Topological constraints and modular structure in the folding and functional motions of GlpG, an intramembrane protease

Nicholas P. Schafer[a], Ha H. Truong[b], Daniel E. Otzen[a], Kresten Lindorff-Larsen[c,1], and Peter G. Wolynes[b,1]

[a]Interdisciplinary Nanoscience Center, Department of Molecular Biology and Genetics, Aarhus University, DK-8000 Aarhus, Denmark; [b]Department of Chemistry, Center for Theoretical Biological Physics, Rice University, Houston, TX 77005; and [c]The Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark

Contributed by Peter G. Wolynes, January 6, 2016 (sent for review December 7, 2015; reviewed by James U. Bowie and Heedeok Hong)

We investigate the folding of GlpG, an intramembrane protease, using perfectly funneled structure-based models that implicitly account for the absence or presence of the membrane. These two models are used to describe, respectively, folding in detergent micelles and folding within a bilayer, which effectively constrains GlpG's topology in unfolded and partially folded states. Structural free-energy landscape analysis shows that although the presence of multiple folding pathways is an intrinsic property of GlpG's modular functional architecture, the large entropic cost of organizing helical bundles in the absence of the constraining bilayer leads to pathways that backtrack (i.e., local unfolding of previously folded substructures is required when moving from the unfolded to the folded state along the minimum free-energy pathway). This backtracking explains the experimental observation of thermodynamically destabilizing mutations that accelerate GlpG's folding in detergent micelles. In contrast, backtracking is absent from the model when folding is constrained within a bilayer, the environment in which GlpG has evolved to fold. We also characterize a near-native state with a highly mobile transmembrane helix 5 (TM5) that is significantly populated under folding conditions when GlpG is embedded in a bilayer. Unbinding of TM5 from the rest of the structure exposes GlpG's active site, consistent with studies of the catalytic mechanism of GlpG that suggest that TM5 serves as a substrate gate to the active site.

membrane proteins | micelle folding | bilayer folding | folding mechanism | intramembrane proteolysis

GlpG is a rhomboid protease that sits and functions in the cell membrane. GlpG's homologs are found across all kingdoms of life. GlpG has been the subject of several biophysical experimental studies aimed toward understanding membrane protein folding and the relationships among protein structure, dynamics, and function (1–5). An extensive experimental φ-value analysis found φ-values significantly different from zero, indicative of structural changes during the rate-limiting step of folding, in transmembrane helices 1 through 5 (TM1-5) and the intervening loops (4). Most of the nonzero φ-values, particularly in TM3-5 and in the large loop L1, were negative, meaning that although the corresponding mutation destabilizes the native state, the mutation nonetheless accelerates folding. The preponderance of negative φ-values was puzzling and unprecedented, and at the time, these effects were tentatively ascribed to nonnative interactions in the transition state ensemble. In this work, we show that, in fact, simple models with perfectly funneled energy landscapes that lack nonnative interactions are able to explain the origin of these negative φ-values and how the values arise when folding in detergent micelles rather than bilayers.

α-Helical membrane protein folding is thought to occur in two stages in vivo (6). The first stage, setting up the proper topology of transmembrane helices, is handled by the translocon (7, 8). In the present context, topology refers to specifying the directions in which a membrane protein's constituent transmembrane helices traverse the bilayer. The second stage, converting from properly inserted but dissociated helices into a functional folded structure, occurs spontaneously and is, in some ways, analogous to soluble protein folding. However, we know, ranging from the hydrophobic effect (9, 10) to water-mediated (11) and screened electrostatic interactions (12), the solvent plays a role in determining what types of noncovalent interactions are stabilizing and destabilizing. Whereas soluble proteins fold in polar and isotropic aqueous solutions, membrane proteins fold in largely apolar and anisotropic environments. These environmental differences complicate applying directly methods developed for studying soluble protein folding to the study of membrane protein folding. Nonetheless, experimentalists have been able to apply a variety of methods to study the kinetics and thermodynamics of membrane protein folding through the use of detergent micelles as a membrane-mimicking environment. Experiments that probe the folding mechanisms of membrane proteins have used micelles composed of a mixture of anionic and nonionic detergents (4, 13, 14), which not only keep membrane proteins soluble but also, through use of mixed micelles, allow the equilibrium between folded and unfolded states to be tuned. Micelles predominantly composed of nonionic detergents, such as *n*-dodecyl-β-D-maltopyranoside (DDM), preferentially stabilize a folded state that has been shown to be functional and is therefore likely to be structurally similar to the folded state in vivo. Micelles predominantly composed of anionic detergents, on the other hand, preferentially stabilize an unfolded state that contains significant amounts of secondary

