Interplay between hydrophobicity and the positive-inside rule in determining membrane-protein topology

Assaf Elazar[a,1], Jonathan Jacob Weinstein[a,1], Jaime Prilusky[b], and Sarel Jacob Fleishman[a,2]

[a]Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot 76100, Israel; and [b]Bioinformatics and Biological Computing Unit, Weizmann Institute of Science, Rehovot 76100, Israel

Edited by Barry Honig, Howard Hughes Medical Institute, Columbia University, New York, NY, and approved July 19, 2016 (received for review April 12, 2016)

The energetics of membrane-protein interactions determine protein topology and structure: hydrophobicity drives the insertion of helical segments into the membrane, and positive charges orient the protein with respect to the membrane plane according to the positive-inside rule. Until recently, however, quantifying these contributions met with difficulty, precluding systematic analysis of the energetic basis for membrane-protein topology. We recently developed the dsTβL method, which uses deep sequencing and in vitro selection of segments inserted into the bacterial plasma membrane to infer insertion-energy profiles for each amino acid residue across the membrane, and quantified the insertion contribution from hydrophobicity and the positive-inside rule. Here, we present a topology-prediction algorithm called TopGraph, which is based on a sequence search for minimum dsTβL insertion energy. Whereas the average insertion energy assigned by previous experimental scales was positive (unfavorable), the average assigned by TopGraph in a nonredundant set is −6.9 kcal/mol. By quantifying contributions from both hydrophobicity and the positive-inside rule we further find that in about half of large membrane proteins polar segments are inserted into the membrane to position more positive charges in the cytoplasm, suggesting an interplay between these two energy contributions. Because membrane-embedded polar residues are crucial for substrate binding and conformational change, the results implicate the positive-inside rule in determining the architectures of membrane-protein functional sites. This insight may aid structure prediction, engineering, and design of membrane proteins. TopGraph is available online (topgraph.weizmann.ac.il).

membrane insertion | topology prediction | positive-inside rule | Bellman–Ford search

The plasma membrane is a complex physical environment comprising a hydrophobic core and a polar exterior, which is more negatively charged on its cytoplasmic side (1). Two hallmarks of membrane proteins are hydrophobic segments that span the membrane core, and positive charges at the membrane–cytoplasm interface (the positive-inside rule; ref. 2); these features drive insertion and orient segments relative to the membrane plane, respectively. Furthermore, recent work has shown that positive charges placed close to engineered segments can drive membrane insertion even of marginally polar segments (3, 4), suggesting a role for the positive-inside rule in insertion, and emphasizing the importance of accurate models of membrane-protein energetics for protein engineering and for understanding the physical basis of membrane-protein topology.

Topology prediction is a stringent test of our models of membrane-protein energetics. The most parsimonious membrane-topology predictor would locate the membrane-spanning segments and determine their orientations by a sequence search for minimum insertion energy. To achieve that, however, the insertion energy scale must at a minimum exhibit two properties: (*i*) to drive membrane insertion, the hydrophobic amino acids must make large contributions to insertion; and (*ii*) to orient the protein with respect to the membrane plane, positively charged residues must be strongly favored in the cytoplasm over the extracellular space. These properties, however, were not observed in previous insertion scales (5–9); instead, topology predictors have relied on machine learning, and used experimentally determined membrane-protein structures to train predictors with hundreds of fitting parameters (10–15). Although the accuracy of these predictors is high (80–90%; refs. 10–16), statistics-based predictors cannot be used to systematically investigate the interplay between different energy contributions to topology. Moreover, such methods are less useful than energy-based methods in predicting topology in targets that lack homology to previously characterized proteins (16), and cannot be used to design new proteins.

In a landmark study, von Heijne and coworkers measured translocon-mediated apparent membrane-insertion energetics of hydrophobic segments engineered into the leader peptidase (Lep) protein, and derived an insertion scale for every amino acid across the membrane (9, 17). Elofsson and coworkers subsequently incorporated this scale into a hidden Markov model-based topology predictor (10, 18). Although this predictor is a substantial improvement over the statistics-based methods in reducing the number of fitted parameters, the authors noted significant uncertainties; for instance, the Lep scale reported only a small driving force of 0.5 kcal/mol for membrane insertion of the most hydrophobic residues, Leu and Phe, compared with at least 2 kcal/mol in other scales (7, 19), leading Elofsson and coworkers to observe that the Lep scale assigns positive (unfavorable) insertion energies to a large fraction of membrane-spanning segments (10). Furthermore, although it is established that protein orientation with respect to the membrane plane is determined by the positive-inside rule, the Lep measurements reported only a small bias of around 0.5 kcal/mol in favor of Arg and Lys in the cytoplasm compared with the extracellular domain (20), and thereby could not be used to predict orientation. To overcome these problems, the predictor relied on corrections, parameter fitting, and empirical rules in addition to the Lep energetics.

To derive higher-confidence insertion energetics we recently developed an experimental method, called deep sequencing TOXCAT-β-lactamase (dsTβL), and measured apparent insertion free energies ($\Delta G^{app}_{insertion}$) into the bacterial plasma membrane for

Significance

Topology prediction is crucial for structure prediction, design, and analysis of membrane proteins. We describe a graphical algorithm, called TopGraph, which is based on a sequence search for minimum energy of insertion using the dsTβL experimental insertion scale rather than statistics derived from known structures. Unlike many existing predictors, TopGraph exhibits high accuracy even on large transporters with no structural homologues. Furthermore, results suggest that the positive-inside rule, which is known to orient segments with respect to the membrane, can also drive insertion of marginally hydrophobic segments in large membrane domains.

