Cell

Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing

Thomas A. Hopf,[1,2] Lucy J. Colwell,[3] Robert Sheridan,[4] Burkhard Rost,[2] Chris Sander,[4] and Debora S. Marks[1,*]
[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[2]Department of Informatics, Technische Universität München, 85748 Garching, Germany
[3]MRC Laboratory of Molecular Biology, Hills Road, CB2 0QH Cambridge, UK
[4]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York City, NY 10065, USA
*Correspondence: evfold_membrane@cbio.mskcc.org

DOI 10.1016/j.cell.2012.04.012

SUMMARY

We show that amino acid covariation in proteins, extracted from the evolutionary sequence record, can be used to fold transmembrane proteins. We use this technique to predict previously unknown 3D structures for 11 transmembrane proteins (with up to 14 helices) from their sequences alone. The prediction method (EVfold_membrane) applies a maximum entropy approach to infer evolutionary covariation in pairs of sequence positions within a protein family and then generates all-atom models with the derived pairwise distance constraints. We benchmark the approach with blinded de novo computation of known transmembrane protein structures from 23 families, demonstrating unprecedented accuracy of the method for large transmembrane proteins. We show how the method can predict oligomerization, functional sites, and conformational changes in transmembrane proteins. With the rapid rise in large-scale sequencing, more accurate and more comprehensive information on evolutionary constraints can be decoded from genetic variation, greatly expanding the repertoire of transmembrane proteins amenable to modeling by this method.

INTRODUCTION

Membrane proteins allow cells to interact with the extracellular environment and to communicate with other cells. More than 25% of all human proteins have integral membrane domains; many of these are medically important, with nearly half of all drug targets containing a membrane domain (Bakheet and Doig, 2009; Overington et al., 2006). Knowing the three-dimensional (3D) structure of a membrane protein facilitates the characterizations its molecular mechanism and accelerates the development of pharmacological agents targeting it (Katritch et al., 2012). Despite great progress in determining structures by experimental methods (Chen et al., 2010; Cherezov et al., 2007; Choe et al., 2011; Long et al., 2007; Miller and Long, 2012; Rasmussen et al., 2011; Rasmussen et al., 2007), the 3D structures of most transmembrane proteins remain unknown, and comparative modeling maximally covers 10% of all human transmembrane proteins. Efficient and accurate computational approaches that predict 3D structures of membrane proteins would be a valuable tool to complement existing experimental approaches.

Well-established methods of structure prediction, such as energy minimization and database fragment searches, have previously addressed the problem of prediction of transmembrane protein structures. However, these calculations were limited in both protein size ($\leq$7 transmembrane helices) and accuracy, despite added information on helix-helix contact predictions from experimentally nonhomologous structures and a few known experimentally determined contacts (Barth et al., 2009; Yarov-Yarovoy et al., 2006).

It is possible that constraints on the function and structure of proteins are reflected in conserved interactions between pairs, or groups, of amino acids. If so, then evolutionary correlations may be observed between specific sequence positions. Previous work has attempted to use correlations between residues, among other methods, to predict structural proximity and functional features (Cronet et al., 1993; Fatakia et al., 2009; Fuchs et al., 2007; Göbel et al., 1994; Horn et al., 1998; Nemoto et al., 2004). The most accurate of these strategies use global statistical methods, such as maximum entropy (Marks et al., 2011; Morcos et al., 2011), Bayesian networks (Burger and van Nimwegen, 2010), or covariance estimation (Jones et al., 2012; Meinshausen and Buhlmann, 2006). However, only recently it was reported that a maximum entropy analysis of residue correlations in sequence families could provide sufficient information about proximity of residues in 3D to compute correct folds of protein structures in 15 example cases, using EVfold (Marks et al., 2011).

Here, we report the development of an algorithm, EVfold_membrane, which enables de novo prediction of 3D structures of unknown α-helical transmembrane proteins from evolutionary constraints, using neither fragments, threading, nor homologous 3D structures. We predict the structures of 11 transmembrane proteins of unknown structure, including eight pharmacological targets (Figure 1, Table 1, and Figure S1 available online). To verify

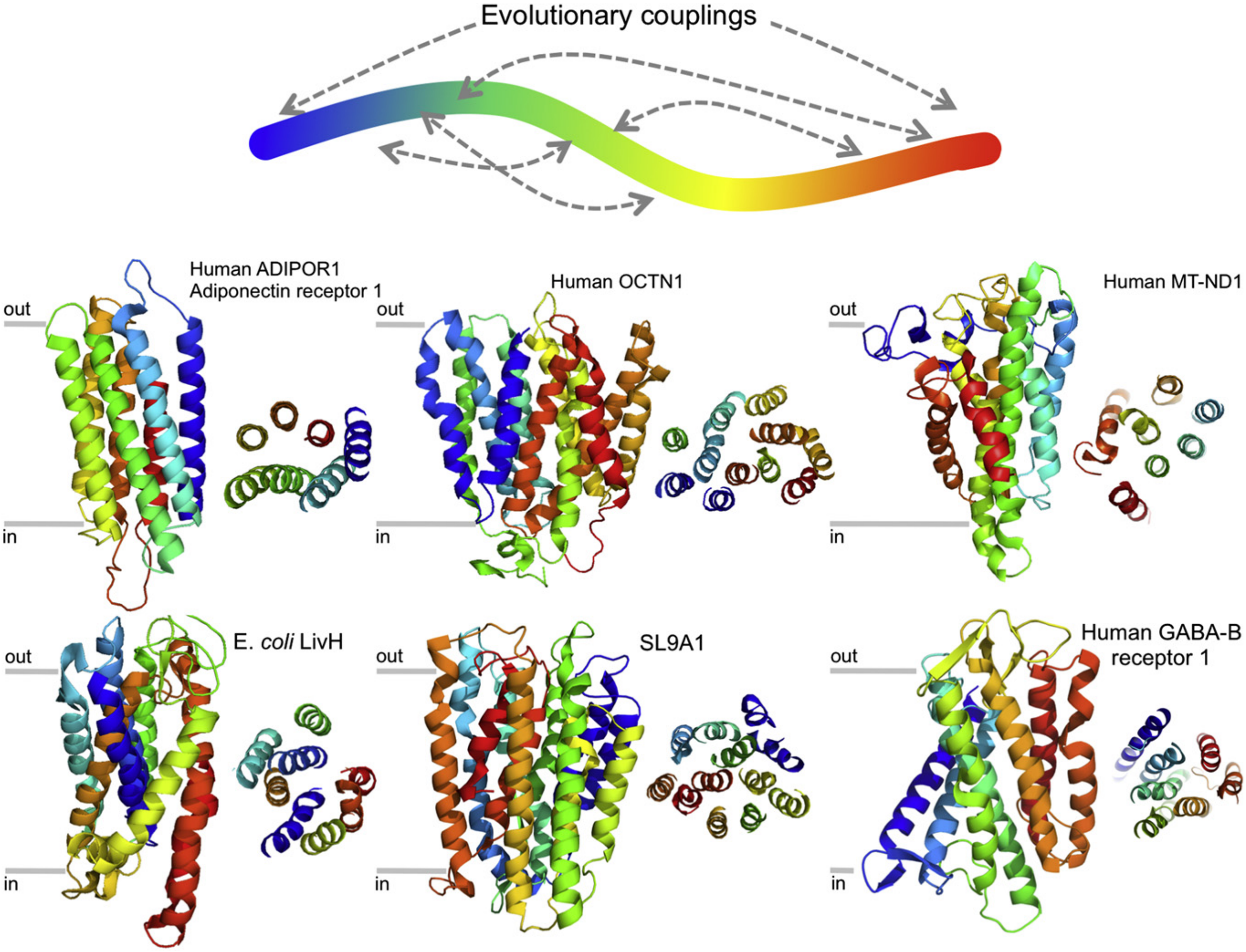

Figure 1. De Novo Predicted 3D Models of Membrane Proteins with No Known Structure

Cartoon shows evolutionary couplings as calculated by EVfold; membrane placed as distance constraints on extended polypeptide before folding. Top-ranked models of a representative set of six transmembrane proteins from diverse families, which have no members with known 3D structures. Models are cartoon representation with rainbow coloring blue N terminal to red C terminal, seen from the side (left) and noncytoplasmic side (right). Naming conventions, 3D coordinates, and input files in are shown in Tables 1 and S1 and Data S1–S5.

that our predicted structures are plausible, we systematically test our ability to predict, in blinded fashion, the structures of a diverse set of 25 transmembrane proteins with known 3D structures (Table 1) and find an unprecedented level of agreement with the cognate crystal structures (TM scores > 0.5 for 22 out of 25 of the benchmarked proteins). We find that functionally important regions of each protein tend to be more accurately predicted than the protein as a whole and that residues that are subject to multiple pair constraints tend to be in substrate binding pockets, oligomerization interfaces, and/or involved in conformational changes.

RESULTS

Global Statistical Approach for Protein Structures from Sequences

Our hypothesis is that evolution conserves interactions between residues that are important to maintaining structure and function by constraining the sets of mutations that are accepted at interacting sites. To find these constraint couplings for each membrane protein, we build a multiple sequence alignment (Remmert et al., 2012) with sufficiently diverse sequences to detect evolutionary covariation and minimize statistical noise. To maximize the power of detection, we developed a method to optimize the trade-off between the number of sequences aligned (i.e., depth) and alignment specificity, a proxy for functional similarity to the query sequence, which is quantified by the sequence range (i.e., breadth) covered by the alignment (Figures 2A, S2, and Experimental Procedures). For example, for bovine Ant1, which catalyzes the exchange of cytoplasmic ADP with mitochondrial ATP, we use a stringency value (E) of 10^{-40}, ensuring that 70% of its residues in the sequence are covered by the alignment. In general, for a protein of length L, we require at least 3L sequences and coverage of at least 0.7 × L of the residues in the sequence of interest.

