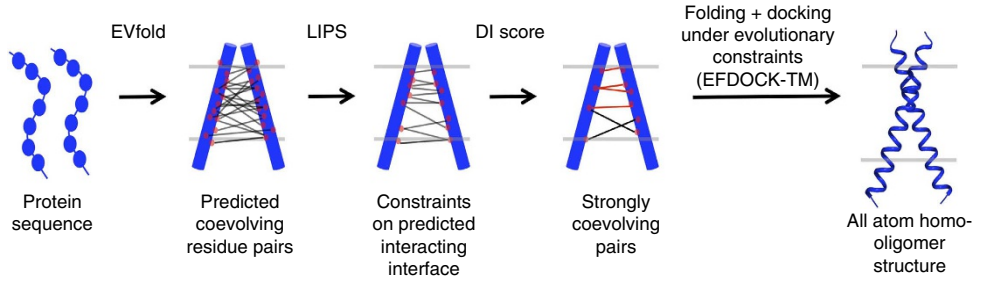


Figure 3 | Evolutionary-guided *de novo* structure prediction of TMH homo-oligomer using EFDOCK-TM. Stepwise selection of inter-chain contacts and structural models. Step 1: EVfold—this method[42] is used to predict co-evolving residue pairs likely in physical contacts in the protein structure (black lines). Step 2: LIPS—this method[49] is used to predict the helical surface that has the highest propensity to self-associate. Predicted contacts which do not belong to this surface are filtered out. Step 3: DI score—predicted contacts with low co-evolutionary DI scores are filtered out (true interacting sites are depicted in red, and false interacting sites are in black). Step 4: Convergence and energy-guided selection of models—folding and docking EFDOCK-TM simulations enumerating all possible combinations of constraints defined by subsets of predicted contacts are performed, leading to different levels of convergence. A total of 10,000 trajectories and final models are generated per simulation. Representative models are selected among the centres of the five largest families obtained by clustering the 10% lowest-energy models generated by the most-converged simulation.

homomeric structures, we implemented a computational framework that enriches co-evolving residue pairs in true inter-chain contacts (Fig. 3). Briefly, (1) contacts between pairs of co-evolving residues along the entire TM region are predicted from sequence using the method EVfold[42]; (2) contacting residues present on TMH surfaces predicted to face lipids using the method LIPS[49] are removed; (3) the remaining contacts with the strongest co-evolutionary signals, which are often enriched in inter-chain contacts (Fig. 2), are selected as pairwise distance constraints; (4) simulations using EFDOCK-TM, a protocol that we developed in this study and implemented within RosettaMembrane[23,34,48] for folding and docking TMH oligomers (Methods) are performed in parallel with randomly selected subsets of constraints until all possible combinations of constraints are enumerated. Representative low-energy models of TMH homo-oligomers were selected from the best converged simulations generating the least clusters (Methods).

To assess whether our protocol is effective at enriching evolutionary constraints in inter-chain contacts, we analysed the true positive rate of predicted interactions across the binding interfaces of all characterized TMH homo-oligomer structures as a function of each selection step described in Fig. 3. As shown in Supplementary Fig. 2, the rate of true inter-chain contacts increased from an average of 16 to 74% after filtering by LIPS-interacting surfaces and DI score. Except for the tetrameric M2 proton channel of influenza B virus (Inf B) and pentameric phospholamban (PLN), which have too few homologues for reliable detection of residue co-evolution, at least one true interacting site was successfully predicted among the top three strongly co-evolving residue pairs selected for guiding the folding simulations. The high rate of true positive inter-chain contacts in the selected constraints suggests that they may be effective at guiding the *de novo* structure prediction of TMH oligomers towards the native state.

Benchmark. EFDOCK-TM was tested on a data set composed of 17 characterized and published TMH homodimer structures and three representative higher-order homo-oligomeric structures determined by high-resolution NMR spectroscopy (Methods). All simulations were performed as in blind predictions: no information from native structures was used and the representative models were selected using the same combination of objective criteria (that is, by all-atom energy and cluster size, Methods).

Atomic accuracy prediction of right-handed TM homodimers. The experimentally determined interface structures of all six receptors forming right-handed TM homodimers (that is, glycophorin A, Bnip3 and the tyrosine kinase receptors ErbB1, ErbB2, ErbB4 and EphA1) were predicted with atomic accuracy using EFDOCK-TM guided by evolutionary constraints. Consistent with our constraint selection strategy, the best-converged simulations contained a relatively high rate of true inter-chain contacts (64%) and were selected for model accuracy analysis (Table 1; Supplementary Fig. 3). For all targets except ErbB1, the largest cluster of low-energy models in our selected simulations was a family of accurate 'near-native' models, indicating that a large fraction of the simulations converged towards the native conformation when guided by evolutionary constraints (Table 1; Supplementary Fig. 3). The representative selected models displayed atomic accuracy with an average interface Cα-root mean squared deviation (r.m.s.d.) of only 0.83 ± 0.38 Å to the well-defined regions of the NMR structure (Methods; Table 1). The backbone structures of the selected representative models in the interacting regions of the homodimers were all superimposable to those of the NMR centre model (Fig. 4a–f). Consistent with atomic accuracy predictions, 84% of the native contacts stabilizing the homodimer interfaces were predicted correctly (Table 1). Critical for right-handed homodimer interactions, inter-monomer weak Cα hydrogen bond networks were predicted with atomic accuracy and near-native interfacial side-chain conformations were also consistently observed in these models (Fig. 5; Supplementary Movies 1–6).

Of particular interest is the ErbB2 receptor sequence, which, despite bearing multiple sites for putative weak hydrogen bonds, is not stabilized by such polar networks in the available experimental structure. Modelling predicted inter-chain contacts as direct physical interactions (that is, allowed pairwise Cα distances from 3 to 6 Å as in NMR structure determination, 'hard potential'; Supplementary Fig. 3a) allowed large families of accurate models to be generated by EFDOCK-TM (interface Cα-r.m.s.d. of 0.56 Å, Table 1; Fig. 4c). Unlike alternative methods PREDDIMER[30] or CATM[24], EFDOCK-TM predicted correctly most side-chain conformations and residue contacts at this rather unusual right-handed interface (Supplementary Movie 3).

Near-atomic accuracy prediction of left-handed TM dimers. Our automated pipeline for constraint selection was also effective at enriching inter-chain contacts at left-handed homodimeric binding interfaces. Except for ErbB3 whose binding interface could not be predicted by LIPS, the most converged simulations were highly enriched in true positive inter-chain contacts

Table 1 | Accurate evolutionary-guided *de novo* prediction of TMH homo-oligomeric structures using EFDOCK-TM.

Uniprot name	EFDOCK-TM interface Ca-r.m.s.d.*	EFDOCK-TM native contact (%)†	PREDDIMER interface Ca-r.m.s.d.‡	PREDDIMER native contact (%)§	CATM interface Ca-r.m.s.d.‖	HFDOCK-TM interface Ca-r.m.s.d.¶	HFDOCK-TM native contact (%)#
ErbB1	1.36 (0.90)	67	1.38	100	0.78	1.03	100
ErbB1 alt**	1.03 (ND)	89	1.22	60	NA	3.70	0
ErbB2	0.56 (2.22)	83	1.93	17	2.43	1.98	33
ErbB3**	4.55 (4.43)	0	4.31	0	NA	4.50	0
ErbB4	0.80 (0.68)	100	1.32	67	0.81	0.69	100
EphA1	0.73 (1.03)	91	1.65	55	1.26	1.02	91
EphA1 alt	1.62 (ND)	100	2.53	25	1.48	1.54	100
EphA2**	1.45 (1.58)	80	2.08	20	NA	4.77	20
GlyA	0.85 (1.00)	75	2.58	25	1.11	0.89	75
Bnip3	0.56 (0.61)	67	1.71	22	0.56	NA	NA
Bnip3 alt1	0.52 (ND)	86	1.84	28	NA	NA	NA
Bnip3 alt2	0.50 (ND)	75	1.59	17	NA	NA	NA
FGFR3**	1.99 (4.71)	50	4.66	0	NA	2.01	60
DAP12**	1.24 (3.82)	64	3.74	0	NA	4.31	0
Zeta-Zeta**	0.53 (1.65)	67	1.95	0	NA	NA	NA
PDGFR**	0.74 (1.02)	91	1.89	36	NA	3.52	0
APP**	1.16 (4.11)	70	4.45	0	NA	4.38	30
APP alt	2.33 (ND)	40	4.36	0	NA	3.59	60
Dimers: transmembrane region and juxtamembrane region							
ErbB1	0.81	100	NA	NA	NA	NA	NA
APP**	0.99	60	NA	NA	NA	NA	NA
Higher-order oligomers: transmembrane region							
M2 InfluA	0.70 (1.26)	82	NA	NA	NA	0.83	71
M2 InfluB	2.40	100	NA	NA	NA	2.40	100
PLN	0.80	87	NA	NA	NA	NA	NA