Author contributions: A.E., J.J.W., and S.J.F. designed research; A.E., J.J.W., and S.J.F. performed research; A.E., J.J.W., and S.J.F. analyzed data; J.P. developed and tested the webserver; and A.E., J.J.W., and S.J.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]A.E. and J.J.W. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: sarel@weizmann.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1605888113/-/DCSupplemental.

each of the 20 amino acids at 27 positions across the membrane (21). The dsTβL scale was in good agreement with what theory and previous experiments suggested; for instance, the hydrophobicity at the membrane core was in line with biophysical measurements (22) and 3–4 times larger than measured with Lep (23). Furthermore, dsTβL reported large asymmetries (~2 kcal/mol) for the localization of the positively charged residues Arg and Lys in the cytoplasmic side relative to the extracellular side of the membrane. Here, we develop a topology predictor, called TopGraph, which uses dsTβL to assign insertion energies to segments of a query protein and search for the topology of minimum insertion energy. We test TopGraph on experimentally determined topology databases. We further analyze large membrane domains of transporters, which have been challenging for statistics-based predictors, and describe cases where the positive-inside rule apparently drives the insertion of polar segments into the membrane.

Results

Assessing Topology-Prediction Accuracy. We used three published datasets to test topology-prediction accuracy. First, prediction accuracy of membrane-span locations was assessed using the Reeb dataset, which is based on a nonredundant set of 188 high-resolution structures (with a pairwise sequence-identity cutoff of 20%; ref. 16). For each of the query sequences in the Reeb dataset we defined "overlap10" to represent whether the correct number of membrane-spanning segments was predicted, and whether at least 10 residues of each predicted segment overlapped with an inserted segment in the experimentally determined structure. Second, to assess orientation-prediction accuracy we used a set of 609 *Escherichia coli* inner membrane proteins, for which the location of the C terminus (cytoplasmic or periplasmic) was experimentally determined (24). Third, we assessed discrimination of soluble and membrane-spanning proteins using an annotated nonredundant set (<30% pairwise sequence identity) of 3,400 soluble proteins and 311 membrane proteins of known structure (13, 25). We compared the performance of TopGraph and TOPCONS, a topology predictor that uses the consensus of five statistical predictors (13, 26), in these three tests, and additionally analyzed the overlap10 performance of the Lep scale on the Reeb dataset.

A Graphical Algorithm for Membrane-Topology Prediction. We set ourselves the goal of predicting membrane-protein topology based on insertion energies and without invoking statistical inference to predict insertion propensities. Given a query sequence we start by using a sliding window to extract all subsequences of lengths 21–30 amino acid residues. The dsTβL scales (21) do not report on secondary-structure propensity nor on the existence of signal peptides, which are often cleaved post translationally. We therefore eliminate all signal peptides predicted by TOPCONS (13) and any subsequence that is predicted to be nonhelical (27), as well as subsequences that contain several charged or polar residues (*SI Methods*). To the remaining segments we assign apparent insertion free energies according to the dsTβL scale (21) in each of the two orientations relative to the membrane (locating the C-terminus either in the cytoplasm or outside). Because segments vary in length, we estimate the location z of every amino acid position i in the segment relative to the membrane midplane:

$$z(i) = \frac{30}{n} i - 15, \qquad [1]$$

where n is the total number of residues in the segment and i is the amino acid position relative to the segment's start; $z(i)$ ranges from −15 to +15 Å, for cytoplasmic and extracellular locations, respectively. The segment's apparent insertion free energy is then given by:

$$\Delta G^{app}_{insertion} = \sum_{i=1}^{n} \Delta G^{z(i)}_{AA}, \qquad [2]$$

where $\Delta G^{z(i)}_{AA}$ is the apparent insertion free energy for amino acid type AA at location $z(i)$ according to dsTβL.

Before running predictions, we modified the dsTβL profiles for the positively charged residue Lys and for the hydrophobic residues Val, Leu, Ile, and Met (Fig. S1 and Table S1). Specifically, in the original dsTβL report (21), Val, Leu, and Met showed slightly nonsymmetric profiles, whereas the other hydrophobics, Ile and Phe, were close to symmetric, as expected. We therefore changed the hydrophobic residues' profiles so that all were symmetric, and maintained the insertion energy at the membrane midplane as in the original dsTβL scales. Furthermore, the energy contribution of Lys at the cytoplasmic side of the membrane in the original dsTβL scale was slightly positive (+0.2 kcal/mol), thereby penalizing lysine-containing membrane-spanning segments. We therefore modified the Lys profile to be slightly negative (−0.1 kcal/mol) at the cytoplasm interface. These corrections increase the deviation between the polynomial functions used to fit the dsTβL data, but the most extreme deviation is only 0.84 kcal/mol (Leu at the membrane–cytoplasm interface; Fig. S1 and Table S1). In preliminary prediction runs we noticed that these changes do not affect prediction accuracy significantly.