Table 1. Predicted Proteins of Known and Unknown Experimental Structure

Uniprot Name	Length	TMH[a]	E-val[b]	Model Length	#Seq[c]	Top #[d]	TM[e]	Cα-rmsd[f]	Best #[d]	TM[e]	Cα-rmsd[f]	PDB[g]
Known Structure												
ADIC_SALTY	445	12	E-20	394	24284	240_15	0.67	4.2 (300)	240_15	0.67	4.2 (300)	3ncyA
ADRB2_HUMAN	413	7	E-20	296	35593	160_5	0.67	3.3 (201)	160_5	0.67	3.3 (201)	2rh1A
ANT1_BOVIN	298	6	E-40	285	9828	200_20	0.48	3.8 (136)	270_17	0.51	4.0 (152)	1okcA
AMTB_ECOLI	428	10	E-5	396	4407	270_17	0.67	3.9 (262)	280_5	0.67	3.6 (260)	1xqfA
AQP4_HUMAN	323	6	E-10	215	6469	80_19	0.50	2.9 (100)	100_14	0.51	3.4 (110)	3gd8A
BTUC_ECOLI	326	10	E-10	299	12926	250_19	0.67	3.2 (209)	250_19	0.67	3.2 (209)	1l7vA
C3NQD8_VIBCJ	461	12	E-20	431	13864	250_11	0.62	4.6 (306)	290_8	0.63	4.3 (305)	3mktA
C6E9S6_ECOBD	485	14	E-10	412	63730	180_9	0.63	4.2 (299)	180_9	0.63	4.2(299)	3rkoN
COX1_BOVIN	514	12	E-40	486	73822	150_6	0.66	4.5 (360)	150_11	0.66	4.4 (354)	1occA
COX3_BOVIN	261	7	E-3	182	10705	50_9	0.69	2.8 (151)	50_9	0.69	2.8 (151)	1occC
CYB_BOVIN	379	8	E-3	335	43891	120_4	0.58	4.1 (203)	100_9	0.64	3.7 (231)	1pp9B
FIEF_ECOLI	300	6	E-5	197	9722	200_10	0.59	2.8 (119)	40_7	0.63	2.8 (131)	3h90A
GLPG_ECOLI	276	6	E-5	169	5263	120_11	0.64	2.6 (126)	120_11	0.64	2.6 (126)	3b45A
GLPT_ECOLI	452	12	E-30	402	24912	330_12	0.67	3.8 (283)	330_13	0.67	4.0 (297)	1pw4A
METI_ECOLI	217	6	E-15	206	30400	120_17	0.46	3.5 (93)	120_6	0.48	3.4 (94)	3dhwA
MIP_BOVIN	263	6	E-10	212	6468	150_12	0.55	3.1 (116)	130_20	0.58	2.9 (124)	1ymgA
MSBA_SALTY	330	6	E-3	310	29034	100_12	0.57	3.3 (180)	110_12	0.61	3.5 (208)	3b60A
O67854_AQUAE	513	12	E-3	463	4500	280_4	0.55	5.1 (274)	170_20	0.58	4.8 (286)	2a65A
OPSD_BOVIN	348	7	E-20	274	35901	110_16	0.70	3.3 (214)	110_16	0.70	3.3 (214)	1hzxA
Q87TN7_VIBPA	485	8	E-10	407	4097	270_12	0.59	4.0 (242)	260_19	0.60	4.2 (258)	3pjzA
Q8EKT7_SHEON	516	12	E-10	447	12063	100_14	0.40	4.6 (160)	240_19	0.43	4.8(183)	2xutA
Q9K0A9_NEIMB	315	10	E-10	297	4244	270_9	0.44	3.6 (131)	120_9	0.49	3.9 (138)	3zuxA
SGLT_VIBPA	543	14	E-5	487	9563	310_11	0.49	4.6 (214)	340_10	0.53	4.8 (264)	2xq2A
TEHA_HAEIN	328	10	E-3	304	1861	70_15	0.51	4.1 (154)	210_17	0.56	4.0 (175)	3m71A
URAA_ECOLI	429	14	E-3	393	14992	250_12	0.50	4.8 (194)	250_5	0.50	4.5 (189)	3qe7A
Unknown Structure							Structural Similarity to			Z[h]	Cα-rmsd[f]	PDB[g]
ADR1_HUMAN	375	7	E-5	223	3410	150_14	bacteriorhodopsin			12	4.5 (204)	3haoA
NU1M_HUMAN	318	8	E-10	282	17558	210_18	mit. complex 1 subunit L			10	5.0 (170)	3rkoL
S22A4_HUMAN	551	12	E-30	373	21704	220_11	L-fucose permease FucP			10	6.0 (267)	3o7qA
ABCG2_HUMAN	655	7	E-10	274	5404	210_3						
ELOV4_HUMAN	314	7	E-3	233	1436	190_6						
SL9A1_HUMAN	815	13	E-10	367	6020	210_17	acriflavine res. prot. AcrB			4	4.7 (165)	2gifA
MSMO1_HUMAN	293	5	E-20	220	897	70_13						
S13A1_HUMAN	595	15	E-20	543	1836	none[i]						
EAMA_ECOLI	299	10	E-5	276	31753	250_10						
LIVH_ECOLI	308	8	E-3	282	23968	230_16	permease protein BtuC			6	4.1 (140)	1l7vA
GABR1_HUMAN	961	7	E-5	298	2871	190_19	β2 adrenergic receptor			6	6.0 (191)	3p0gA

[a] Number of transmembrane helices.

[b] E value for HHblits sequence search.

[c] Number of sequences in multiple sequence alignment.

[d] Number of evolutionary constraints used and model number of blind top-ranked and best-generated model, respectively.

[e] TM score.

[f] Cα, root-mean-square deviation in Å (number of residues).

[g] Accession code and chain of similar PDB structure that has negligible sequence similarity.

[h] DALI Z score.

[i] No model looks plausible (large protein, few sequences).

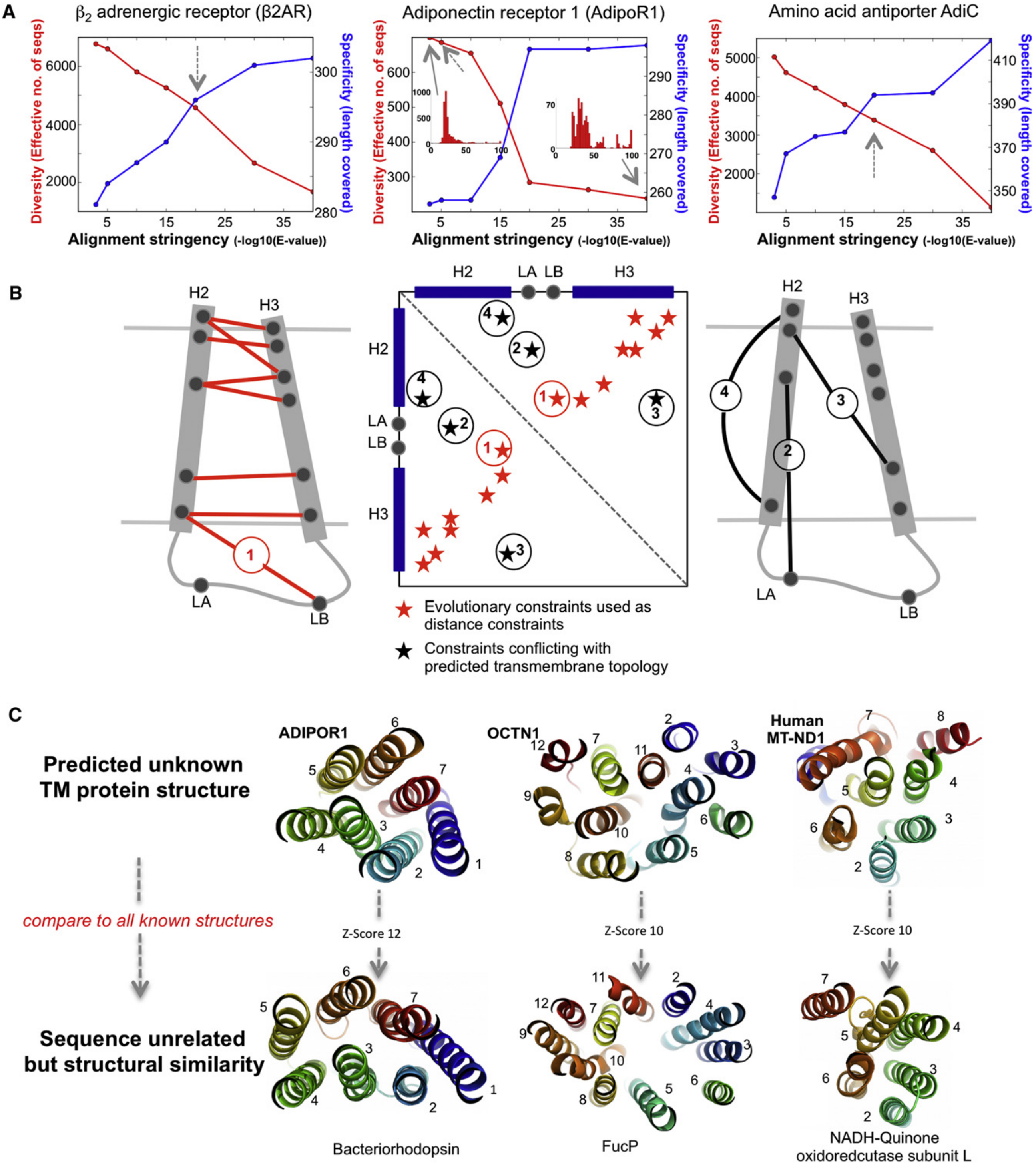

Figure 2. From Sequence Alignment to Folded Structures

(A) Building the alignment for the EC calculation for the specific query protein requires a trade-off between specificity and diversity. To investigate this blindly, we scan a range of alignment depths using different expectation values and calculate the effective number of sequences returned (diversity) and the number of residues in our query protein sequence that do not have more than 30% gaps in the alignment column of the alignment (coverage). Dashed arrows point to chosen stringency for folding. Contrast in the distribution of sequence space at different alignment depths in histograms of the range of number of sequences with the 0%–100% identity to query protein sequence (insets, middle).

(B) Schematic showing constraint conflict resolution between predicted coevolution and predicted secondary structure/membrane topology. In all cases we follow the predicted membrane topology and discard coevolving residue pairs that conflict with this prediction. The predicted toy contact map (middle) shows

To discover residue interactions that are conserved by evolution, we developed an algorithm that extracts patterns of amino acid coevolution from these sequence alignments (Lapedes et al., 1997; Marks et al., 2011; Morcos et al., 2011). The algorithm, which uses entropy maximization, transforms the set of observed amino acid correlations in all pairs of sequence positions to a set of position couplings (residue couplings) that best explains the observed data. This set of globally consistent residue couplings is likely to be causative, i.e., likely to reflect residue interactions constrained in evolution. Our statistical approach is thus in a class of algorithms that address the classic problem of deriving ‘‘causation from correlation.’’ Our ‘‘global’’ statistical approach is different from ‘‘local’’ approaches such as mutual information (MI) and variants thereof (Fodor and Aldrich, 2004; Livesay et al., 2012). The MI of pairs of columns in a sequence alignment is local in that it quantifies covariation for each pair independently of all other pairs, potentially leading to inconsistencies. The simplest inconsistencies in local models are transitive correlations, e.g., correlations between a noncontact pair A-C in a triplet A-B-C that arise from transitive influence in contact pairs A-B and B-C. Thus, pairs with high MI scores are not necessarily constrained by a direct interaction effect, even if they are correlated.