All TM dimers	**R.m.s.d.**	**Native contact (%)**	**All TM dimers††**	**R.m.s.d.**	**Native contact (%)**
EFDOCK-TM	1.22 ± 0.93	73 ± 24	EFDOCK-TM	1.04 ± 0.52	76 ± 17
PREDDIMER	2.51 ± 1.21	26 ± 28	PREDDIMER	2.40 ± 1.16	28 ± 28
HFDOCK-TM	2.71 ± 1.54	51 ± 40	HFDOCK-TM	2.57 ± 1.51	56 ± 38
Left-handed dimers	**R.m.s.d.**	**Native contact (%)**	**Left-handed dimers‡‡**	**R.m.s.d.**	**Native contact (%)**
EFDOCK-TM	1.52 ± 1.21	63 ± 27	EFDOCK-TM	1.14 ± 0.45	71 ± 14
PREDDIMER	3.04 ± 1.39	15 ± 23	PREDDIMER	2.86 ± 1.39	17 ± 24
HFDOCK-TM	3.88 ± 0.94	16 ± 23	HFDOCK-TM	3.78 ± 0.98	18 ± 24
Right-handed dimers	**R.m.s.d.**	**Native contact (%)**	**Right-handed dimers§§**	**R.m.s.d.**	**Native contact (%)**
EFDOCK-TM	0.97 ± 0.58	80 ± 18	EFDOCK-TM	0.83 ± 0.38	84 ± 13
PREDDIMER	2.09 ± 0.90	36 ± 30	PREDDIMER	1.84 ± 0.45	40 ± 29
CATM	1.20 ± 0.62	NA	CATM	1.20 ± 0.62	NA
HFDOCK-TM	1.53 ± 1.01	80 ± 26	HFDOCK-TM	1.19 ± 0.48	83 ± 26

NA, not applicable; ND, not determined; PLN, phospholamban; r.m.s.d., root mean squared deviation; TM, transmembrane; TMH, transmembrane protein helical structure.
For the method EFDOCK-TM, 'interface Ca-r.m.s.d.' numbers are reported for simulations performed with or without (parentheses) evolutionary constraints. The results are compared with those obtained by the methods PREDDIMER[30], CATM[24] and HFDOCK-TM (identical to EFDOCK-TM but with homologue template-derived constraints). Averages are provided as the mean value ± s.d. R.m.s.d. values are in Angstrom.
*The lowest interface r.m.s.d. among the centres of the five largest clusters of EFDOCK-TM models.
†The percent native residue–residue contact correctly predicted in the selected EFDOCK-TM models.
‡The lowest interface r.m.s.d. among the models predicted by the method PREDDIMER.
§The percent native residue–residue contact correctly predicted in the selected PREDDIMER models.
‖The interface r.m.s.d. of the models predicted and reported by the method CATM.
¶The lowest interface r.m.s.d. among the centres of the five largest clusters of HFDOCK-TM models using structural homolog instead of evolutionary-derived constraints (HFDOCK-TM).
#The percent native residue–residue contact correctly predicted in the selected HFDOCK-TM models.
**Left-handed TMH homodimers; others are right-handed TMH homodimers.
††The average interface r.m.s.d. and % native contact of all targets, except ErbB3 for which the native binding interface was not correctly predicted by LIPS.
‡‡The average interface r.m.s.d. and % native contact of all left-handed targets, except ErbB3.
§§The average interface r.m.s.d. and % native contact of all right-handed targets, except APP alt for which the native binding interface was not correctly predicted by LIPS.

(average true positive rate of 83%), thereby generating large clusters of accurate models (Table 1; Supplementary Figs 2 and 3). All representative models had an interface r.m.s.d. smaller than 2 Å with an average interface r.m.s.d. of only 1.14 ± 0.45 Å over interface regions spanning an average of 19 amino acids (that is, covering almost the entire TM helix). The backbone structures of our selected representative models were all superimposable to their corresponding NMR structures (Table 1; Fig. 4g–l) and most buried inter-chain packing interactions and 71% of interfacial residue contacts were predicted correctly (Fig. 6; Supplementary Movies 7–11). Of particular interest were DAP12 and zeta–zeta homodimers, which belong to the immune receptor family, participate in the assembly of large hetero-oligomers and are stabilized by several polar interactions. The representative models of zeta–zeta recapitulated all ion pair and hydrogen bond interaction networks with atomic accuracy (interface Cα-r.m.s.d. of