We represent subsequences in the query and their apparent insertion energies as a graph, where nodes N stand for each subsequence (Fig. 1*A*). Nodes N_i and N_j are connected with a directed edge $N_i \rightarrow N_j$ if and only if N_i precedes and does not overlap with N_j in the query sequence and the two segments are inverted with respect to one another; that is, one segment's N terminus is cytoplasmic, and the other's C terminus is cytoplasmic. In addition, a virtual source node is connected to all other nodes, and every edge is weighted according to its successor node's $\Delta G^{app}_{insertion}$ (Eq. **2**); that is, the weight of edge $N_i \rightarrow N_j$ is the insertion

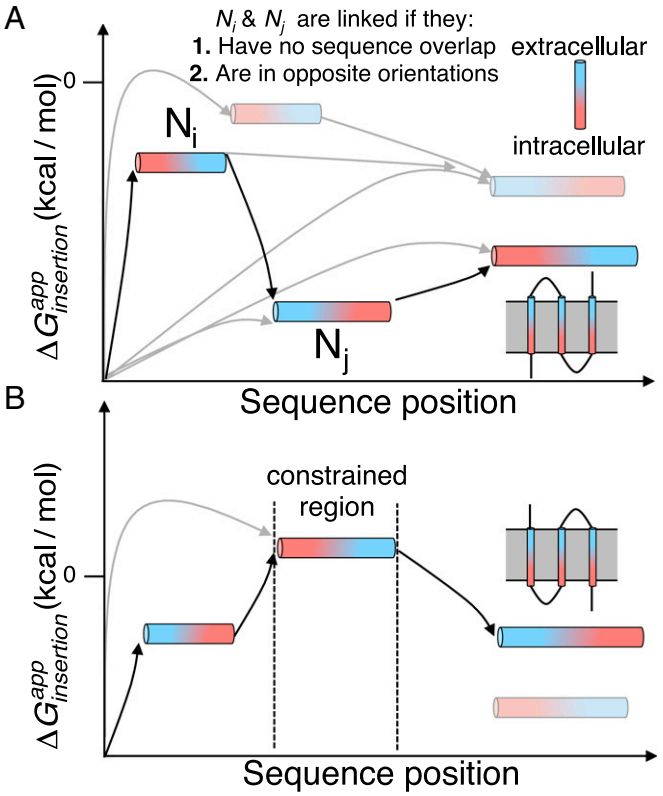

Fig. 1. Schematic representation of the graphical topology-prediction algorithm TopGraph. (*A*) Cylinders represent sequence segments (nodes in the graph) in either insertion orientation. Curved gray arrows represent edges connecting nodes; curved black arrows denote the minimum-energy path. Faded cylinders represent nodes not in the minimum-energy path. (*B*) A constraint (dashed lines) eliminates all edges bypassing it, guaranteeing, by construction, that only segments satisfying the constraint are chosen. *Insets* show final predicted topology.

energy of the segment represented by N_j plus the contributions from positive residues (Lys, Arg, and His) within a five amino acid stretch C terminal to N_i, and similarly, five amino acids N terminal to N_j (*SI Methods*). In this graphical representation, the minimum-energy path starting from the source predicts not only the location of membrane-spanning segments, as in previous predictors, but also the orientation of the protein with respect to the membrane and the length of each inserted segment. To search for the minimum-energy path we use the Bellman–Ford algorithm (28), which takes under 10 s to find minimal paths on a representative 265 amino acid protein using a standard CPU.

Constraints on the locations of membrane-spanning segments within the query can improve prediction accuracy. In the benchmark below we test the unconstrained prediction as well as two types of constraints: from multiple-sequence alignments (MSA) of homologous sequences, and from the TOPCONS predictor (13). To maintain the validity of the apparent insertion energies we do not use information other than from the query sequence itself to assign segment energies; rather, we use the information from MSAs or from TOPCONS only to determine where membrane spans are likely to be located, and compute the query's insertion energy by optimizing the inserted segments' precise locations and orientations within a stretch that includes five positions on either side of the segment determined using the MSA or TOPCONS. TopGraph^{MSA} conducts the search in two steps: it first predicts membrane-spanning locations using the MSA, and subsequently uses this information as location constraints in a search for minimum-energy paths in the query sequence (Fig. S2). In TopGraph^{TOPCONS}, by contrast, the locations of membrane-spanning segments are predicted using TOPCONS (13), and are then used to constrain the locations of membrane-spanning segments in a search for minimum-energy paths in the query (Fig. 1*B*). Alternative predictors could be used to constrain the locations of membrane-spanning segments with no loss in generality.

All three TopGraph variants predict the locations, lengths, orientations, and insertion energies of the query sequence. We note, additionally, that the graphical representation lends itself to imposing other types of constraints, which may be inferred from experimental or computational data; for instance, if a certain segment N_k is known to span the membrane, all edges $N_i \rightarrow N_j$ that bypass N_k may be eliminated (Fig. 1*B*). Conversely, nodes representing segments that are known not to cross the membrane may be eliminated, and prior data regarding the orientation of the protein in the membrane can be used to select the lowest-energy path through the graph in the known orientation. The ability to define a variety of topological constraints could aid the study of membrane proteins with incomplete structural data, such as on probe accessibility or proteolysis resistance (29), and we implemented a webserver providing free access to TopGraph including such manually constrained prediction (topgraph.weizmann.ac.il).