In contrast, our entropy maximization approach builds a probability model for the entire sequence, such that the scores for each pair of residues are consistent with other pairs, thereby preventing high scoring from transitive relationships in the data. Starting with a simple covariance matrix between all pairs of columns in the alignment, entropy maximization gives rise to a formalism that is similar to the well-known inverse Ising model of ferromagnetism (in which there are two states) except that, for protein sequences, each site (i.e., sequence position) can be assigned to 1 of 21 discrete states (20 amino acids or a gap), as in the Potts model in physics. The numerical parameters in the entropy maximization method (analogous to the spin-spin interactions in the Ising model Hamiltonian) can be computed efficiently by inverting a covariance matrix. This algorithmic entropy maximization solution is similar to partial correlation methods in Gaussian graphical models for continuous distributions (Dempster, 1972). In entropy maximization, after the covariance matrix inversion, the residue pair scores, or evolutionary coupling scores, are consistent with the correlation data between pairs of positions and single column data, including conservation, while making a minimum set of other assumptions. Although constrained interactions can arise from diverse evolutionary requirements, we find that many reflect interactions between residues close in space and are thus highly productive as distance constraints for protein folding (Marks et al., 2011).

The structure of transmembrane proteins is additionally constrained by the presence of the membrane. Hence, we can blindly remove predicted coevolved pairs for which 3D proximity is unlikely (Figures 2B and S3 and Experimental Procedures). The resulting set of evolutionary constraints and the predicted secondary structure are interpreted as distance constraints on extended polypeptide chains (Data S1). Distance geometry and out-of-the-box simulated annealing using the CNS software (Brünger et al., 1998) are used to fold the chain ab initio to produce ~500 3D all-atom coordinate models for each protein. To assess the set of predicted structures for each protein, we apply an automated membrane-specific ranking of the computed models that combines the quality of secondary structure formation, lipid accessibility of the residues, and a measure of violation of the evolutionary constraints and cluster the structures, excluding predictions not represented in the larger clusters (Experimental Procedures).

Prediction of Unknown 3D Structures of α-Helical Transmembrane Proteins

A survey of targets in the DrugBank database (Knox et al., 2011) for transmembrane proteins together with large families from the CAMPS (Neumann et al., 2012) yielded 18 nonredundant families with >1,000 sequences, with ≥5 predicted transmembrane helices, and without a known 3D structure for any family member. We selected 11 of these targets for detailed analysis, covering diverse sizes and functional types, with several of the families having more than one drug target (Tables 1 and S1 and Experimental Procedures). Coordinates for the remaining seven families are available at http://evfold.org/transmembrane. These proteins are implicated in many diseases, including diabetes, obesity, Crohn's disease, breast cancer, Leber hereditary optic neuropathy, Alzheimer's disease, and Parkinson's disease (Holland et al., 2011; Pei et al., 2011; Peltekova et al., 2004; Yamauchi et al., 2003; Doyle et al., 1998; Natarajan et al., 2012; Howell et al., 1991; Jaksch et al., 1996; Aldahmesh et al., 2011; Zhang et al., 2001). We predicted 400–600 all-atom 3D models for each protein (Data S2 and Experimental Procedures). The predicted structures of five of the proteins had similar folds to other known 3D membrane protein structures (Figure 2C) despite negligible sequence similarity, a recurring theme seen in structural genomics and earlier work (Holm and Sander, 1993; Murzin, 1993). Predicted structures of three membrane proteins show some structural similarities to those of other sequence-distant members of the same PFAM clan. A search against all known 3D structures with our top-ranked model of the human OCTN1, a 12 helical transporter sugar transporter, yields several significant hits to structures in the major facilitator superfamily, including FucP (PDB: 3o7q; Dang et al., 2010) and GlpT (PDB: 1pw4; Law et al., 2008) (Figures 1 and 2C and Tables 1 and S1). FucP and GlpT sequences were not in our alignment and have only 10% and 7% sequence identity to OCTN1, respectively, below the level allowing inference of structural homology. Similarly, the 8 transmembrane helical *E. Coli* LIVH, a high-affinity

evolutionary constraints that conflict with the predicted membrane topology that are removed (black stars). Evolutionary constraints that do not conflict with the predicted membrane topology are not removed, irrespective of any knowledge about their distance in 3D space (constraint 1).

(C) The top-ranked model from the set of each de novo predicted structure was compared to the entire PDB using the structural alignment program DALI (Holm and Sander, 1995). Three of the six predicted 3D TM protein structures with significant structural similarities to known transmembrane protein folds are shown. See also Figure S2.

leucine transporter, is structurally similar to the bacterial B12 uptake protein BtuC (PDB: 1l7v; Locher et al., 2002) despite only 8% sequence identity between the proteins (Tables 1 and S1). Third, the predicted structures of the GABA receptor 1, a protein involved in synaptic inhibition and a pharmacological target, are structurally similar to other GPCRs despite a negligible sequence identity (10%) (Figures 1 and S1 and Table 1). Although this result is not so surprising, the sequence diversity in GPCRs is sufficiently high that de novo computation of the 3D structure from evolutionary couplings may be of interest, in addition to model building by remote homology (Katritch et al., 2012) (Figures 1 and S1 and Table 1). In the predicted models of the GABA receptor, a lack of well-ordered structure formed by the extracellular loops and a lack of b sheet formation by the predicted β strands indicate potential model errors. Nevertheless, high-scoring predicted residue pair interactions in the extracellular region, specifically between loops 2/3 and 3/4, are located close to the putative extracellular ligand-binding domain. Given the moderate number of sequences in this GABA receptor family, we expect the current accuracy to be limited, but the models may serve as a useful starting point for further iteration using hybrid approaches and different alignment depths.

The five top-ranked predicted adiponectin receptor 3D structures are surprisingly similar (~4.5 Å Cα-rmsd over 204 residues) to the bacteriorhodopsin crystal structure (PDB: 3hao), with highly significant Dali (Holm and Sander, 1995) Z scores between 7 and 13, despite negligible sequence identity (8%) (Figures 1 and 2C and Data S2). Although adiponectin receptor is a 7 transmembrane protein, it was not previously thought to have structural or functional similarity to G protein-coupled receptors and is inverted with respect to the membrane (Yamauchi et al., 2003). Assuming that our predictions are accurate, it remains an open question whether the similarity of AdipoR1 to the GPCR fold is an example of divergent evolution or the result of convergent evolution to an exceptionally robust 7 helical fold.

We also find significant structural similarity of predicted structures of the human MT-ND1 subunit to the recently solved structure of one of the major membrane subunits of respiratory complex I (*E. coli*, 3rko-C; NuoL subunit) (Efremov and Sazanov, 2011); again, the sequences of MT-ND1 and the NuoL subunit are unrelated, with <8% identity. However, we do not find high topological similarity to the coarse grained model of the bacterial NuoL subunit (homologous to MT_ND1), which was solved at low resolution without residue assignment (Efremov et al., 2010), and the NuoL subunit is almost double the size of our modeled protein. Nevertheless, our MT-ND1 structures overlay optimally on precisely the regions of bacterial subunits that are structurally duplicated within each protein (in NuoL, TM helices 3–7 and 8–15), further supporting the idea that this is a repeating structural evolutionary module (Efremov and Sazanov, 2011). Because these mitochondrial subunits are functionally related and spatially coincident throughout evolution, the structural relationship between them may plausibly result from divergent evolution of the sequence. Taken together, these examples of structural relationships between the predicted models and the structures of functionally related but sequence-distant proteins provide support for the accuracy of the de novo prediction.

Benchmark: Blinded Prediction of Transmembrane Proteins of Known 3D Structure

To evaluate the performance of the prediction protocol, we computed the 3D structures of α-helical membrane proteins of known structure from the proteins' sequences alone, i.e., ignoring all aspects of known 3D structures, including sequence-similar fragments. We selected all α-helical membrane proteins from all Pfam families that have >1,000 sequences, sufficient sequence coverage, and more than 4 helices. This resulted in a set of 25 membrane proteins with up to 487 residues (up to 14 transmembrane helices) in 23 structurally diverse families. This set includes the human β2 adrenergic receptor (GPCR family), the *S. typhimurium* arginine/agmatine antiporter ADIC (amino acid/polyamine transporter superfamily), and the *E. coli* glycerol-3-phosphate transporter (GlpT; major facilitator superfamily) (Table 1 and Data S3–S5).

The EVfold_membrane protocol provides a ranked set of predicted structures for each protein, which we then compare to a cognate crystal structure. The combined score used for ranking the generated models reliably identifies structures of high accuracy and, in some cases, even the best model in the top ten (Tables 1 and S1 and Figure S4). Overall, 21 of our test set of 23 diverse α-helical transmembrane proteins are reliably predicted, with template modeling (TM) scores of 0.5–0.7 and Cα-rmsd 2.6–4.8 Å over > 70% of the length (Figures 3A and 3B and Tables 1 and S1). Template modeling score (range 0.0–1.0) is considered reasonable when >0.5 and is comparable across proteins of varying lengths (Zhang and Skolnick, 2004). This blindly predicted set allows assessment of the relationship between the number of evolutionary constraints that are not spatially close in the cognate crystal structure (false positives) and the accuracy of our 3D structure prediction (Figure S5). The highest-ranked evolutionary constraints (1–20) contain ~2% false positives, and the proportion of true positives decreases monotonically as a function of the number of constraints (Figure S3). However, the accuracy of folding, as measured by TM score, is remarkably robust to variation in the proportion of true positives and is stable over many different folding experiments in which the number of constraints is steadily increased (Figure 3C). Details of the distribution of predicted contacts along the protein chain and the precise nature of false positives, such as mutual effective cancellation, may contribute to this robustness.