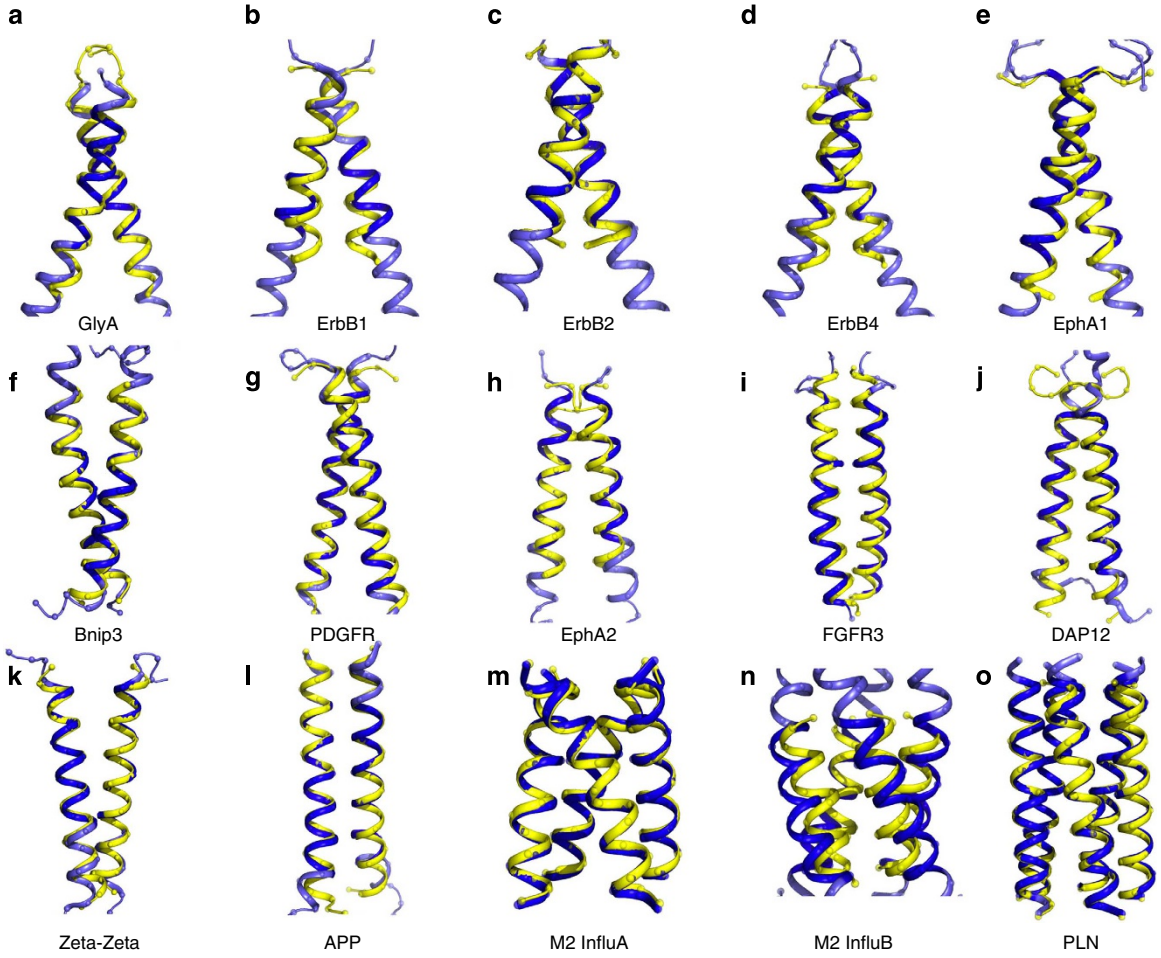


Figure 4 | Consistent prediction of near-native homo-oligomeric TMH structures. Backbone superposition between the centre NMR model (blue) and the representative EFDOCK-TM model (yellow) of the following receptors: GlyA (**a**), ErbB1 (**b**), ErbB2 (**c**), ErbB4 (**d**), EphA1 (**e**), Bnip3 (**f**), PDGFR (**g**), EphA2 (**h**), FGFR3 (**i**), DAP12 (**j**), Zeta-zeta (**k**), APP (**l**), M2 influA (**m**), M2 influB (**n**) and PLN (**o**).

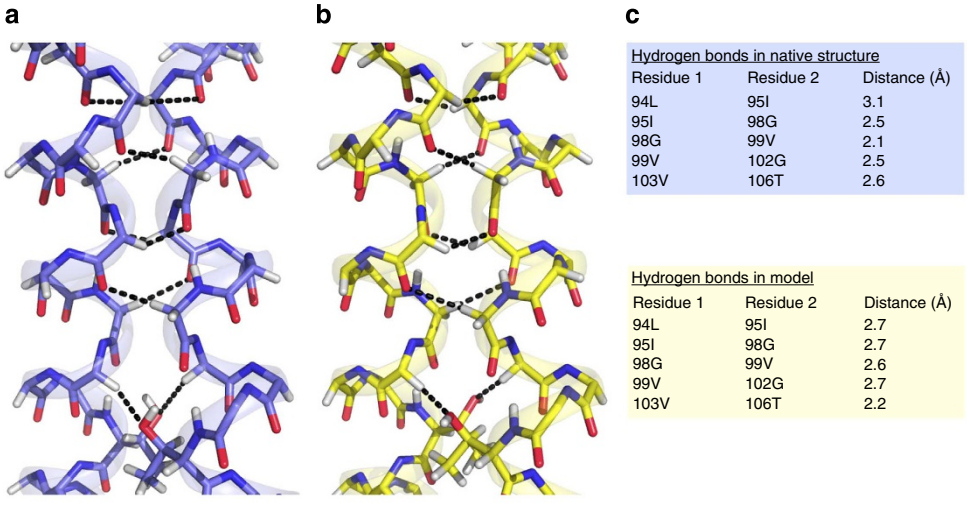


Figure 5 | Atomic accuracy in right-handed homodimer structure prediction. (**a**) Backbone representation of glycophorin A centre NMR model (blue). (**b**) Backbone representation of glycophorin A representative EFDOCK-TM model (yellow). The inter-chain weak hydrogen bond network is highlighted (black dashed lines). The carbon C-alpha root mean square deviation (that is, Cα-r.m.s.d.) of the model over the binding interface region (residues 94–107) is reported at the bottom. (**c**) Weak hydrogen bond distances in Angstrom are reported for both the native and modelled structures.

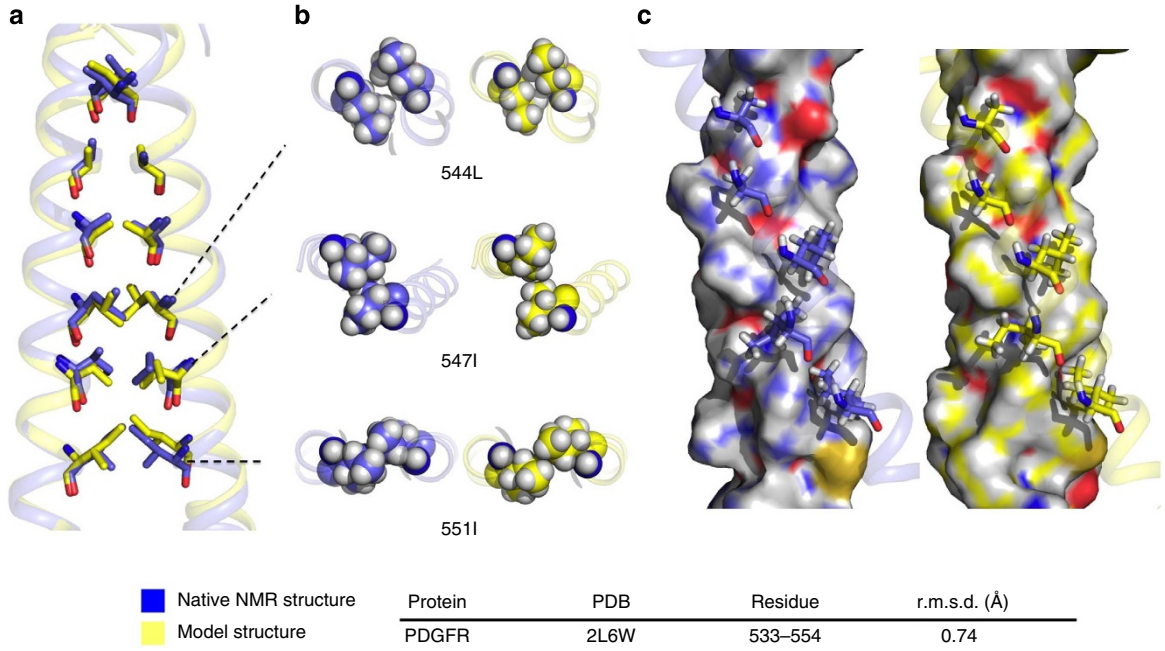
only 0.5 Å with up to 100% predicted native contacts for the lowest-energy model; Supplementary Movie 12). While the near-native models of DAP12 correctly identified Thr 16 and Asp 20 as hot-spot binding sites at the homodimer interface, they were stabilized by low-energy but non-native polar interaction networks (Supplementary Movie 9). Detailed inspection of the NMR structure revealed that the native conformations of these residues do not form hydrogen bonds according to Rosetta-Membrane energy function, which may explain why such a configuration was not selected in our simulations.

Modelling TM–JM domain improves the TM structure accuracy. JM regions (adjacent to TM domains) in SPMRs have been shown to modulate the conformations accessible to the TM domains and to play important roles in propagating signals[19,50]. Therefore, we tested the ability of EFDOCK-TM to improve the prediction of the TM region by folding and docking simultaneously TM and JM regions of the characterized Erbb1 and APP TM–JM dimers. For both receptors, simulations were performed with predicted interacting sites in the TM domain and without any constraints in the JM region. Including the JM region significantly enriched the simulations in near-native models for both receptors (that is, additional or higher-ranked clusters of near-native models, Table 1; Supplementary Fig. 3). Interestingly, APP's JM domain prevents the GSA motif located near the N-terminal end of the homodimer to interact, which is a low-energy non-native state generated in our simulations of TM homodimers (Supplementary Fig. 4).