Prediction Accuracy Increases with Use of Prior Data. The purist TopGraph predictor, with no use of prior data predicts the locations of membrane segments in single-span proteins with high accuracy (94%; Fig. 2*A*). This high accuracy is not surprising given that the dsTβL scale is based on experimental data on a single-span membrane protein (21). Multispan membrane proteins are accordingly predicted less accurately, and above four segments prediction accuracy drops to 46%; the overall prediction accuracy across the entire set is 78%. When either of the two lowest-energy predicted paths is compared with the known topology, prediction accuracy increases from 70% to 80% for proteins with two to four membrane spans, and more modestly for larger membrane domains. The preprocessing filters that remove signal peptides, highly charged and nonhelical segments make a substantial contribution to prediction accuracy by eliminating, on average, two-thirds of the segments with $\Delta G^{app}_{insertion} < 5$ kcal/mol in each target sequence (Fig. S3). Nevertheless, prediction accuracy is high even in proteins, in which less than 20% of the

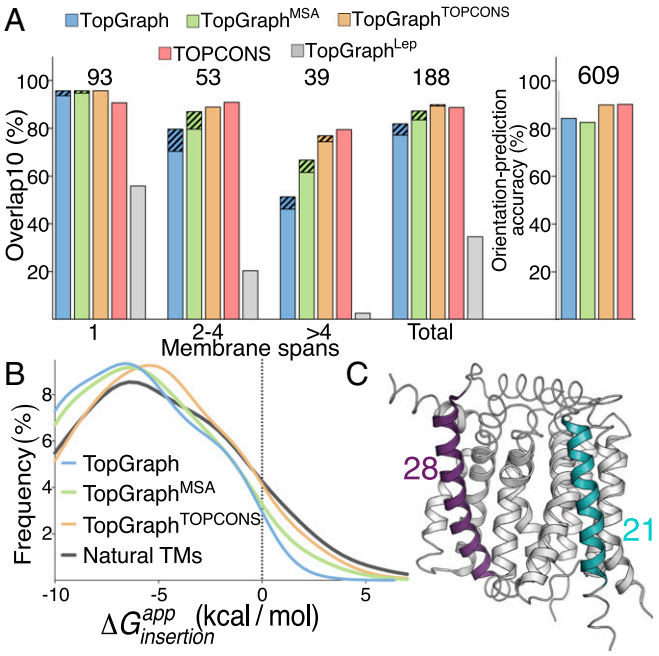

Fig. 2. Topology-prediction benchmark. (*A*, *Left*) Fraction of proteins where all predicted membrane spans overlap with experimentally observed membrane-spanning segments over at least 10 residues (overlap10) and there are no additional predicted segments. The number of proteins in each group is noted above the bars; dashed lines represent accuracy when considering either of the two best-energy predictions. (*Right*) Orientation-prediction accuracy of the C-terminal position (cytoplasmic or extracellular). (*B*) Distribution of insertion energies in individual membrane-spanning segments. Natural TMs reports the insertion-energy distribution of membrane-spanning segments annotated by the structure-based PDBTM (31). (*C*) Experimentally determined structure (PDB entry: 4K1C; ref. 42) annotated according to the TopGraph^{MSA} prediction: thin ribbon, extramembrane; thick ribbon, membrane spanning; two membrane-spanning segments with minimal and maximal predicted lengths are colored in turquoise and purple, respectively, and their lengths are noted.

sequence is eliminated by these filters; specifically, all of these proteins are predicted correctly according to the overlap10 metric.

For comparison, we replaced the dsTβL profiles with those from the Lep study (9) and tested prediction accuracy using the same algorithm as used for dsTβL (Fig. 2*A*). Bernsel et al. previously noted that the average energy assigned to single-spanning domains by the Lep scales is only slightly negative (approximately –0.3 kcal/mol) and that segments in multispanning domains are assigned positive energies on average (10), TopGraph^{Lep} prediction accuracy is correspondingly modest (56%) for single-spanning domains; it drops to 20% for proteins with two to four membrane spans, and there are nearly no correct predictions (3%) for larger membrane domains, with 34% overall prediction accuracy. These results are consistent with previous observations that the Lep insertion energetics are small for hydrophobic residues (9, 10, 17–19, 21, 30); because single-span membrane domains are typically more hydrophobic than multispan domains (10), the Lep scale predicts location more accurately in the former than in the latter.

We hypothesized that TopGraph^{MSA} may improve prediction accuracy relative to the purist TopGraph. The basis for this hypothesis is that homologous proteins are likely to have the same topology. Furthermore, although any given membrane protein must encode sufficiently favorable membrane-insertion free energy, individual segments in any protein may have lower insertion propensity than aligned segments in homologs. TopGraph^{MSA} retains TopGraph's high accuracy in single-pass membrane proteins (95%) and indeed improves on unconstrained TopGraph, reaching 61% accuracy for membrane proteins with more than

four spans and overall prediction accuracy of 84% (Fig. 2*A*). Furthermore, when considering either of the two lowest-energy paths, prediction accuracy improves to 87%, on par with TOPCONS (89%). TopGraph^{TOPCONS} shows nearly identical performance to TopGraph^{MSA} with overall prediction accuracy of 89%.

Energy-Based Prediction of Protein Orientation with Respect to the Membrane Plane. The dsTβL scale differs from other scales in showing large asymmetries for the localization of the positively charged residues, Arg, Lys, and His, in the cytoplasm compared with the extracellular space (21); this asymmetry is a prerequisite for energy-based prediction of membrane-protein orientation. Indeed, TopGraph correctly predicts orientation in 84% of the proteins in a benchmark of 609 bacterial proteins of experimentally determined orientation (24); overall accuracy is 82% and 90%, for TopGraph^{MSA} and TopGraph^{TOPCONS}, respectively, compared with 90% for TOPCONS (Fig. 2*A*).