Currently, state-of-the-art approaches for de novo folding are based primarily on searching for sequence-similar fragments in 3D structure databases followed by fragment assembly using specially designed empirical force fields. The key limitation is the enormous size of the conformational search space. Our approach overcomes this limitation by using the information in the evolutionary constraints and its direct translation to 3D coordinates via distance geometry, leading to a considerable performance advantage relative to earlier methods. The advantage is apparent in terms of (1) protein size range, (2) prediction accuracy, (3) efficiency of conformational search, and (4) lack of dependence on fragments and helix-helix contacts from previously solved 3D structures. More than 50% of membrane proteins have 8–14 transmembrane helices. Here, we report models of proteins with up to 14 helices and anticipate that our

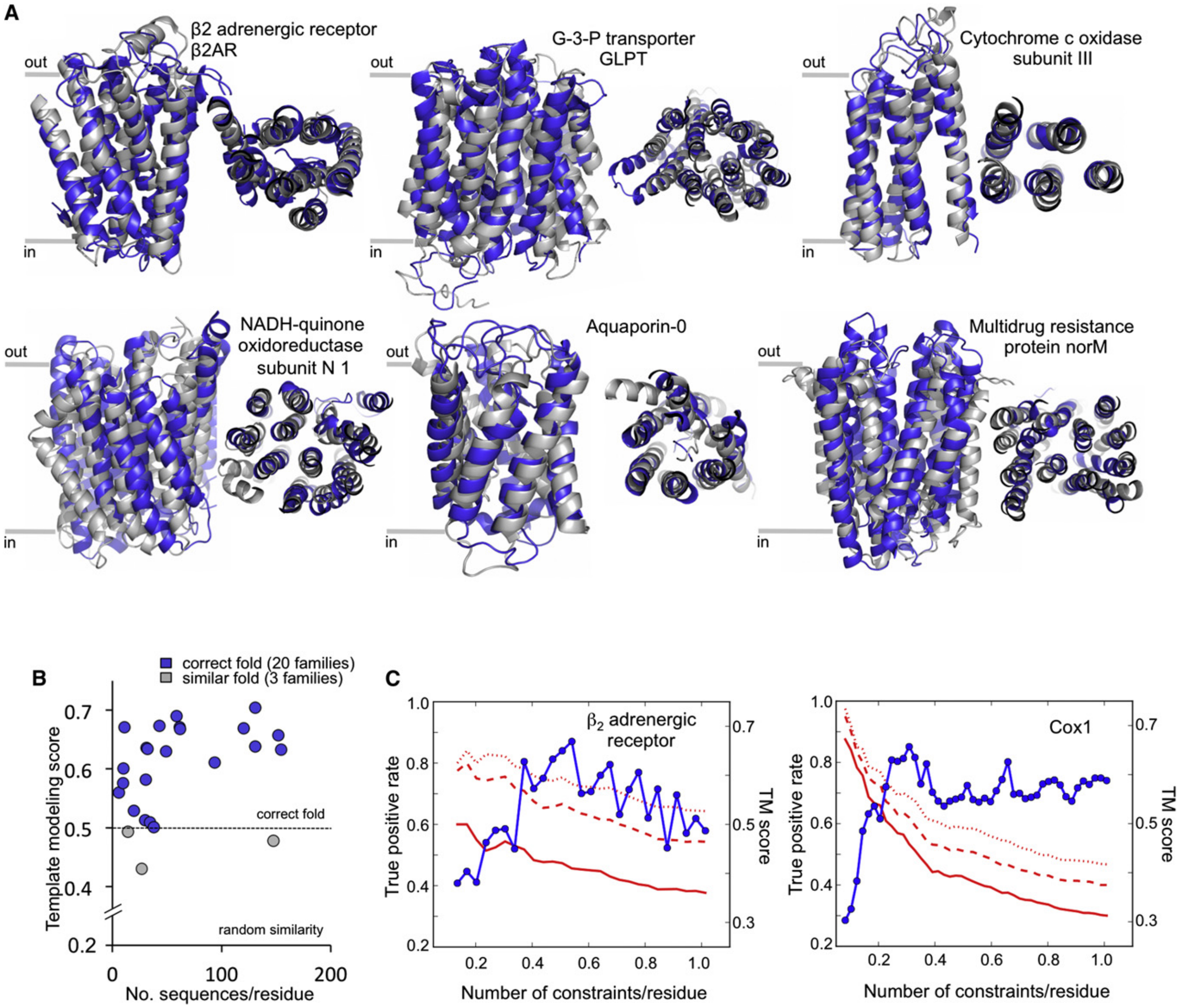

Figure 3. Accuracy of Blinded 3D Structure Prediction for Candidates with Known Structure

(A) Structural superpositions of predicted structures (blue) onto experimental structures (gray). First panel for each protein: side view from within the membrane; second panel: top-down view from noncytoplasmic side. All figures were rendered with PyMOL.

(B) Accuracy of 3D structure prediction for candidates with known structure. Template modeling score (TM score) (Zhang and Skolnick, 2004) of the best model for each protein plotted against the number of sequences in the multiple sequence alignment, normalized by modeled protein length.

(C) 3D prediction accuracy is surprisingly stable as the true positive rate of evolutionary constraints decreases, going down the list of ranked ECs. The TM score of the best prediction (blue) and the true positive rate (red) are plotted for increasing numbers of evolutionary constraints (divided by the number of residues in the protein to allow comparison between proteins). Distance cut-offs to define true contacts of true positive rate are 5 Å (red dots), 7 Å (red dashes), and 8Å (red). See also Figures S3 and S4 and Data S2–S5.

method will allow the prediction of even larger membrane proteins, as we see no deterioration of accuracy with size (up to almost 500 residues) and obtain accurate 3D fold with as little as one constraint per residue over the entire size range. In contrast, previous prediction tools have been used to generate models for proteins with only four to seven helices reported (Barth et al., 2009). To compare accuracy, we predicted structures for five of the same proteins predicted by Barth et al. (2009) (Table S2). Our method reached the threshold coordinate accuracy of 4 Å over comparable or significant larger regions (e.g., 89% rather than 40% of residues for bovine rhodopsin), and it explored conformational search space more efficiently (e.g., ∼500 candidate models compared to 200,000 generated in Barth et al. [2009]). As a result of this efficiency gain, in current practice, EVfold all-atom models can be generated on a laptop in a few minutes per structure, without the need for supercomputers. A possible conceptual and practical advantage of the EVfold_membrane method is information about the roles of residues and residue interactions in protein function as a result of extracting coupling information at the protein level filtered through functional selection over a myriad of evolutionary experiments.

Although the results from our validation set of proteins are encouraging, they raise the question of whether we can predict the success of our approach for any given protein of interest, based on sequence information alone. In general, the accuracy of the predicted model increases with the number of sequences in the alignment normalized for the length of the protein (Figure 3B). For instance, the predicted structures of two proteins, a proton/peptide symporter and a bile acid symporter, have the lowest TM scores (0.4–0.5) compared to their cognate crystal structures and have among the lowest number of sequences per residue in their input alignments (26 and 3, Table S1). Conversely, the predicted structure of bovine rhodopsin has 131 sequences per residue and an excellent TM score of 0.7. Thus, the number of sequences, the diversity of sequences, and the coverage of the length of the protein will no doubt be important metrics in estimating the likely accuracy of predictions and will be used to develop metrics for more accurate and more subfamily-specific structure calculations.

Evolutionary Constraints Include Homo-Oligomer Contacts

Not all residue interactions that are strongly constrained by evolution are close in the 3D structure of the monomeric protein. Residue pairs close in transmembrane protein homo-oligomers may thus appear in conflict with other monomer constraints and/or the predicted 3D fold.

For example, in the computed structure of the ABC transporter *S. typhimurium* MsbA, evolutionary couplings between transmembrane helix 2 and transmembrane helices 5 and 6 are false positives with respect to monomer structure but true positives with respect to the crystal structure dimer interface (PDB: 3b60; Ward et al., 2007) (Figure 4A). Similarly, *E. coli* MetI has a cluster of evolutionary couplings with residues that are not in contact in the monomer but form contacts in the dimer (PDB: 3dhw; Kadaba et al., 2008). If successfully identified, the removal of the conflicting oligomer evolutionary couplings from the folding calculation improves the accuracy of prediction for the monomer (test done in MsbA and MetI; data not shown).

We also predict oligomer contacts for proteins of unknown 3D structures, such as AdipoR1. To identify potential dimerization contacts, we noted that some evolutionary constraints are inconsistent with the monomer predicted structure and may therefore be involved in the putative dimerization interface. Interpreting these evolutionary constraints as distance constraints between residues in two separate monomer structures shows that the AdipoR1 dimer interface involves contacts between the loop from helices 4 to 5 and both helices 1 and 7 (Figure 4B). Consistent with our prediction of the dimerization region are experimental observations that mutations in the GXXXG motif on transmembrane helix 5 of AdipoR1 disrupt dimerization (Kosel et al., 2010). Q335 on the transmembrane helix 7 is unusually strongly constrained, in spite of a low 19% conservation level as a single residue, as a partner in more than 11 evolutionary couplings, some of which may be across this putative interface (Figure 4B). These examples suggest that homo-oligomer contact detection using evolutionary coupling pairs may yield valuable testable information. It remains an algorithmic challenge to identify such evolutionary couplings between the components of oligomers in a more automated fashion.

Evolutionary Constraints Reflect Conformational Change

Many proteins can adopt different distinct conformations as part of their function (Tokuriki and Tawfik, 2009). Can we correctly predict more than one 3D conformation of a protein by extracting and analyzing evolutionary couplings from one set of protein sequences? We investigated this challenge by an analysis of known structures and genuine prediction. GlpT and OCTN1 belong to the functionally diverse subfamilies of the large major facilitator superfamily, secondary membrane transporters that move substrate across the membrane by alternating between two alternative conformations of the channel—one open to the cytoplasm and the other open to the periplasm or extracellular space (Boudker and Verdon, 2010; Huang et al., 2003).

Comparing the predicted model of Glpt to the crystal structure 1pw4 (cytoplasm-open conformation), we noticed that the predicted cytoplasmic side of the transporter channel is not as open as in the crystal structure (Figure 3A). The Glpt evolutionary couplings differ from contacts made in the GlpT crystal structure in an apparently false positive set that would, however, be in contact in the suspected alternative cytoplasm-closed conformation (Figure 5A). Similarly, a set of contacts can be identified that are consistent only with the cytoplasm-open conformation (selection rules in Supplemental Information). To test whether the two alternative sets of evolutionary couplings for GlpT protein would be sufficient to predict the two different conformations, we refolded GlpT with both sets separately (Figure 5A and Table S3). As expected, when we exclude evolutionary coupling pairs between the domains on the periplasmic side, we obtain models in a closed-to-cytoplasm conformation, similar in overall structure to the known closed conformation structure of the L-fucose-proton symporter FucP (PDB: 3o7q) and to a homology model of LacY (Radestock and Forrest, 2011) but distinct from the known open GlpT structure of GlpT (PDB: 1pw4). The arrangements of transmembrane helices 5 and 8 and transmembrane helices 2 and 11 in the two folded models differ as expected for “rocking” changes between alternative transporter conformations (Lemieux et al., 2004). Therefore, plausibly, the evolutionary constraints in the sequence family of GlpT, when decomposed into two overlapping sets, reflect two alternative conformations of the channel.