Higher-order TMH homo-oligomers are predicted accurately. We then tested the ability of our approach to predict the structure of higher-order symmetric oligomers, that is, the tetrameric domains of the M2 influenza A and B viroporins and the pentameric domain of PLN. M2 influenza A and PLN were predicted with atomic accuracy (interface r.m.s.d. $\leq$0.80 Å) (Table 1). Although M2 influenza B and PLN exhibit almost no sequence variation within their homologues, EFDOCK-TM identified the native conformation without constraints. Our representative models of all three targets superimposed well to the experimental structures (Fig. 4m–o) with more than 80% of native interhelical packing contacts predicted correctly (Table 1; Supplementary Movies 13 and 14).

Evolutionary contacts improve TM structure prediction. To assess whether evolutionary constraints improve the accuracy of the predictions, identical simulations were performed without constraints. The prediction accuracy of three left-handed and one right-handed dimers (FGFR3, DAP12, APP and ErbB2) markedly increased upon addition of evolutionary constraints (the largest clusters of left-handed models generated without constraints were all non-native; Table 1). The number of near-native models among the five largest clusters also largely increased, especially for left-handed targets (Supplementary Fig. 5).

Alternative experimental TM structures are well predicted. Conformational regulation is a hallmark of membrane receptor function, and several TMH homo-oligomers were characterized in different conformational/functional states. EphA1 TM homodimer structures were solved at low- and high-pH conditions and differ significantly (Cα-r.m.s.d. of 2.7 Å)[51]. The top two lowest-energy clusters among the five largest recapitulated the low- and high-pH structures with Cα-r.m.s.d. of 0.7 and 1.6 Å, respectively (Supplementary Movie 15). Therefore, both conformations of EphA1 would be accurately predicted and selected in a blind prediction. The EGFR TM sequence bears an N- and a C-terminal GAS motif, suggesting two possible binding modes. Indeed, the isolated EGFR TM homodimer was solved in two very different conformations, each stabilized by one of the two motifs (pdb codes: 2M20 and 2M0B; Methods). Experimental and computational studies on the full-length receptor suggest that these conformations represent an active and inactive state occupied by the TM region during receptor signalling[8,9,21].



Protein	PDB	Residue	r.m.s.d. (Å)
PDGFR	2L6W	533–554	0.74

Figure 6 | Atomic accuracy in left-handed homodimer structure prediction. (**a**) Superposition of PDGFR centre NMR model (blue) and representative EFDOCK-TM model (yellow) with buried side chains in sticks. (**b**) Local packing interactions are highlighted for a few buried residues in spheres along the binding interface. (**c**) Surface representation of the binding interface of one monomer highlighting specific binding grove geometries. The Cα-r.m.s.d. of the model to the native structure over the binding interface region (residues 533–554) is reported at the bottom.

Both functionally relevant conformations were accurately predicted by one of the top five clusters of EFDOCK-TM models, with Cα-r.m.s.d. of 1.4 Å (67% native contacts) and 1.0 Å (89% native contacts) (Supplementary Movies 1, 2 and 16).

The tetramer M2 proton channel of influenza A virus (M2A) was crystallized in both open and closed states[52,53]. Two of the top five clusters including the largest one accurately captured the closed state of M2A, and the open state was predicted by the second and eighth largest clusters with a Cα-r.m.s.d. of 2.90 Å over the TM region, consistent with a lesser-packed and higher-energy open conformation (Supplementary Fig. 6). APP was experimentally observed in two very different conformations, right-handed and left-handed structures which, unlike PREDDIMER[30], were both predicted correctly by EFDOCK-TM (Table 1). Unlike alternative techniques, our method could also predict multiple experimentally observed conformations of BNIP3 with atomic accuracy (Table 1).

EFDOCK-TM outperforms alternative techniques. The accuracy of our predictions considerably exceeded that of the method PREDDIMER, which models TM dimer interfaces using ideal helices[30], especially for left-handed dimers (Cα-r.m.s.d. of 2.86 Å and 17% correctly predicted contacts compared with 1.14 Å and 71% for EFDOCK-TM; Table 1). CATM is another technique developed to predict the structure of the GAS motif containing right-handed TM homodimers[24]. Unlike EFDOCK-TM with evolutionary constraints, CATM was not able to predict and select near-native models of the difficult target ErbB2 (Table 1). With an average Cα-r.m.s.d. of only 0.83 Å for right-handed TM dimers, EFDOCK-TM outperformed CATM (Cα-r.m.s.d. of 1.20 Å).

EFDOCK-TM outperforms homology-modelling approaches. In principle, accurate structural models can be generated from close homologue structures using comparative modelling techniques[31,34]. Here we assessed whether available structures are sufficient to derive accurate sequence alignments, templates and constraints to complement the evolutionary-derived EFDOCK-TM approach. Consistent with the very low number of available TM oligomer structures, close structural homologues sharing at least 30% sequence identity and aligning well with target sequences were found for only four receptors (Erbb4, EphA1, influA and influB) in our benchmark (Supplementary Table 2). Consequently, structural accuracy of the threaded templates was low with an average Cα-r.m.s.d. of only 3.4 Å. When relaxed using RosettaMembrane homology-modelling techniques[34], only four templates could be refined to high accuracy (Supplementary Table 2). While interhelical interactions derived from these templates were of high accuracy for right-handed TM dimers, those derived for left-handed-associated TMs contained only 14% true positive contacts instead of 93% for EFDOCK-TM (Supplementary Fig. 7). When these template-derived contacts were used as constraints in our folding and docking simulations (HFDOCK-TM), right-handed and left-handed models displayed interface r.m.s.d. of 1.2 and 3.8 Å instead of 0.8 and 1.2 Å for EFDOCK-TM models, respectively (Table 1). These results indicate that while associated TM structures may be accurately modelled using constraints from close structural homologues, the scarcity of TMH oligomeric structures currently prevents the wide application of our HFDOCK-TM approach.

EFDOCK-TM models are suitable for protein engineering. To assess whether EFDOCK-TM models provide structural templates accurate enough for rational protein design applications, we performed native sequence recovery calculations. If native interactions are optimized for stability of the binding interface, then the native sequence should be recapitulated in design calculations. Since the selection of residues in design calculations is very sensitive to the structural environment, native sequence recovery can be used as a strong indication of the accuracy of structural models[34]. To identify which sites are likely optimized for stability, we redesigned the experimental NMR structures and then performed the same calculations using EFDOCK-TM or PREDDIMER models as starting structures. Consistent with the higher accuracy of our EFDOCK-TM models, 86% of the native residues recapitulated using the NMR structures were also recapitulated using the EFDOCK-TM structures compared with only 47% using PREDDIMER models (Supplementary Fig. 8). These results indicate that the EFDOCK-TM models should be accurate enough to guide rational protein design applications, for example, to redesign the binding affinity/specificity of TMH oligomers or to design inhibitors regulating receptor associations and functions.

Blind predictions of disease-associated receptor variants. The FGFRs and the stem cell growth factor receptor c-Kit are tyrosine kinase receptors binding to the largest family of growth factor ligands, for which a large number of point mutations have been associated with various diseases[1]. A few of these mutations are located in the TM and JM regions of these receptors[1] and are discussed below.

Mutations Y372C on the TM–lipid interface and a TM mutation C379R in FGFR1 were found to be strongly associated with osteoglophonic dysplasia, a rare genetic disease characterized by abnormal bone growth. In all five representative FGFR1 models, residues 372 and 379 were predicted to reside on the interacting interface (Supplementary Fig. 9).

Two mutations S372C and Y375C were identified in FGFR2 in patients with Beare-Stevenson syndrome, an autosomal-dominant condition characterized by the furrow skin disorder. Y375C is located in the TM region, while S372C belongs to the extracellular JM/TM linker. The largest cluster of FGFR2 models formed a nearly parallel left-handed homodimer with closely packed S372 and Y375 at the interface (Supplementary Fig. 9). The mutation G384R reported in patients with craniosynostosis is located near the middle of the TM region but predicted to be lipid-exposed in our top-ranked models.