The three TopGraph variants output apparent insertion free energies that are based on the dsTβL scale (21). Applied to the Reeb dataset (16), nearly all segments (99%) predicted using the purist TopGraph exhibit negative apparent insertion energies with a mean of −6.9 kcal/mol (Fig. 2*B*). Using the more accurate predictors TopGraph^{MSA} and TopGraph^{TOPCONS} the mean shifts to −6.4 and −5.7 kcal/mol, respectively, and 95% of segments exhibit negative insertion energies. We computed the per-segment insertion free energies of verified membrane-spanning segments, by constraining locations to those observed in membrane–protein structures (31), and derived a very similar distribution of insertion energies (Fig. 2*B*), and further found that 98% of the membrane spans had apparent insertion energies below +5 kcal/mol. Our analysis suggests that individual membrane spans, even in large domains in which intersegment interactions can drive insertion, must encode sufficiently high insertion propensity. These insertion energies are in agreement with theoretical treatments, which predict an average of approximately −5 kcal/mol for membrane insertion of a single segment (1, 32). The values stand, furthermore, in contrast to the analysis of membrane segments using the Lep insertion scale (9), which computes average insertion energy of +0.8 kcal/mol (10).

The relatively large magnitude of per-helix insertion energies predicted by TopGraph implies that it may discriminate soluble from membrane-spanning proteins. Indeed, in a set of 3,400 proteins (13), we find that a cutoff of $\Delta G^{app}_{insertion} = -3$ kcal/mol correctly discriminates membrane from soluble proteins with sensitivity of 96% and specificity of 93% (Table S2), comparable to other predictors (10, 33). We note that on average 99% of the sequence in soluble proteins is eliminated by the secondary-structure and polar-residues filters (Fig. S3), drastically simplifying prediction. We further find that individual membrane-spanning segments differ from segments in soluble proteins in that a large majority encode both hydrophobicity and orientation preference (the positive-inside rule; Fig. S4).

Most previous membrane-topology predictors search the sequence with a fixed-length window (typically ∼21 amino acid positions); TopGraph, by contrast, optimizes the lengths of the inserted segments. Fig. 2*C* and Fig. S5 show several TopGraph^{MSA} predictions for large membrane domains plotted on their molecular structures, demonstrating that TopGraph^{MSA} accurately locates membrane spans even in proteins with large extramembrane domains. Furthermore, the predictor correctly assigns long and short membrane-spanning segments within the same protein. Accurate length assignment could in the future aid ab initio structure prediction in membrane domains (34–36).

The Positive-Inside Rule Can Drive Insertion of Polar Segments in Large Membrane Domains. Many transporters and receptors have membrane-embedded polar and charged residues, suggesting that a purely energy-based predictor, such as TopGraph, might not assign membrane topology correctly in these cases. We nevertheless found that TopGraph^{MSA} correctly predicted the locations of experimentally validated membrane-spanning segments, even if they were assigned positive insertion energies. Indeed, out of 20 proteins of 6 or more membrane-spanning segments in the Reeb dataset (16), for which TopGraph^{MSA} produced correct predictions, 13 had at least one segment of marginal hydrophobicity ($\Delta G^{app}_{insertion} > -1.5$ kcal/mol), and of these, 8 had at least one polar segment ($\Delta G^{app}_{insertion} > 0$ kcal/mol). Furthermore, in nine cases at least 20% of the polar segment was exposed to the membrane environment; therefore, in many cases polar segments are not fully shielded from the surrounding hydrophobic lipid in the native structure (Table S3).

To investigate how TopGraph^{MSA} correctly predicts topology even in these challenging cases we compared its lowest-energy prediction to a simulated topology, in which the polar segment was computationally constrained to be excluded from the membrane and the lowest-energy topology was recalculated (Table S3). In 63% of the cases we found that the exclusion of a polar segment led to significant worsening in the total apparent insertion energy (increase of 2.6–8.5 kcal/mol relative to the unconstrained topology). We therefore looked for sequence features outside the polar segment that would explain this gap, and found that by excluding the polar segment from the membrane, the distribution of Lys and Arg residues across the entire protein became roughly balanced between cytoplasm and extracellular space; in the unconstrained TopGraph^{MSA} prediction, by contrast, the majority of Lys and Arg residues were near the cytoplasm, where they would be favored by the positive-inside rule (ref. 2; Fig. S6). Accordingly, most of the energy gap between the correct prediction from TopGraph^{MSA} and the simulated topology, where the marginally hydrophobic segment is excluded from the membrane, was due to contributions from Lys and Arg residues.