As human OCTN1 (unknown structure) is also from the major facilitator superfamily, we wondered whether evolutionary couplings in OCTN1 also contained information about alternative conformations. We compared our top-ranked model of OCTN1 to all structures in the PDB and found significant hits to known structures in the major facilitator superfamily, including those of GlpT and FucP. The predicted OCTN1 models, as above for GlpT, looked like an intermediate conformation between outward-open and inward-open, consistent with the expectation that both states are constrained by evolution (Figure 1 and

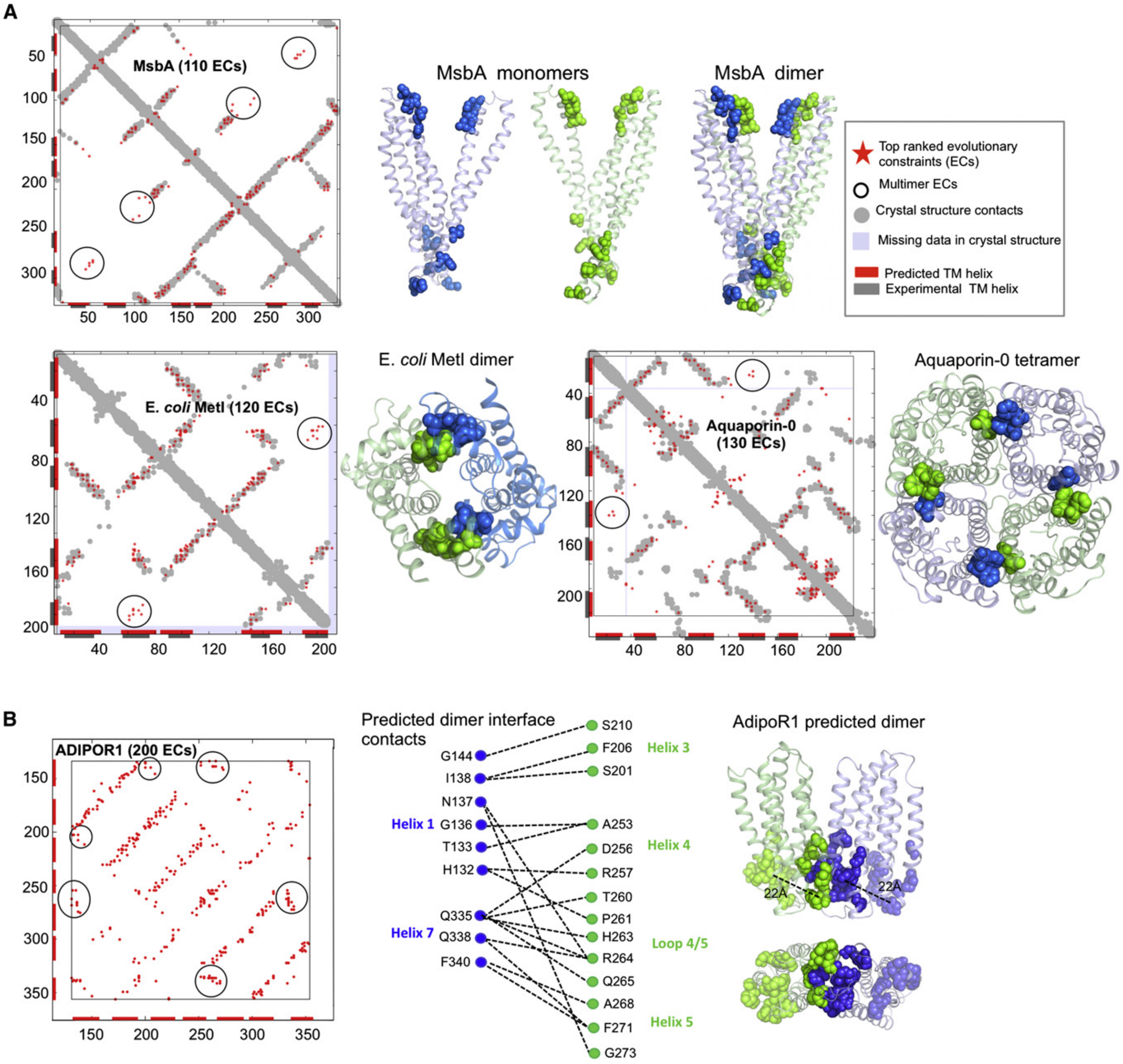

Figure 4. Evolutionary Constraints on Residue Pairs in Oligomerization Interfaces

Contact maps of top-ranked predicted ECs (red stars in A and B) overlaid on crystal structure contacts (gray, known only in A). Residue pairs coevolving due to intermonomer contacts in the homo-oligomer (black circles) in an overlay of top-ranked predicted evolutionary constraints (red) experimental structure contacts (gray), where known, on contact maps for each protein. In the monomer (blue or green ribbon with blue or green residue balls), the corresponding residue pairs would be false positive contacts (blue with blue or green with green do not make contact in the monomer) but would be true positives in the homo-oligomer structure (contacting blue-green pairs).

(A) Four examples of inference of oligomer contacts from ECs of known 3D structures.

(B) Predicted dimer contacts of AdipoR1, shown on predicted monomer structures. EC pairs (black circles) at a large distance in monomer structure (~23 Å, green with green, blue with blue) are close (green-blue contact pair) in predicted dimers. Predicted dimer cartoon (right) is a rough estimate, produced by manual-visual docking of monomers, satisfying the majority of predicted dimer interface EC pairs (middle).

See also Figure S5.

Data S2). Examination of the distribution of EC pairs suggests that they contain information for two conformations of the transporter (Figure 5C).

Given that our evolutionary constraints contain information about the different states of members of the major facilitator superfamily, we anticipate that evolutionary constraints might

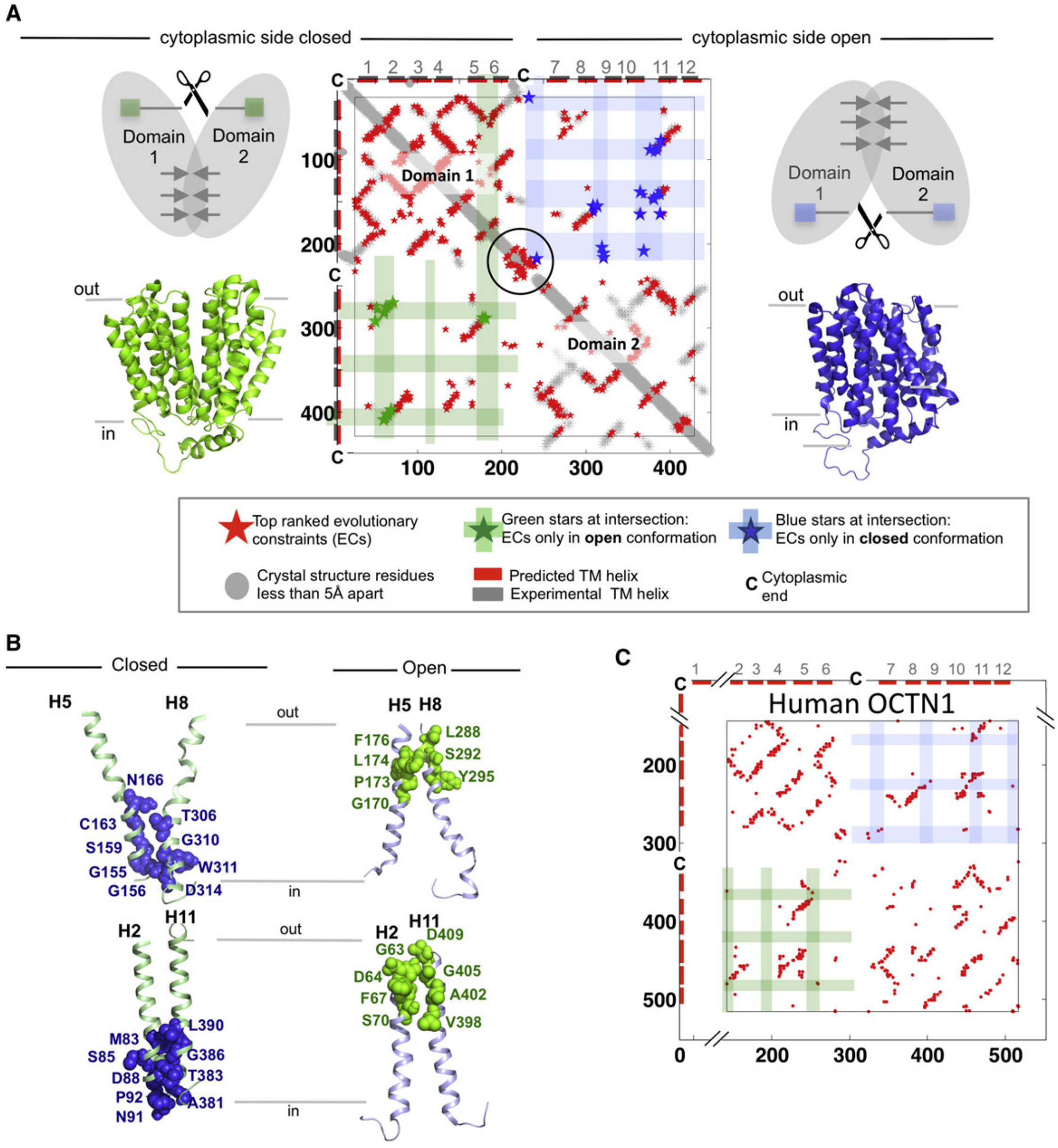

Figure 5. Coevolved Pairs Consistent with Open and Closed Conformations of Proteins in the Major Facilitator Family

(A) (Center) Contact map for *E. coli* GlpT, residues less than 5 Å apart in the crystal structure (gray, PDB: 1pw4) overlaid with the top 350 ECs (red stars). The similarity of the upper-left and lower-right quadrants reflect the similarity of the structure and sequences of the two domains. Upper-right and lower-left quadrants show the predicted interdomain contacts (all stars). Stripes in lower-left and upper-right quadrants cover interdomain contacts involving the periplasmic end of the helices/loops (green strips, lower-left) and the cytoplasmic ends of the helices/loops (blue strips, upper-right). Predicted ECs located where stripes of the same color cross each are likely interdomain contacts (green and blue stars) (Table S3). (Right and left) Refolded GlpT from extended polypeptide excluding blue constraints for cytoplasmic side open (right) and excluding green constraints for cytoplasmic side closed (left). The schematics (right and left top) indicate contacts used (arrows) and not used (scissors) in refolding to get the two alternative conformations. Open conformation (right) is similar to crystal structure (Table 1) and is reproduced via refolding; closed conformation structure (left) is previously unknown and predicted here via refolding.