A single TM mutation G388R in FGFR4 is strongly associated with tumour cell motility. This allele is highly abundant in patients with advanced tumour metastasis. In the two largest clusters of models, G388 formed either weak hydrogen bonds or Van der Waals contacts at the dimer interface (Supplementary Fig. 9).

The F522C mutation in the TM domain of c-Kit leads to ligand-independent autophosphorylation and is associated with Mastocytosis. Residue F522 is located at the TM/JM junction of the homodimer and adopt a wide diversity of conformations in our simulations, that is, buried at the binding interface or exposed to the lipid hydrophobic/headgroup interface. Another mutation A533D has been related to diffuse cutaneous mastocytosis. In four among the five top-ranked models, A533 forms contacts across the binding interface at the middle of the TM region of the predicted homodimers (Supplementary Fig. 9).

Among the eight mutated sites analysed, six of them were located along the receptor predicted binding interface in the top-ranked models, suggesting a critical role of these positions in controlling protein inter-monomer interaction, associations and function. The two mutation sites predicted to be exposed to the

lipids in a fraction of our models were G384R in FGFR2 and F522C in c-Kit. While the G384R mutation may not directly interfere with the interhelical interactions, the presence of an arginine residue in the middle of a TM domain may perturb the membrane insertion of the receptor (due to hydrophobic mismatch that cannot be compensated by lipid deformation when arginine is located at the centre of the lipid membrane) and result in non-functional receptor variants. The F522C mutation may perturb the interhelical interaction or the optimal orientation of the helices in the membrane and affect the regulation of the receptor.

Discussion

SPMR functioning as oligomers represent a large fraction of the human membrane proteome and are involved in crucial functions which when deregulated can lead to severe diseases[1,2,10]. Experimental evidence indicates that TM and JM domains of many SPMRs participate to signal propagation across the membrane by changing orientation or oligomerization states and by coupling extracellular to cytoplasmic domains[1,8,9,50]. To properly regulate signal transduction, TMHs must be able to switch between states without large energy penalties[1,7]. Consequently, TM associations are usually weaker, stabilized by fewer interhelical contacts (Supplementary Fig. 10), conformationally more flexible in single-pass than in multi-pass membrane proteins and very challenging to characterize structurally. In addition, little experimental data on TM–TM or JM–JM interactions is currently available, making the structure prediction of SPMR oligomers also a real challenge. To address this problem, we have developed EFDOCK-TM, a method to predict TMH homo-oligomeric structures from sequence guided by co-evolutionary constraints. By combining TMH-binding surface and residue contact predictions from sequence, we show that sequence co-evolutionary patterns that reflect both intra- and inter-monomer constraints in homo-oligomers can be considerably enriched in inter-chain contacts (Fig. 3; Supplementary Fig. 2). When combined with an efficient technique that we implemented to fold and dock TMH oligomers, a small number (2.6 on average) of selected predicted contacts are sufficient to accurately predict the native structure for 21 of 22 TMH homo-oligomers (average Cα-r.m.s.d. of 1.0 Å; Table 1; Figs 4–6; Supplementary Movies 1–16). Our benchmark includes right- and left-handed homodimers and representative higher-order oligomers spanning a wide range of conformations and binding interactions. By contrast, accurate predictions of TMH homo-oligomers so far required numerous experimental constraints[19,22] or were restricted to a limited subclass of right-handed homodimers (RHs) stabilized by GXXXG motifs[24]. Left-handed TM dimers (LHs) are particularly challenging to model (Table 1) and the near-atomic accuracy predictions achieved by EFDOCK-TM are unprecedented.

While EFDOCK-TM without constraints was able to predict near-native structures of most RHs, the addition of constraints was critical to consistently predict LH structures with high accuracy (Table 1; Supplementary Fig. 3). Most RH structures in our benchmark are mainly stabilized by backbone–backbone contacts through weak hydrogen bonds, while LH-binding interfaces often involve greater number but weaker VDW/aromatics contacts between large side chains. Since the latter samples many more conformational degrees of freedom than backbone atoms, side-chain/side-chain contacts may be more difficult to simultaneously optimize. Also, the strong orientational dependencies of hydrogen bond energies make these interactions more specific and may greatly facilitate the search for low-energy native conformations of RH compared with LH structures. Future work will be needed to assess how such differences in chemical interactions between LH and RH structures impact the conformational energy landscape relevant to their function.

Discriminating between intra- and inter-chain contacts from co-evolutionary sequence patterns of self-associating proteins is challenging because it requires a systematic and accurate structural interpretation of constraint violations, which can become a daunting task when simulations are performed with a large number of contacts. Interestingly, as shown in Supplementary Fig. 2, it is the synergistic combination and not each individual constraint selection step alone (EVfold, LIPS and high DI score at predicted binding surfaces) that is effective at enriching for inter-chain contacts. With such a high contact precision, only a small number of constraints (average of 2.6) are needed to select sufficient true positive inter-chain contacts (average of 1.9 per target). Because only very few selected constraints are necessary, accurate prediction may be achievable for a large number of TMH homo-oligomers even when, as observed for many eukaryotic proteins, only relatively few homologue sequences are available (as low as $3 \times L$ with L: modelled protein length). Close structural homologues could be identified for a small number of targets, and the constraints and resulting models derived from these homologue templates (HFDOCK-TM) were of similar accuracy than those generated by EFDOCK-TM (Table 1). However, EFDOCK-TM largely outperformed HFDOCK-TM for most of the targets for which close structural homologues were not available (Table 1). Therefore, while HFDOCK-TM may be a useful approach when close structural homologues can be identified, EFDOCK-TM should be more widely applicable for characterizing the self-associations of eukaryotic membrane receptors.

While full-length receptor high-resolution structures have not been characterized to date, current biochemical evidences suggest that TM–TM interactions determined from isolated TM domains and those in full-length receptors are similar[8,9,19–21]. For example, seminal studies on the EGFR indicate that the TM domain during receptor signalling adopts two conformations (i.e., inactive and active) that are also observed by NMR spectroscopy on the isolated TM region[8,9] (Methods). Therefore, these data support a model where TM sequences encode an ensemble of functionally relevant conformations that are 'selected' by extramembrane domains and ligands during signalling. Because inter-chain contacts at the TMH-binding interface may have evolved to stabilize multiple conformations[8,9,16,18], performing simulations using a soft constraint potential allows EFDOCK-TM to populate multiple minima compatible with similar binding surfaces. Alternative conformations may correspond to functionally relevant states of SPMRs as reflected by our ability to predict multiple TM conformations of the EphA1, EGFR, Bnip3 and APP receptors.

As demonstrated in previous studies[34,54] and in our native sequence recovery calculations (Supplementary Fig. 8), near-atomic structure accuracy, which allows a majority of binding contacts to be accurately predicted (Table 1), is sufficient to design accurate physical interactions. Therefore, engineering TMH inhibitors binding with high affinity and selectivity to receptor-binding surfaces modelled using our method should be feasible. This would extend computed helical anti-membrane protein (CHAMP)-based approaches[1,55] to target a large diversity of TMHs for which no structural homologues or specific sequence/structure motifs can be identified.

In summary, we have developed a general method that can accurately predict from sequence the structure of a large diversity of TMH homo-oligomers, which to our knowledge is unprecedented. Our approach may prove useful for uncovering the

determinants of molecular recognition and regulatory mechanisms of SPMR signalling.