A representative example is provided by the homotrimeric 11-transmembrane (TM) archaeal ammonium transporter (PDB entry: 2B2F; ref. 37). In this protein, forcing the polar segment TM7 ($\Delta G^{app}_{insertion} = +1.9$ kcal/mol) out of the membrane increases the total apparent insertion energy by 8.5 kcal/mol (Fig. 3). Visual inspection shows that the correct topology positions 14 positive charges in the cytoplasm and 3 in the extracellular space, whereas the topology that excludes TM7 has a more balanced distribution of positive charges (9 and 8, respectively). Indeed, Lys and Arg residues make a large contribution (13.7 kcal/mol) to the difference in insertion energy between the correct and simulated topology. We conclude that the distribution of charges across the entire membrane domain

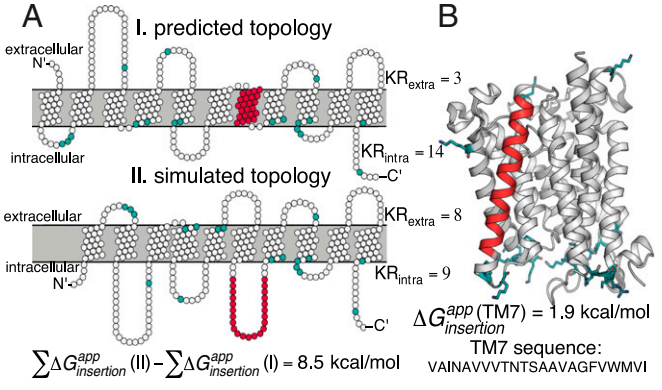

Fig. 3. Case study demonstrating that the positive-inside rule may favor membrane-insertion of polar segments. (*A*) The insertion of the marginally hydrophobic segment TM7 (red) positions a greater number of positive charges (turquoise) inside the cell compared with a hypothetical situation where the segment is forced out of the membrane (bottom). KR$_{extra}$ and KR$_{intra}$ denote the number of extra- and intracellular Lys and Arg residues. (*B*) Molecular structure (PDB entry: 2B2F; 37) annotated according to insertion prediction: TM7 in red; positive charges in turquoise.

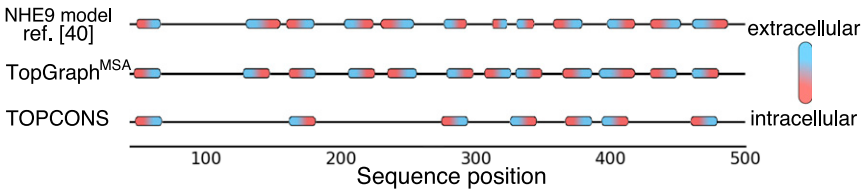

Fig. 4. NHE9 topology prediction. NHE9's topology was predicted using TopGraph^{MSA} and TOPCONS, and compared with an experimentally constrained model (40). Red/blue ellipses represent membrane-spanning segments within the sequences. TopGraph finds all membrane spans, and correctly assigns orientation, whereas TOPCONS misses five segments, and the C-terminal position. Only positions 43–500 are shown. See Fig. S8 for a similar analysis of NHE1.

may drive membrane insertion of weakly hydrophobic or polar segments.

Topology Prediction in a Transporter Family of Unknown Structure: Na^+/H^+ Exchanger as a Case Study. Current topology predictors are trained on known membrane protein structures. We find, for instance, that most sequences (88%) in the Reeb dataset (16) exhibit >40% sequence identity to sequences used for training at least one of the predictors used by TOPCONS (Fig. S7). To analyze a case, on which these predictors had no opportunity to train, we compared TOPCONS and TopGraph^{MSA} predictions for sequences belonging to the mammalian Na^+/H^+ exchanger (NHE) family. Several structures of functionally related proteins to NHE family members are available; the most homologous is the bacterial antiporter NhaA (PDB entry: 1ZCD; ref. 38), which is, however, of only ∼10% sequence identity to NHE family members (38, 39), and NHE family members are indeed of low homology relative to any of the sequences in the TOPCONS training sets. Due to the importance of the NHE family in pH regulation and the implication of NHE mutants in human disease (40), advanced structure–bioinformatics tools together with expert supervision were used to suggest the topology and 3D models for two members: NHE1 and NHE9 (39, 40) on the basis of the NhaA structure. These studies agreed on key topological features: all NHE family members place the C terminus in the cytoplasm and comprise 12 membrane-spanning segments in the region excluding the first 50 amino acid residues (thereby excluding the posttranslationally cleaved N-terminal signal peptide and possibly an additional membrane-spanning segment predicted in some NHE family members).

TopGraph^{MSA} predicts NHE9's topology correctly relative to the NHE9 model structure, finding all 12 membrane spans and placing the C terminus in the cytoplasm (ref. 40; Fig. 4). TOPCONS, by contrast, fails to recognize five NHE9 membrane spans, and incorrectly predicts that the C terminus lies outside the cytoplasm. NHE1 presents a more difficult case for TopGraph^{MSA} and although the C terminus is positioned correctly and 11 of the 12 membrane spans are accurately predicted, TM5, which is buried within the core of the NHE1 model structure (39), is missed. Although this segment is assigned marginally favorable insertion energy ($\Delta G^{app}_{insertion}$ = −0.8 kcal/mol), the secondary-structure prediction algorithm used by TopGraph mistakes TM5 for being nonhelical. TOPCONS also misses TM5 and additionally misses TM7 (Fig. S8). Our analysis is restricted to only two proteins, and we note that an alternative topology and a 3D model for NHE1 were put forward (41); we nevertheless find it encouraging that TopGraph can predict topology more accurately and largely in agreement with expert-guided modeling in these challenging cases of low sequence homology to known structures; improvements in secondary-structure prediction algorithms would further improve TopGraph accuracy. In specific cases, such as NHE1, that lack high-homology structures, but where a large body of experimental data are available, for instance on probe accessibility, topology prediction may be constrained with these data to improve accuracy.