(B) Details from the models in (A). The two pairs of helices (H5/8 and H2/11) in the predicted models of GlpT are thought to change conformation dependent on state of substrate binding (closed at cytoplasm, green ribbons, left; open at cytoplasm, blue ribbons, right). Differences in interhelical angles are driven by the alternative use of top (green) or bottom (blue) contact pairs derived from ECs in refolding (Table S3).

(C) Predicted EC pairs of human OCTN1 (red stars on contact map) determine the overall fold. Stripes in lower-left and upper-right quadrants cover the predicted periplasmic end of the helices/loops (green) and the cytoplasmic ends of the helices/loops (blue). Predicted evolutionary constraints (not differentiated by star color) located where stripes of the same color cross each other are predicted interdomain contacts. 3D structures of alternative conformations of OCTN1 are not shown here. For predicted OCTN1 structure details, see Figure 1, Table 1, and Data S2.

See also Table S3.

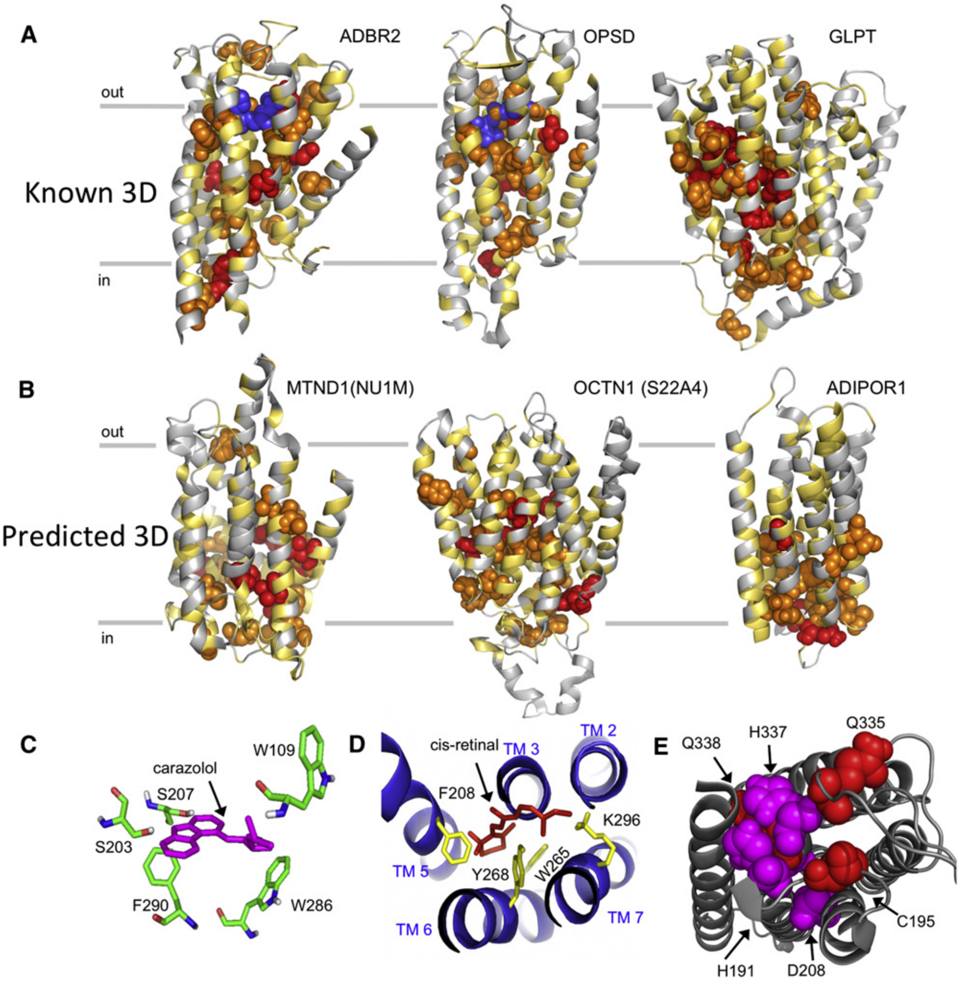

Figure 6. Known Functional Sites Contain Residues Strongly Involved in Evolutionary Constraints

(A and B) The total evolutionary coupling score on individual residues reflects likely functional involvement (top 5%, red spheres; top 6%–15%, orange spheres; all others, yellow ribbon); scores are as in Table S4.

(A) The ligands carazolol in Adrb2 and retinal in Opsd (blue spheres) were positioned in the predicted structure by globally superimposing the most accurate predicted model and the experimental structure plus ligand (experimental structures are not shown; no docking was performed).

(B) Residues with high evolutionary coupling scores mapped on the predicted structures of unknown structure transmembrane proteins.

(C and D) Above average accuracy of blinded prediction of atomic positions of the binding site of Adrb2 (1.6 Å Cα-rmsd over 9 residues, C) and bovine rhodopsin (1.8 Å Cα-rmsd over 10 residues, D).

(E) Likely functional residues (high evolutionary coupling scores) in AdipoR1 on the predicted cytoplasmic side (known functional residues in magenta, predicted functional residues in red).

See also Table S4.

help to unravel the precise conformational changes upon substrate binding and transport.

Evolutionary Constraints Mark Functional Residues

Conservation of amino acids in proteins in single columns is routinely used to infer functional importance of the site and assess the consequences of genetic variation. As our evolutionary analysis reflects both residue-residue correlations and single residue terms, we wondered whether the strength of evolutionary couplings on a residue is an indication of its general functional importance for the protein. To assess this, we calculate the total evolutionary coupling score for a given residue by summing the evolutionary coupling values over all high-ranking pairs involving that residue (Experimental Procedures, Table S4, Data S6). We find that, in Adrb2, Opsd, and GlpT, residues with high total coupling scores line the substrate-binding sites and affect signaling or transport; for instance, W109, D113, and Y141 in Adrb2; K296, W265, and H211 in Opsd; and Y393, H165, and K90 in GlpT (Huang et al., 2003; Law et al., 2008; Valiquette et al., 1995) (Figure 6A). Higher prediction accuracy of atomic coordinates near the active sites for Adrb2 and Opsd than for the average of the protein reflects the multiple constraints, i.e., high total coupling score, on these sites (Figure 6C).

In the unknown structure AdipoR1, residues with a high total coupling score include putative enzymatic residues S187, H191, D208, H337, and H341 (Holland et al., 2011; Pei et al., 2011) together with the top three high-scoring residues, C195, A235, and Q335, which cluster together (within ~4 Å) in the predicted 3D structure, indicating that they are important in the activity of AdipoR1 (Figure 6B). Similarly, clusters of residues with high scoring in OCTN1 make potential salt bridges at the cytoplasmic side of the domains (169R-220E, 397R-450E), cluster in the central transport pore (N210, Y211, C236, E381, and R469), and are potentially involved in conformational changes. Residues with high total coupling scores in our predicted models of human MT-ND1 are clustered in a periplasmic-oriented pocket and along the mitochondrial interface with the hydrophilic domain and the putative quinine-binding site (Figure 6B) (Efremov and Sazanov, 2011). Mutations in MT-ND1 at residues Y30 and M31 are associated with Alzheimer's disease and Leber's hereditary optic neuropathy (LHON) (Johns et al., 1992), and these two residues have particularly high total coupling scores, suggesting that they are functionally constrained by interactions with several other residues.

We hypothesize that many evolutionary coupling pairs, whether or not close in the 3D structure, may be functionally important. The examples presented here, however, are not the result of an exhaustive analysis. Therefore, reliable functional interpretation of evolutionary constraints, whether indicative of intramonomer contacts or not, remains a challenge. Our results here provide some confidence in the validity of the conceptual link between the strength of evolutionary constraints on a residue and its functional importance, whether through location in binding sites or involvement in conformational changes.

DISCUSSION

The process of evolution and the massive sequencing of diverse species have provided the opportunity to compute an important aspect of molecular phenotype, protein 3D structure, and the EVfold method appears to achieve a useful level of accuracy. However, a serious gap remains between predicted and experimental structures. Though an overall Ca-rmsd of 4–5 Å across hundreds of residues does imply the correct identification of the overall fold, it also implies that particular atomic positions, the interdigitation of packed side chains and loop conformations can be incorrect in detail, although they appear more accurate near heavily constrained binding sites. To improve the quality of the predicted contacts and resulting atomic structures, four areas of focus hold particular promise: (1) improved information handling in sequence space, such as improvements in weighting schemes for sequences, evaluation of alignment diversity, inclusion of higher-order terms, and consistency filters to reduce the number of false positive pairs; (2) automated procedures to distinguish between internal and homo-oligomer pair contacts and to identify contacts reflecting alternative conformations; (3) the use of fragments imported from known structures; and (4) the use of advanced energy refinement methods, including molecular dynamics and Monte Carlo simulations (Dror et al., 2011; MacCallum et al., 2011).

Even at the current level of accuracy, a number of applications may have immediate benefit. One is the development of hybrid methods of structure determinations. In NMR spectroscopy, inclusion of evolutionary constraints from sequences may permit structure determination with a smaller number of chemical shifts and NOEs, saving machine time or permitting the solution of larger protein structures than previously reachable. In protein crystallography, the solution of a 3D structure from a native data set alone may become possible, without the need for heavy atom derivatives or MAD phasing, via molecular replacement searches starting with predicted 3D structures. If successful in future work, such methods would significantly increase the productivity of structural biology and the rate of solving new structures.