Methods

Selection of targets for the benchmark data set. Single-TM helix homodimers and representative homo-oligomers with solved experimental structure were selected: right-handed (PDB code: 2M20 (ref. 8), 2JWA[16], 2LCX[56], 2K1K[51] and 2K1L[51], 1AFO[15], 2J5D[57], 2KA1 (ref. 26), 2KA2 (ref. 26) and 2LZ3 (ref. 58)), left-handed TM homodimers (PDB code: 2M0B, 2L9U[59], 2K9Y[28], 2LZL[27], 2L34 (ref. 60), 2HAC[3], 2L6W[61] and 2LOH[62]), tetramers (PDB code: 3LBW M2A closed state[52], 3BKD M2A open state[53] and 2KIX[63]) and pentamer (PDB code: 2KYV[64]). Modelled regions included both residues in the TM and residues in the water–lipid interface regions. The JM region of three targets (amyloid precursor protein, ErbB1 and PLN) was also included in the modelling. The JM regions of ErbB1, APP and PLN were defined between residues 677–690, 686–698 and 1–22, respectively.

Blind prediction of disease-related receptor variants. Four receptors from the tyrosine kinase family (FGFR1, FGFR2, FGFR4 and c-Kit) with no experimental structures were selected that have multiple reported disease-causing mutational variants in the human population[1,65].

Interacting site prediction from protein co-evolution. Co-evolution-based contacts were predicted using EVfold[42], which is based on an inverse covariance matrix-based statistical model. Multiple sequence alignment (MSA) for each candidate protein was performed using HHblits[66] searching against the entire Uniprot database with a range of different *E*-values. Full-length sequences were used directly as query for small targets (full-length $\leq$350 amino-acid residues) but were truncated for large targets to allow for optimal alignment of the TM regions. Truncated sequences consisted in the TM region flanked by one conserved extramembrane globular domain on each side of the TM region. Sequences with lower than 50% TM region coverage were filtered out. Then, following the published protocol of the method EVfold[42], the *E*-value that generated the largest number of sequence homologues with well-aligned TM region was selected for each target. This procedure ensured that the selected MSAs correspond to the optimal tradeoff between the number of sequences in the alignment and the coverage of the region of interest. To infer inter-residue contacts from sequence covariation, the co-evolution-based direct interaction score DI_{ij} between each pair of residues i and j was calculated using EVfold with default settings (Supplementary Methods). All residue pairs were ranked by their co-evolutionary coupling strength and filtered using the following criteria: (1) both residues belong to the TM region; (2) due to the intrinsic symmetry of homo-oligomers, only 'short-range' residue pairs separated in sequence by eight positions or less were considered; (3) both residues belong to the helical surface predicted by LIPS[49] to be the least likely lipid-exposed, and the most likely the interacting surface; (4) only residue pairs with strong co-evolutionary signals (that is, $DI_{ij} \geq 0.1$) were selected. For targets with no $DI_{ij} \geq 0.1$, the top three constraints ranked by DI score were used in the simulation (Supplementary Fig. 3).

Interacting helical surface and site prediction by LIPS. For each target, MSAs were selected using *E*-values as low as those chosen for EVfold that guaranteed a query sequence coverage of at least $0.5 \times L$ (*L*: query protein length). From the selected MSAs, sequences with no gaps within the TM region were extracted and used as input to the method LIPS[49], which splits a helix into seven overlapping surfaces and predicts interacting TMH surfaces based on residue lipophilicity and sequence entropy. For each target, the surface with the lowest LIPS score was selected as the predicted binding interface. If the two top-ranked LIPS surfaces did not differ by more than 0.5 LIPS score, the surface that had the lowest lipophilicity score (which is the least sensitive to the alignment) was selected. The selected helical surface was applied to filter co-evolutionary constraints identified using EVfold. A set of 'LIPS'-predicted interacting sites was also derived by selecting the residues with the lowest LIPS score on the predicted interacting surface. To compare and select simulations based on convergence, the number of residues selected by the LIPS score was equal to the number of EVfold constraints for the same target. Since LIPS alone does not provide information on interactions between different residues, loose distance constraints were simply defined between the same residues on each monomer (see below).

Co-evolutionary signal analysis at binding interfaces. Co-evolutionary DI scores were compared for residue pairs making or not inter-chain contacts along the true or the LIPS-predicted binding interface of all TMH homodimers in the benchmark. Residue pairs were defined as true interacting if the distance between at least one of their respective heavy atoms was smaller or equal to 5 Å in the NMR centre model. To mimic the selection of the helix surface by LIPS, the true binding interface was defined by extrapolating to the entire helix the largest solid angle between the interacting residues and the helix centre axis. For tetramers, the above-mentioned procedure would select the entire helix as the interacting surface so only the subset of residues lying in the centre region of the tetramer, that is, the hexahedron enclosed by the four helical axes, were selected.

Residue pairs present at the binding interface were separated into two categories: inter-chain-interacting sites, and inter-chain non-interacting sites. Two different sequence separation thresholds were applied for residue pairs i and j: (1). $|i—j| \leq 8$, which corresponds to the short-range window used to select predicted interacting sites; (2) $|i—j| \leq 4$, which allows to compare residue pairs forming inter-chain contacts to residue pairs always involved in intra-chain contacts. A paired two-sample *T*-test was used to evaluate the significance of differences in average DI score and high DI score percentage between the two sets of residue pairs in all targets. The high DI score threshold was defined for each target as the lowest DI score among the top three most strongly co-evolving residue pairs for that target. ErbB3 was not included for the predicted LIPS-binding interface because LIPS predicted a non-native binding surface, which did not bear any true interacting residues. In Fig. 2 panels a and c, the average DI score of both sets of residue pairs for InfluA were substracted by 0.1 to facilitate illustration.

The same analysis was extended to a representative set of 103 water-soluble coiled-coils. The 103 coiled-coil structures were selected from the CC+ database (http://coiledcoils.chm.bris.ac.uk/ccplus/search/) by searching coiled-coil homodimers with more than 14 residues, and less than 50% redundancy. DI scores were calculated using EVfold as described above. Coiled-coil binding interface was predicted using MULTICOIL[67]. To analyse the statistical significance of DI score differences between two groups of contacts, a paired two-sample *T*-test was performed. This test is justified because the sample sizes (that is, number of proteins) of the two groups being compared are identical and there is a one-to-one correspondence between the values in the two samples.

Evolutionary-guided structure prediction (EFDOCK-TM). To predict *de novo* the structure of TMH homo-oligomers, we developed EFDOCK-TM, starting from a protocol to fold and dock symmetric water-soluble oligomers[68]. Simulations typically start from a random symmetric coarse-grained conformation generated by fragment insertion, where torsion angles of randomly selected consecutive three- or nine-residue fragments are replaced by those of protein homologues with known structures. For every 1 in 10 fragment insertions, rigid body backbone movements and docking arrangement perturbations are applied simultaneously to one monomer and cloned to the other monomer. Then, all-atom refinement of the coarse-grained models is performed by sampling side-chain conformational degrees of freedom and applying restrained backbone perturbations. The protocol was modified as follows to increase the efficiency of conformational sampling and the accuracy of the membrane protein structural models: (1) TM regions of each protein predicted using OCTOPUS[69] (http://topcons.cbr.su.se/index.php?about=octopus) are inserted into a membrane plane object approximating the lipid membrane prior to fragment insertion. (2) Models are scored using a coarse-grained or all-atom energy function developed for membrane proteins[23]. Unless stated otherwise, predicted interacting sites were implemented as 'soft' distance constraints between c-alpha atoms characterized by an equilibrium distance (d_{eq}) and s.d. of 6.5 (2.5 Å) for co-evolutionary constraints and 7.0 (4.0 Å) for LIPS constraints. Because they are not based on direct contact information, LIPS constraints are less precise in nature so looser distance constraints were applied. Any distance outside the range defined by $d_{eq} \pm$ s.d. is penalized using an harmonic potential. While guiding the simulation towards the native state, 'soft' constraints still allow the TMH oligomeric structure to be refined and selected by the physical model underlying RosettaMembrane's all-atom energy function.