Discussion

Despite four decades of research on membrane-protein energetics, experimental insertion scales lacked sufficient accuracy to predict membrane-protein topology directly from sequence. Instead, predictors have been dominated by statistical models fitted to experimental data. Although statistical predictors are accurate, the use of statistics raises two objections: first, given the low counts and high redundancy among membrane proteins of known structure, training and testing sets often cannot be satisfactorily segregated (16), and such studies might overestimate the expected prediction accuracy for proteins with no homology to known structures. Although this concern may be alleviated with future accumulation of experimental data, the second objection is more fundamental: statistics-based methods cannot be used to tease apart the different energy contributions to topology and have limited use in 3D structure prediction and design—modeling tasks that require accurate energetics. Our recent experimental measurement of apparent insertion energetics using the dsTβL assay quantified the positive-inside rule and agreed with hydrophobicity measurements (21); this higher accuracy allowed us to formulate a prediction algorithm without relying on statistics derived from known membrane-protein structures. The TopGraph analysis shows that prediction accuracy is on par with the consensus predictor TOPCONS. Furthermore, the NHE case study suggests that TopGraph^{MSA} has an advantage over statistical predictors in large membrane domains of low homology to known structures, where the statistical predictors have had no opportunity to train. Additionally, we noted several cases in which the lengths of the predicted segments agreed with experimental structures, a property which may aid 3D structure prediction.

TopGraph allowed us to quantitatively examine aspects of membrane-protein topology. We found that the majority of membrane spans in experimental structures were assigned negative apparent insertion energies and favorable orientation preferences (the positive-inside rule), suggesting that even in large membrane domains, spans must individually encode sufficiently favorable interactions with the membrane for insertion and orientation (Fig. S4). We additionally noticed that more than a third of large membrane domains have polar segments away from their termini that are nevertheless inserted to locate a greater number of positive charges in the cytoplasm. To be sure, a relationship between insertion and the positive-inside rule was recently noted by von Heijne, Elofsson, and coworkers, who showed that positive charges could drive the insertion of proximal segments (3, 4). Our results generalize this observation and suggest that the distribution of charges across the entire protein, rather than only in the proximity of polar segments, may drive the insertion of polar segments located away from the protein's termini. Whereas the orientation bias from a single positive charge (∼2 kcal/mol) (21) is smaller than the average net contribution from the insertion of a typical membrane-spanning segment (5–7 kcal/mol), the fact that a large membrane domain may have a dozen or more positive charges distributed across the entire protein provides a large and previously unnoted driving force for inserting polar segments. Although polar residues in membrane proteins are often linked to crucial functional features, such as oligomerization, substrate binding, and conformational change, high polarity undermines membrane insertion. We therefore speculate that the positive-inside rule has an important role in determining the architectures that underlie membrane-protein function. This insight may in the future help design altered or

new membrane-protein functional sites. TopGraph may be used to test hypotheses on the relative insertion propensities of natural and engineered proteins. Furthermore, our observations of high prediction accuracy recommend the dsTβL scale as the implicit-solvent term in structure prediction, design, and dynamics of membrane proteins.

Methods

Removing Signal Peptide, Highly Charged, and Nonhelical Segments. Signal peptides, nonhelical segments, and polar/charged subsequences were prefiltered as described in *SI Methods*.

N- and C-Terminal Sequence Contributions. The $\Delta G^{app}_{insertion}$ for every subsequence is supplemented by the contribution of Arg, Lys, and His in subsequent five residues, as described in *SI Methods*.

Multiple-Sequence Alignments. Multiple-sequence alignments are generated as described in *SI Methods*.

Topology Prediction Using MSA-Based Location Constraints. The use of MSA-based constraints is described in Fig. S2 and *SI Methods*.

Source Code. The source code is available at https://github.com/FleishmanLab/membrane_prediction. See *SI Methods* for more details.

TOPCONS and Lep-Based Predictions. The acquisition of data from the TOPCONS server and the Lep insertion scales is described in *SI Methods*.

Data Acquisition. The acquisition of the dataset is described in *SI Methods*.

ACKNOWLEDGMENTS. We thank Nir Ben-Tal and Arne Elofsson for critical reading and Meytal Landau and Gal Masrati for suggestions on NHE. The research was supported by the Minerva Foundation with funding from the Federal German Ministry for Education and Research. The S.J.F. laboratory is also supported by a European Research Council's Starter Grant, an individual grant from the Israel Science Foundation (ISF), the ISF's Center for Research Excellence in Structural Cell Biology, career development awards from the Human Frontier Science Program and the Marie Curie Reintegration Grant, an Alon Fellowship, and a charitable donation from Sam Switzer and family.