Beyond structure determination, the predicted models may be useful for pharmacological selection of compounds via docking calculations. The observation of exceptionally strong evolutionary constraints near active sites, as reported here for a few proteins, is a favorable starting point, as the accuracy of protein coordinates in active sites and binding sites is an important requirement for computational drug screening. In molecular biology in general, the placement of constrained pairs in the context of known or predicted 3D structures may also provide useful information to guide functional mutational experiments. Similarly, evolutionarily coupled pairs may be excellent design elements for engineering new proteins in synthetic biology (Russ et al., 2005) and may have a strategic role in the protein folding process (Fersht, 2008).

Inferred evolutionary constraints may also help guide the computational assembly of protein monomers into complexes, with or without low-resolution information from electron diffraction or similar methods. The computational extension to predict the structure of protein complexes is a straightforward generalization using pairwise sequence alignments, with a homologous pair of sequences in place of a single sequence and derivation of evolutionary couplings not within a protein but between two potentially interacting proteins (Pazos and Valencia, 2002; Skerker et al., 2008; Weigt et al., 2009). We see no practical limit in the size of complexes accessible to such computation, provided sufficiently diverse sequence information is available, as the configuration of even large complexes with tens of constituents effectively can be deduced from calculation of all pairwise protein interactions in the complex. The nuclear pore complex, as solved by computational assembly from protein-protein pair information determined experimentally, would be an excellent test case (Fernandez-Martinez et al., 2012).

Looking forward, how much information about 3D folds of transmembrane proteins can be gained if this kind of method is broadly and successfully applied? A current snapshot of protein families, as organized in the PFAM 26.0 database, has about 2 million transmembrane proteins in 1,259 protein families, of which 107 families have one or more 3D structures. An additional 150 families appear to have sufficient sequences to be modeled using evolutionary couplings, including those with β sheet folds (not tested here). Given the current efficiency and rapid development of DNA sequencing technology, perhaps another 500 families would accrue similar levels of sequence information to have their folds determined in about 2 years, with subsequent rapid growth likely.

On a practical level, the simplicity of the theoretical approach and efficiency of the computational implementation, with computation in a couple of hours for proteins up to 500 residues, will allow availability of the EVfold procedure to a broad community of researchers, not limited to structural biologists, in either precomputed or server mode. For the proteins described here, detailed data, such as 3D coordinates and evolutionary constraints, as well as software code for their calculation, are available at http://evfold.org/transmembrane. Computational protein folding using evolutionary constraints may thus drive new experimental approaches that will harness the massive explosion in genomic sequencing by reading the evolutionary footprints of protein structure and function.

EXPERIMENTAL PROCEDURES

Full methods are described in the Extended Experimental Procedures. All EC scores and residue name mappings are in Data S1, and all 3D model coordinates, input files, and analysis are in Data S2–S5 and online at http://evfold.org/transmembrane.

Selection of Membrane Proteins

To test our ability to predict the 3D structure of a-helical multipass membrane proteins ($\geq$5 helices), we compiled a set of 25 proteins from 23 different Pfam families from the database of membrane proteins of known 3D structure (http://blanco.biomol.uci.edu/mpstruc/listAll/list). We optimized the set for nonredundancy and depth of sequence alignment. The set of interesting membrane protein families with no known representative structure was chosen by selecting transmembrane proteins that are drug targets, using the DrugBank (Knox et al., 2011), Pfam (Punta et al., 2012), and CAMPS (Neumann et al., 2012) databases. For this initial study, we selected proteins with at least 2 × L sequences in the family alignment (L = protein length) and with more than 70% coverage (breadth).

Multiple Sequence Alignments

Multiple sequence alignments for each candidate protein were obtained using HHblits (Remmert et al., 2012) sequence searches against the UniProt database at a range of different E values. The alignment used for constraint inference was selected by choosing the E value giving the best trade-off between a maximum number of sequences in the alignment and sufficient coverage of the entire transmembrane domain by most sequences in the alignment (all alignments are available at http://www.evfold.org/transmembrane) (Figure 2A and Figure S2).

Inference of Evolutionary Constraints from Sequence Variation

To predict the 3D structure of membrane proteins, we devised a membrane protein-specific version of the original EVfold method (Marks et al., 2011) and named it EVfold_membrane. First, a set of evolutionary couplings between residue pairs is inferred by computing the parameters in a global maximum entropy probability model of the multiple sequence alignment (Data S1). This set is ranked according to coupling strength and filtered for inconsistency with predicted membrane topology and predicted secondary structure.

Ab Initio Folding from Membrane Protein Sequence

All predictions started from fully extended polypeptide using increasing numbers of evolutionary constraints, from 40 to L constraints (L = length of modeled sequence) in steps of 10, with 20 models generated for each EC bin. Additionally, we added distance and dihedral angle constraints consistent with predicted secondary structure. The folding protocol uses default modules from the CNS software suite (Brunger, 2007), which consists of distance geometry, simulated annealing, and energy minimization stages. Each model takes about 1–2 min of computing time on a single CPU for a protein of average size.

Clustering and Ranking of Predicted Models

Although only a small number of models are generated, we devised a ranking scheme based on simple intuitive requirements for membrane proteins, such as satisfaction of unused constraints (adapted from Miller and Eisenberg, 2008), predicted secondary structure, and predicted lipid exposure agreement in the folded models. Structures are additionally clustered using MaxCluster (Siew et al., 2000) single-linkage clustering to eliminate high-ranked outliers belonging to small clusters.

Assessment of Evolutionary Constraints and Prediction Quality

Predicted evolutionary constraints were compared to observed contacts from crystal structures using contact maps and false positive rate plots (Data S2–S5 and http://evfold.org/transmembrane). Predicted models of known structures were compared to a representative crystal structure (Table S1) using the Cα-rmsd and TM and GDT scores calculated with MaxCluster (Zemla, 2003; Zhang and Skolnick, 2005). Predicted 3D structures for transmembrane proteins of unknown structure, for which no family member structure has been solved yet, are compared for structural homology against a representative set of proteins in the Protein Data Bank (PDB) using structural alignments with DALI (Holm and Sander, 1993) and FATCAT (Ye and Godzik, 2004).

Residues with High Total Evolutionary Coupling Scores

To quantify the strength of evolutionary constraints on a residue, we calculated the total strength of evolutionary constraints per residue. For each residue, we sum the pair scores obtained from the maximum entropy model over all high-ranking pairs that it is involved in, down to a predefined cutoff (Data S1). The score for each residue is normalized by the average score for all residues in the full sequence (Table S4).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, four tables, and six data files and can be found with this article online at doi:10.1016/j.cell.2012.04.012.

ACKNOWLEDGMENTS

We thank Guy Montelione, Rebecca Ward, Nicholas Stroustrup, Steven Long, Nikola Pavletich, Anil Korkut, Edda Kloppman, Marc Offman, Yitzak Pilpel, Andrea Pagnani, Riccardo Zecchina, David Baker, Richard Stein, James Thompson, and Michael Lappe for scientific discussions; Johannes Soeding and Michael Remmert for help with HHblits; Marco Punta and Jaina Mistry for help with PFAM numbers; and Nathan Marks-Forder for graphical abstract image. C.S. and R.S. have support from the Dana-Farber Cancer Institute-Memorial Sloan-Kettering Cancer Center Physical Sciences Oncology Center (NIH U54-CA143798). L.J.C. is supported by an Engineering and Physical Sciences Research Council fellowship (EP/H028064/1). T.A.H. has support from the German National Academic Foundation. D.S.M. thanks the Systems Biology Department at Harvard Medical School for their generous support.

Received: March 9, 2012
Revised: April 12, 2012
Accepted: April 23, 2012
Published online: May 10, 2012

REFERENCES

Aldahmesh, M.A., Mohamed, J.Y., Alkuraya, H.S., Verma, I.C., Puri, R.D., Alaiya, A.A., Rizzo, W.B., and Alkuraya, F.S. (2011). Recessive mutations in ELOVL4 cause ichthyosis, intellectual disability, and spastic quadriplegia. Am. J. Hum. Genet. *89*, 745–750.

Bakheet, T.M., and Doig, A.J. (2009). Properties and identification of human protein drug targets. Bioinformatics *25*, 451–457.

Barth, P., Wallner, B., and Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. Proc. Natl. Acad. Sci. USA *106*, 1409–1414.

Boudker, O., and Verdon, G. (2010). Structural perspectives on secondary active transporters. Trends Pharmacol. Sci. *31*, 418–426.

Brunger, A.T. (2007). Version 1.2 of the Crystallography and NMR system. Nat. Protoc. *2*, 2728–2733.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Burger, L., and van Nimwegen, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput. Biol. *6*, e1000633.

Chen, Y.H., Hu, L., Punta, M., Bruni, R., Hillerich, B., Kloss, B., Rost, B., Love, J., Siegelbaum, S.A., and Hendrickson, W.A. (2010). Homologue structure of the SLAC1 anion channel for closing stomata in leaves. Nature *467*, 1074–1080.

Cherezov, V., Rosenbaum, D.M., Hanson, M.A., Rasmussen, S.G., Thian, F.S., Kobilka, T.S., Choi, H.J., Kuhn, P., Weis, W.I., Kobilka, B.K., and Stevens, R.C. (2007). High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science *318*, 1258–1265.

Choe, H.W., Kim, Y.J., Park, J.H., Morizumi, T., Pai, E.F., Krauss, N., Hofmann, K.P., Scheerer, P., and Ernst, O.P. (2011). Crystal structure of metarhodopsin II. Nature *471*, 651–655.

Cronet, P., Sander, C., and Vriend, G. (1993). Modeling of transmembrane seven helix bundles. Protein Eng. *6*, 59–64.

Dang, S., Sun, L., Huang, Y., Lu, F., Liu, Y., Gong, H., Wang, J., and Yan, N. (2010). Structure of a fucose transporter in an outward-open conformation. Nature *467*, 734–738.

Dempster, A.P. (1972). Covariance Selection. Biometrics *28*, 157–175.

Doyle, L.A., Yang, W., Abruzzo, L.V., Krogmann, T., Gao, Y., Rishi, A.K., and Ross, D.D. (1998). A multidrug resistance transporter from human MCF-7 breast cancer cells. Proc. Natl. Acad. Sci. USA *95*, 15665–15670.

Dror, R.O., Pan, A.C., Arlow, D.H., Borhani, D.W., Maragakis, P., Shan, Y., Xu, H., and Shaw, D.E. (2011). Pathway and mechanism of drug binding to G-protein-coupled receptors. Proc. Natl. Acad. Sci. USA *108*, 13118–13123.

Efremov, R.G., and Sazanov, L.A. (2011). Structure of the membrane domain of respiratory complex I. Nature *476*, 414–420.