For all targets, at least two types of simulations were performed: using all co-evolution constraints, and LIPS constraints. If three or more co-evolutionary constraints were selected for a given target, simulations using all possible randomly selected subsets of constraints (at least two) were also performed. The most converged simulation (that is, generating the least clusters of low-energy models) was selected for model analysis. For each simulation, 10,000 trajectories and models (which guaranteed convergence of the simulations) were generated, and the lowest-energy 10% models were clustered along the TM region using the Rosetta clustering method with a cluster radius of 1.5 Å. As in blind predictions, the representative model selected for each target was the most accurate among the centres of the five most populated clusters. Following another blind prediction selection strategy, the clusters were also ranked by the all-atom energy of the cluster centre. The model for the right-handed alternative conformation of the target APP was selected as one of the top five lowest-energy cluster centres.

Structure homology-guided structure prediction (HFDOCK-TM). To predict the structure of TMH homo-oligomers using constraints derived from homologues, we first performed sequence/structure alignments using HHpred[70] to identify the TMH homo-oligomer homologues with known structures, which align best with a target sequence, as described previously[34]. Different target sequence lengths were tried, and query sequences corresponding to the TM region were found to generate the best alignments and were used for all targets. The structural homologue whose sequence aligned best (that is, highest HHpred score and no gap in the aligned TM) with that of each target was selected as the structural template. To assess the accuracy of the alignment and the resulting homology models, the target sequence was threaded onto the homologue structures and relaxed to identify low-energy conformations using the homology-based modelling technique of RosettaMembrane[34]. To extract constraints from these homologues, the closest contact at each helical turn of the binding interface in the homologue structure

were selected and implemented as distance constraints in the folding and docking simulations that we defined as HFDOCK-TM. The total number of template-derived constraints in the HFDOCK-TM simulations lay between 4 and 7 for each target. The treatment of constraints and the analysis of the simulations using HFDOCK-TM were identical than with EFDOCK-TM. The accuracy (that is, true positive rate) of the template-derived constraints was very similar when considering a larger set of constraints (that is, top two closest contacts per helical turn at the homologue-binding interface).

Assessment of homo-oligomer structure prediction accuracy. The r.m.s.d. of the binding interface region was calculated to assess the accuracy of the predictions. The interface region was defined by the residues for which experimental inter-monomeric NMR constraints were obtained (from 9 to 24 residues were selected for each target). The r.m.s.d. was calculated between the representative EFDOCK-TM model and the centre NMR model. All calculations were performed using an open-source python script MATCH (http://boscoh.com/protein/matchpy.html).

Native contact calculation. Native residue–residue contacts were defined if the distance between any of the heavy atoms from two residues on each protein monomer was smaller than 4 Å in the experimental structures. Residue–residue contacts in predicted models were calculated using the same criteria.

Comparison with alternative techniques CATM and PREDDIMER. The accuracy of published CATM models[24] is reported in Table 1. PREDDIMER[30] models for all TM dimers in our data set were generated using the webserver: http://model.nmr.ru/preddimer/. The most accurate among all models output by the server is reported in Table 1. Accuracy of PREDDIMER and EFDOCK-TM models was analysed using identical criteria.

Interhelical contact density comparison in TMH proteins. The average number of interhelical contacts per helix in self-associated single-pass TMH protein complexes was calculated using all 20 native structures constituting the benchmark. Residue–residue contact was defined if the distance between two heavy atoms is within a certain threshold. Various distance thresholds were applied (4, 4.5 and 5 Å) for comparison. The same calculation was performed on a representative set of 75 non-redundant multi-pass TM helical domains (sequences identity less than 30%) with resolution better than 3.5 Å from the Protein Data Bank.

Native sequence recovery. For each target, the NMR structures, EFDOCK-TM and PREDDIMER models were selected and all residues making contacts (between two and three residues per helical turn) at the binding interface were randomized to all 20 amino acids and redesigned to identify the combination that minimizes the energy of the homo-oligomer structures using the design mode of Rosetta-Membrane as previously described[23]. The percentage of native residues recovered by design at the binding interface was calculated for each starting structure, and the intersection of these native recovered sites between NMR and EFDOCK-TM or PREDDIMER models was reported.

Method release. Softwares to run and analyse the EFDOCK-TM simulations will be released at the time of publication, and detailed information to run the simulations and reproduce the results is provided in Supplementary Methods.

References

1. Moore, D. T., Berger, B. W. & DeGrado, W. F. Protein-protein interactions in the membrane: sequence, structural, and biological motifs. *Structure* **16,** 991–1001 (2008).
2. Lemmon, M. A. & Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **141,** 1117–1134 (2010).
3. Call, M. E. *et al.* The structure of the zetazeta transmembrane dimer reveals features essential for its assembly with the T cell receptor. *Cell* **127,** 355–368 (2006).
4. Call, M. E. & Wucherpfennig, K. W. Common themes in the assembly and architecture of activating immune receptors. *Nat. Rev. Immunol.* **7,** 841–850 (2007).
5. Maurel, D. *et al.* Cell-surface protein-protein interaction analysis with time-resolved FRET and snap-tag technologies: application to GPCR oligomerization. *Nat. Methods* **5,** 561–567 (2008).
6. Langosch, D. & Arkin, I. T. Interaction and conformational dynamics of membrane-spanning protein helices. *Protein Sci.* **18,** 1343–1358 (2009).
7. Matthews, E. E., Zoonens, M. & Engelman, D. M. Dynamic helix interactions in transmembrane signaling. *Cell* **127,** 447–450 (2006).
8. Endres, N. F. *et al.* Conformational coupling across the plasma membrane in activation of the EGF receptor. *Cell* **152,** 543–556 (2013).
9. Arkhipov, A. *et al.* Architecture and membrane interactions of the EGF receptor. *Cell* **152,** 557–569 (2013).
10. Cymer, F. & Schneider, D. Transmembrane helix-helix interactions involved in ErbB receptor signaling. *Cell Adh. Migr.* **4,** 299–312 (2010).
11. Li, E., You, M. & Hristova, K. FGFR3 dimer stabilization due to a single amino acid pathogenic mutation. *J. Mol. Biol.* **356,** 600–612 (2006).
12. Toffalini, F. & Demoulin, J. B. New insights into the mechanisms of hematopoietic cell transformation by activated receptor tyrosine kinases. *Blood* **116,** 2429–2437 (2010).
13. Yang, Y. *et al.* Structural basis for dimerization of ICAM-1 on the cell surface. *Mol. Cell* **14,** 269–276 (2004).
14. Zhang, X., Gureasko, J., Shen, K., Cole, P. A. & Kuriyan, J. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* **125,** 1137–1149 (2006).
15. MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. A transmembrane helix dimer: structure and implications. *Science* **276,** 131–133 (1997).
16. Bocharov, E. V. *et al.* Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state. *J. Biol. Chem.* **283,** 6950–6956 (2008).
17. Jura, N. *et al.* Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment. *Cell* **137,** 1293–1307 (2009).
18. Mineev, K. S. *et al.* Spatial structure of the transmembrane domain heterodimer of ErbB1 and ErbB2 receptor tyrosine kinases. *J. Mol. Biol.* **400,** 231–243 (2010).
19. Zhu, J. *et al.* The structure of a receptor with two associating transmembrane domains on the cell surface: integrin alphaIIbbeta3. *Mol. Cell* **34,** 234–249 (2009).
20. Lau, T. L., Kim, C., Ginsberg, M. H. & Ulmer, T. S. The structure of the integrin alphaIIbbeta3 transmembrane complex explains integrin transmembrane signalling. *EMBO J.* **28,** 1351–1361 (2009).
21. Lu, C. *et al.* Structural evidence for loose linkage between ligand binding and kinase activation in the epidermal growth factor receptor. *Mol. Cell. Biol.* **30,** 5432–5443 (2010).
22. Soto, C. S., Hannigan, B. T. & DeGrado, W. F. A photon-free approach to transmembrane protein structure determination. *J. Mol. Biol.* **414,** 596–610 (2011).
23. Barth, P., Schonbrun, J. & Baker, D. Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc. Natl Acad. Sci. USA* **104,** 15682–15687 (2007).
24. Mueller, B. K., Subramaniam, S. & Senes, A. A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Calpha-H hydrogen bonds. *Proc. Natl Acad. Sci. USA* **111,** E888–E895 (2014).
25. Walters, R. F. & DeGrado, W. F. Helix-packing motifs in membrane proteins. *Proc. Natl Acad. Sci. USA* **103,** 13658–13663 (2006).
26. Sulistijo, E. S. & Mackenzie, K. R. Structural basis for dimerization of the BNIP3 transmembrane domain. *Biochemistry* **48,** 5106–5120 (2009).
27. Bocharov, E. V. *et al.* Structure of FGFR3 transmembrane domain dimer: implications for signaling and human pathologies. *Structure* **21,** 2087–2093 (2013).
28. Bocharov, E. V. *et al.* Left-handed dimer of EphA2 transmembrane domain: helix packing diversity among receptor tyrosine kinases. *Biophys. J.* **98,** 881–889 (2010).
29. Ried, C. L., Kube, S., Kirrbach, J. & Langosch, D. Homotypic interaction and amino acid distribution of unilaterally conserved transmembrane helices. *J. Mol. Biol.* **420,** 251–257 (2012).
30. Polyansky, A. A., Volynsky, P. E. & Efremov, R. G. Multistate organization of transmembrane helical protein dimers governed by the host membrane. *J. Am. Chem. Soc.* **134,** 14390–14400 (2012).
31. Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294,** 93–96 (2001).
32. Fleishman, S. J. & Baker, D. Role of the biomolecular energy gap in protein design, structure, and evolution. *Cell* **149,** 262–273 (2012).
33. Gray, J. J. High-resolution protein-protein docking. *Curr. Opin. Struct. Biol.* **16,** 183–193 (2006).
34. Chen, K. Y., Sun, J., Salvo, J. S., Baker, D. & Barth, P. High-resolution modeling of transmembrane helical protein structures from distant homologues. *PLoS Comput. Biol.* **10,** e1003636 (2014).
35. Thomas, J., Ramakrishnan, N. & Bailey-Kellogg, C. Graphical models of residue coupling in protein families. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **5,** 183–197 (2008).
36. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6,** e28766 (2011).
37. Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28,** 184–190 (2012).
38. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108,** E1293–E1301 (2011).
39. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl Acad. Sci. USA* **110,** 15674–15679 (2013).

40. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79,** 1061–1078 (2011).
41. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E. Stat. Nonlin. Soft. Matter Phys.* **87,** 012707 (2013).
42. Hopf, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149,** 1607–1621 (2012).
43. Nugent, T. & Jones, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl Acad. Sci. USA* **109,** E1540–E1547 (2012).
44. Morcos, F., Jana, B., Hwa, T. & Onuchic, J. N. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl Acad. Sci. USA* **110,** 20533–20538 (2013).
45. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3,** e02030 (2014).
46. Dago, A. E. *et al.* Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc. Natl Acad. Sci. USA* **109,** E1733–E1742 (2012).
47. Madaoui, H. & Guerois, R. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proc. Natl Acad. Sci. USA* **105,** 7708–7713 (2008).
48. Barth, P., Wallner, B. & Baker, D. Prediction of membrane protein structures with complex topologies using limited constraints. *Proc. Natl Acad. Sci. USA* **106,** 1409–1414 (2009).
49. Adamian, L. & Liang, J. Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Struct. Biol.* **6,** 13 (2006).
50. Defour, J. P. *et al.* Tryptophan at the transmembrane-cytosolic junction modulates thrombopoietin receptor dimerization and activation. *Proc. Natl Acad. Sci. USA* **110,** 2540–2545 (2013).
51. Bocharov, E. V. *et al.* Spatial structure and pH-dependent conformational diversity of dimeric transmembrane domain of the receptor tyrosine kinase EphA1. *J. Biol. Chem.* **283,** 29385–29395 (2008).
52. Acharya, R. *et al.* Structure and mechanism of proton transport through the transmembrane tetrameric M2 protein bundle of the influenza A virus. *Proc. Natl Acad. Sci. USA* **107,** 15075–15080 (2010).
53. Stouffer, A. L. *et al.* Structural basis for the function and inhibition of an influenza virus proton channel. *Nature* **451,** 596–599 (2008).
54. Chen, K. Y., Zhou, F., Fryszczyn, B. G. & Barth, P. Naturally evolved G protein-coupled receptors adopt metastable conformations. *Proc. Natl Acad. Sci. USA* **109,** 13284–13289 (2012).
55. Yin, H. *et al.* Computational design of peptides that target transmembrane helices. *Science* **315,** 1817–1822 (2007).
56. Bocharov, E. V., Mineev, K. S., Goncharuk, M. V. & Arseniev, A. S. Structural and thermodynamic insight into the process of 'weak' dimerization of the ErbB4 transmembrane domain by solution NMR. *Biochim. Biophys. Acta* **1818,** 2158–2170 (2012).
57. Bocharov, E. V. *et al.* Unique dimeric structure of BNip3 transmembrane domain suggests membrane permeabilization as a cell death trigger. *J. Biol. Chem.* **282,** 16256–16266 (2007).
58. Chen, W. *et al.* Familial Alzheimer's mutations within APPTM increase Abeta42 production by enhancing accessibility of epsilon-cleavage site. *Nat. Commun.* **5,** 3037 (2014).
59. Mineev, K. S. *et al.* Spatial structure and dimer--monomer equilibrium of the ErbB3 transmembrane domain in DPC micelles. *Biochim. Biophys. Acta* **1808,** 2081–2088 (2011).
60. Call, M. E., Wucherpfennig, K. W. & Chou, J. J. The structural basis for intramembrane assembly of an activating immunoreceptor complex. *Nat. Immunol.* **11,** 1023–1029 (2010).
61. Muhle-Goll, C. *et al.* Hydrophobic matching controls the tilt and stability of the dimeric platelet-derived growth factor receptor (PDGFR) beta transmembrane segment. *J. Biol. Chem.* **287,** 26178–26186 (2012).
62. Nadezhdin, K. D., Bocharova, O. V., Bocharov, E. V. & Arseniev, A. S. Dimeric structure of transmembrane domain of amyloid precursor protein in micellar environment. *FEBS Lett.* **586,** 1687–1692 (2012).
63. Wang, J., Pielak, R. M., McClintock, M. A. & Chou, J. J. Solution structure and functional analysis of the influenza B proton channel. *Nat. Struct. Mol. Biol.* **16,** 1267–1271 (2009).
64. Verardi, R., Shi, L., Traaseth, N. J., Walsh, N. & Veglia, G. Structural topology of phospholamban pentamer in lipid bilayers by a hybrid solution and solid-state NMR method. *Proc. Natl Acad. Sci. USA* **108,** 9101–9106 (2011).
65. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr. Protoc. Bioinform.* **Chapter 1,** Unit1 13 (2012).
66. Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9,** 173–175 (2012).
67. Wolf, E., Kim, P. S. & Berger, B. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.* **6,** 1179–1189 (1997).
68. Das, R. *et al.* Simultaneous prediction of protein folding and docking at high resolution. *Proc. Natl Acad. Sci. USA* **106,** 18978–18983 (2009).
69. Viklund, H. & Elofsson, A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* **24,** 1662–1668 (2008).
70. Soding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–W248 (2005).

Acknowledgements

We thank the members of the Barth lab for insightful discussions during this study and critical comments on the manuscript. This work was supported by a grant from the National Institute of Health (1R01GM097207-01A1) and by a supercomputer allocation from XSEDE (MCB120101) to P.B. We thank Aaron Kelly for his involvement in the initial stages of the project.

Author contributions

P.B. designed the study. P.B. and Y.W. developed the methods. Y.W. performed the benchmark. P.B. and Y.W. analysed and discussed the results. P.B. wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Wang, Y. & Barth, P. Evolutionary-guided *de novo* structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy. *Nat. Commun.* 6:7196 doi: 10.1038/ncomms8196 (2015).