1. White SH, Wimley WC (1999) Membrane protein folding and stability: Physical principles. *Annu Rev Biophys Biomol Struct* 28:319–365.
2. von Heijne G (1989) Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* 341(6241):456–458.
3. Virkki MT, et al. (2014) The positive inside rule is stronger when followed by a transmembrane helix. *J Mol Biol* 426(16):2982–2991.
4. Öjemalm K, Halling KK, Nilsson I, von Heijne G (2012) Orientational preferences of neighboring helices can drive ER insertion of a marginally hydrophobic transmembrane helix. *Mol Cell* 45(4):529–540.
5. Kessel A, Ben-Tal N (2002) Free energy determinants of peptide association with lipid bilayers. *Pept Interact* 52:205–253.
6. Schramm CA, et al. (2012) Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions. *Structure* 20(5):924–935.
7. Moon CP, Fleming KG (2011) Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc Natl Acad Sci USA* 108(25):10174–10177.
8. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157(1):105–132.
9. Hessa T, et al. (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* 450(7172):1026–1030.
10. Bernsel A, et al. (2008) Prediction of membrane-protein topology from first principles. *Proc Natl Acad Sci USA* 105(20):7177–7181.
11. Käll L, Krogh A, Sonnhammer ELL (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 21(Suppl 1):i251–i257.
12. Reynolds SM, Käll L, Riffle ME, Bilmes JA, Noble WS (2008) Transmembrane topology and signal peptide prediction using dynamic Bayesian networks. *PLOS Comput Biol* 4(11):e1000213.
13. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res* 43(W1):W401–7.
14. Viklund H, Elofsson A (2008) OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 24(15):1662–1668.
15. Viklund H, Bernsel A, Skwark M, Elofsson A (2008) SPOCTOPUS: A combined predictor of signal peptides and membrane protein topology. *Bioinformatics* 24(24):2928–2929.
16. Reeb J, Kloppmann E, Bernhofer M, Rost B (2015) Evaluation of transmembrane helix predictions in 2014. *Proteins* 83(3):473–484.
17. Hessa T, et al. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433(7024):377–381.
18. Peters C, Tsirigos KD, Shu N, Elofsson A (2016) Improved topology prediction using the terminal hydrophobic helices rule. *Bioinformatics* 32(8):1158–1162.
19. Johansson ACV, Lindahl E (2009) Protein contents in biological membranes can explain abnormal solvation of charged and polar residues. *Proc Natl Acad Sci USA* 106(37):15684–15689.
20. Öjemalm K, Botelho SC, Stüdle C, von Heijne G (2013) Quantitative analysis of SecYEG-mediated insertion of transmembrane α-helices into the bacterial inner membrane. *J Mol Biol* 425(15):2813–2822.
21. Elazar A, et al. (2016) Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. *eLife* 5:e12125, 10.7554/eLife.12125.
22. Vajda S, Weng Z, DeLisi C (1995) Extracting hydrophobicity parameters from solute partition and protein mutation/unfolding experiments. *Protein Eng* 8(11):1081–1092.
23. Öjemalm K, et al. (2011) Apolar surface area determines the efficiency of translocon-mediated membrane-protein integration into the endoplasmic reticulum. *Proc Natl Acad Sci USA* 108(31):E359–E364.
24. Daley DO, et al. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. *Science* 308(5726):1321–1323.
25. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat Methods* 8(10):785–786.
26. Bernsel A, Viklund H, Hennerdal A, Elofsson A (2009) TOPCONS: Consensus prediction of membrane protein topology. *Nucleic Acids Res* 37(Web Server issue, SUPPL. 2):W465–W468.
27. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16(4):404–405.
28. Cormen TH, Leiserson CE, Rivest RL (1997) *Introduction to Algorithms* (MIT Press, Cambridge, MA), Vol 6.
29. Fleishman SJ, Unger VM, Ben-Tal N (2006) Transmembrane protein structures without X-rays. *Trends Biochem Sci* 31(2):106–113.
30. Shental-Bechor D, Fleishman SJ, Ben-Tal N (2006) Has the code for protein translocation been broken? *Trends Biochem Sci* 31(4):192–196.
31. Tusnady GE, Dosztanyi Z, Simon I (2005) PDB_TM: Selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res* 33(Database issue):D275–D278.
32. Ben-Tal N, Ben-Shaul A, Nicholls A, Honig B (1996) Free-energy determinants of alpha-helix insertion into lipid bilayers. *Biophys J* 70(4):1803–1812.
33. Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 23(5):538–544.
34. Barth P, Wallner B, Baker D (2009) Prediction of membrane protein structures with complex topologies using limited constraints. *Proc Natl Acad Sci USA* 106(5):1409–1414.
35. Barth P, Schonbrun J, Baker D (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc Natl Acad Sci USA* 104(40):15682–15687.
36. Yarov-Yarovoy V, Schonbrun J, Baker D (2006) Multipass membrane protein structure prediction using Rosetta. *Proteins* 62(4):1010–1025.
37. Andrade SLA, Dickmanns A, Ficner R, Einsle O (2005) Crystal structure of the archaeal ammonium transporter Amt-1 from Archaeoglobus fulgidus. *Proc Natl Acad Sci USA* 102(42):14994–14999.
38. Hunte C, et al. (2005) Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH. *Nature* 435(7046):1197–1202.
39. Landau M, Herz K, Padan E, Ben-Tal N (2007) Model structure of the Na+/H+ exchanger 1 (NHE1): Functional and clinical implications. *J Biol Chem* 282(52):37854–37863.
40. Kondapalli KC, et al. (2013) Functional evaluation of autism-associated mutations in NHE9. *Nat Commun* 4(May):2510.
41. Nygaard EB, et al. (2011) Structural modeling and electron paramagnetic resonance spectroscopy of the human Na+/H+ exchanger isoform 1, NHE1. *J Biol Chem* 286(1):634–648.
42. Waight AB, et al. (2013) Structural basis for alternating access of a eukaryotic calcium/proton exchanger. *Nature* 499(7456):107–110.
43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
44. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–1659.
45. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
46. Omasits U, Ahrens CH, Muller S, Wollscheid B (2014) Protter: Interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* 30(6):884–886.