Efremov, R.G., Baradaran, R., and Sazanov, L.A. (2010). The architecture of respiratory complex I. Nature *465*, 441–445.

Fatakia, S.N., Costanzi, S., and Chow, C.C. (2009). Computing highly correlated positions using mutual information and graph theory for G protein-coupled receptors. PLoS ONE *4*, e4681.

Fernandez-Martinez, J., Phillips, J., Sekedat, M.D., Diaz-Avalos, R., Velazquez-Muriel, J., Franke, J.D., Williams, R., Stokes, D.L., Chait, B.T., Sali, A., and Rout, M.P. (2012). Structure-function mapping of a heptameric module in the nuclear pore complex. J. Cell Biol. *196*, 419–434.

Fersht, A.R. (2008). From the first protein structures to our current knowledge of protein folding: delights and skepticisms. Nat. Rev. Mol. Cell Biol. *9*, 650–654.

Fodor, A.A., and Aldrich, R.W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins *56*, 211–221.

Fuchs, A., Martin-Galiano, A.J., Kalman, M., Fleishman, S., Ben-Tal, N., and Frishman, D. (2007). Co-evolving residues in membrane proteins. Bioinformatics *23*, 3312–3319.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins *18*, 309–317.

Holland, W.L., Miller, R.A., Wang, Z.V., Sun, K., Barth, B.M., Bui, H.H., Davis, K.E., Bikman, B.T., Halberg, N., Rutkowski, J.M., et al. (2011). Receptor-mediated activation of ceramidase activity initiates the pleiotropic actions of adiponectin. Nat. Med. *17*, 55–63.

Holm, L., and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. J. Mol. Biol. *233*, 123–138.

Holm, L., and Sander, C. (1995). Dali: a network tool for protein structure comparison. Trends Biochem. Sci. *20*, 478–480.

Horn, F., Bywater, R., Krause, G., Kuipers, W., Oliveira, L., Paiva, A.C., Sander, C., and Vriend, G. (1998). The interaction of class B G protein-coupled receptors with their hormones. Receptors Channels *5*, 305–314.

Howell, N., Bindoff, L.A., McCullough, D.A., Kubacka, I., Poulton, J., Mackey, D., Taylor, L., and Turnbull, D.M. (1991). Leber hereditary optic neuropathy: identification of the same mitochondrial ND1 mutation in six pedigrees. Am. J. Hum. Genet. *49*, 939–950.

Huang, Y., Lemieux, M.J., Song, J., Auer, M., and Wang, D.N. (2003). Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli. Science *301*, 616–620.

Jaksch, M., Hofmann, S., Kaufhold, P., Obermaier-Kusser, B., Zierz, S., and Gerbitz, K.D. (1996). A novel combination of mitochondrial tRNA and ND1 gene mutations in a syndrome with MELAS, cardiomyopathy, and diabetes mellitus. Hum. Mutat. *7*, 358–360.

Johns, D.R., Neufeld, M.J., and Park, R.D. (1992). An ND-6 mitochondrial DNA mutation associated with Leber hereditary optic neuropathy. Biochem. Biophys. Res. Commun. *187*, 1551–1557.

Jones, D.T., Buchan, D.W., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics *28*, 184–190.

Kadaba, N.S., Kaiser, J.T., Johnson, E., Lee, A., and Rees, D.C. (2008). The high-affinity E. coli methionine ABC transporter: structure and allosteric regulation. Science *321*, 250–253.

Katritch, V., Cherezov, V., and Stevens, R.C. (2012). Diversity and modularity of G protein-coupled receptor structures. Trends Pharmacol. Sci. *33*, 17–27.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. (2011). DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs. Nucleic Acids Res. *39* (Database issue), D1035–D1041.

Kosel, D., Heiker, J.T., Juhl, C., Wottawah, C.M., Blüher, M., Mörl, K., and Beck-Sickinger, A.G. (2010). Dimerization of adiponectin receptor 1 is inhibited by adiponectin. J. Cell Sci. *123*, 1320–1328.

Lapedes, A.S., Giraud, B.G., Liu, L.C., and Stormo, G.D. (1997). Correlated Mutations in Protein Sequences: Phylogenetic and Structural Effects (Santa Fe, New Mexico: Santa Fe Institute).

Law, C.J., Almqvist, J., Bernstein, A., Goetz, R.M., Huang, Y., Soudant, C., Laaksonen, A., Hovmöller, S., and Wang, D.N. (2008). Salt-bridge dynamics control substrate-induced conformational change in the membrane transporter GlpT. J. Mol. Biol. *378*, 828–839.

Lemieux, M.J., Huang, Y., and Wang, D.N. (2004). The structural basis of substrate translocation by the Escherichia coli glycerol-3-phosphate transporter: a member of the major facilitator superfamily. Curr. Opin. Struct. Biol. *14*, 405–412.

Livesay, D.R., Kreth, K.E., and Fodor, A.A. (2012). A critical evaluation of correlated mutation algorithms and coevolution within allosteric mechanisms. Methods Mol. Biol. *796*, 385–398.

Locher, K.P., Lee, A.T., and Rees, D.C. (2002). The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism. Science *296*, 1091–1098.

Long, S.B., Tao, X., Campbell, E.B., and MacKinnon, R. (2007). Atomic structure of a voltage-dependent K+ channel in a lipid membrane-like environment. Nature *450*, 376–382.

MacCallum, J.L., Pérez, A., Schnieders, M.J., Hua, L., Jacobson, M.P., and Dill, K.A. (2011). Assessment of protein structure refinement in CASP9. Proteins *79* (*Suppl 10*), 74–90.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS ONE *6*, e28766.

Meinshausen, N., and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. Ann. Stat. *34*, 1436–1462.

Miller, C.S., and Eisenberg, D. (2008). Using inferred residue contacts to distinguish between correct and incorrect protein models. Bioinformatics *24*, 1575–1582.

Miller, A.N., and Long, S.B. (2012). Crystal structure of the human two-pore domain potassium channel K2P1. Science *335*, 432–436.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. USA *108*, E1293–E1301.

Murzin, A.G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. EMBO J. *12*, 861–867.

Natarajan, K., Xie, Y., Baer, M.R., and Ross, D.D. (2012). Role of breast cancer resistance protein (BCRP/ABCG2) in cancer drug resistance. Biochem. Pharmacol. *83*, 1084–1103.

Nemoto, W., Imai, T., Takahashi, T., Kikuchi, T., and Fujita, N. (2004). Detection of pairwise residue proximity by covariation analysis for 3D-structure prediction of G-protein-coupled receptors. Protein J. *23*, 427–435.

Neumann, S., Hartmann, H., Martin-Galiano, A.J., Fuchs, A., and Frishman, D. (2012). Camps 2.0: Exploring the sequence and structure space of prokaryotic, eukaryotic, and viral membrane proteins. Proteins *80*, 839–857.

Overington, J.P., Al-Lazikani, B., and Hopkins, A.L. (2006). How many drug targets are there? Nat. Rev. Drug Discov. *5*, 993–996.

Pazos, F., and Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins *47*, 219–227.

Pei, J., Millay, D.P., Olson, E.N., and Grishin, N.V. (2011). CREST—a large and diverse superfamily of putative transmembrane hydrolases. Biol. Direct *6*, 37.

Peltekova, V.D., Wintle, R.F., Rubin, L.A., Amos, C.I., Huang, Q., Gu, X., Newman, B., Van Oene, M., Cescon, D., Greenberg, G., et al. (2004). Functional variants of OCTN cation transporter genes are associated with Crohn disease. Nat. Genet. *36*, 471–475.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. Nucleic Acids Res. *40* (Database issue), D290–D301.

Radestock, S., and Forrest, L.R. (2011). The alternating-access mechanism of MFS transporters arises from inverted-topology repeats. J. Mol. Biol. *407*, 698–715.

Rasmussen, S.G., Choi, H.J., Fung, J.J., Pardon, E., Casarosa, P., Chae, P.S., Devree, B.T., Rosenbaum, D.M., Thian, F.S., Kobilka, T.S., et al. (2011). Structure of a nanobody-stabilized active state of the β(2) adrenoceptor. Nature *469*, 175–180.

Rasmussen, S.G., Choi, H.J., Rosenbaum, D.M., Kobilka, T.S., Thian, F.S., Edwards, P.C., Burghammer, M., Ratnala, V.R., Sanishvili, R., Fischetti, R.F., et al. (2007). Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature *450*, 383–387.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175.

Russ, W.P., Lowery, D.M., Mishra, P., Yaffe, M.B., and Ranganathan, R. (2005). Natural-like function in artificial WW domains. Nature *437*, 579–583.

Siew, N., Elofsson, A., Rychlewski, L., and Fischer, D. (2000). MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics *16*, 776–785.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043–1054.

Tokuriki, N., and Tawfik, D.S. (2009). Protein dynamism and evolvability. Science *324*, 203–207.

Valiquette, M., Parent, S., Loisel, T.P., and Bouvier, M. (1995). Mutation of tyrosine-141 inhibits insulin-promoted tyrosine phosphorylation and increased responsiveness of the human beta 2-adrenergic receptor. EMBO J. *14*, 5542–5549.

Ward, A., Reyes, C.L., Yu, J., Roth, C.B., and Chang, G. (2007). Flexibility in the ABC transporter MsbA: Alternating access with a twist. Proc. Natl. Acad. Sci. USA *104*, 19005–19010.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. Proc. Natl. Acad. Sci. USA *106*, 67–72.

Yamauchi, T., Kamon, J., Ito, Y., Tsuchida, A., Yokomizo, T., Kita, S., Sugiyama, T., Miyagishi, M., Hara, K., Tsunoda, M., et al. (2003). Cloning of adiponectin receptors that mediate antidiabetic metabolic effects. Nature *423*, 762–769.

Yarov-Yarovoy, V., Schonbrun, J., and Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. Proteins *62*, 1010–1025.

Ye, Y., and Godzik, A. (2004). FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res. *32*(Web Server issue), W582–W585.

Zemla, A. (2003). LGA: A method fdor finding 3D similarities in protein structures. Nucleic Acids Res. *31*, 3370–3374.

Zhang, K., Kniazeva, M., Han, M., Li, W., Yu, Z., Yang, Z., Li, Y., Metzker, M.L., Allikmets, R., Zack, D.J., et al. (2001). A 5-bp deletion in ELOVL4 is associated with two related forms of autosomal dominant macular dystrophy. Nat. Genet. *27*, 89–93.

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

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309.