Significance

Membrane proteins perform diverse functions in the cell while being embedded in lipid bilayers, but the presence of the anisotropic, nonpolar membrane environment has slowed progress in understanding how these proteins fold and function. Herein, we study GlpG, an intramembrane protease, using computationally efficient models to fill in structural details that are currently invisible to experimental techniques and inaccessible to atomistic simulations. We find that GlpG's modular functional architecture leaves an imprint throughout its folding and functional landscape, leading to multiple possible folding pathways and the population of near-native states with functional significance. We propose a mechanism by which destabilizing mutations can accelerate folding in detergent micelles, a previously puzzling experimental observation.

Author contributions: N.P.S., H.H.T., D.E.O., K.L.-L., and P.G.W. designed research; N.P.S. and H.H.T. performed research; N.P.S. and H.H.T. contributed new reagents/analytic tools; N.P.S., H.H.T., D.E.O., K.L.-L., and P.G.W. analyzed data; and N.P.S., H.H.T., D.E.O., K.L.-L., and P.G.W. wrote the paper.

Reviewers: J.U.B., University of California, Los Angeles; and H.H., Michigan State University.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: lindorff@bio.ku.dk or pwolynes@rice.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1524027113/-/DCSupplemental.

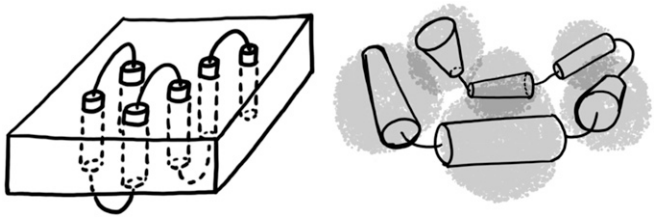

Fig. 1. Schematic diagrams of the unfolded state of α-helical membrane proteins in bilayers (*Left*) and detergent micelles (*Right*). The transmembrane helices (cylinders) are connected by loops. Transmembrane helices are either embedded in a membrane (rectangular prism) or are surrounded by detergent micelles (transparent gray spheres). In this work, we use an implicit membrane model to simulate folding within a bilayer and assume that folding in detergent micelles corresponds to folding without constraints on the alignment of helices. In both cases, we assume that the unfolded state has near-native levels of secondary structure, as has been observed in experiments on the SDS-denatured state of membrane proteins.

structure. This ability to tune the equilibrium means that stopped-flow kinetic experiments can be combined with protein-engineering techniques to determine folding mechanisms at the single-residue level (4, 13, 15), in analogy to what has been done for soluble proteins (16–18). Because carrying out these types of experiments in bilayers is still difficult, it is presently unknown how folding mechanisms determined in micelles compare with those in membranes. Confining proteins to a 2D membrane is expected to constrain unfolded and partially folded ensembles to having structures with helices that are largely properly aligned and embedded in the membrane; such topological restrictions would be relaxed in a micellar environment.

Theoretical (19, 20) and experimental (3, 4) work suggests that at least some membrane proteins can reversibly fold and unfold without the aid of the translocon or chaperones in vitro. It is therefore likely that membrane protein folding landscapes are funneled, much like globular protein landscapes (21, 22). Structure-based models with perfectly funneled energy landscapes have proven useful for investigating the folding and binding of proteins (23, 24). In this study, we use a structure-based model to investigate folding of a membrane protein in two different situations: in the absence and presence of an implicit membrane energy term that biases conformations to have the correct topology with respect to the membrane. Simulations with the implicit membrane term are thus taken to model folding in a bilayer, whereas simulations without the implicit membrane energy are taken to model folding in detergent micelles. Although this way of modeling micelles and bilayers is an oversimplification, it captures the significantly increased topological freedom of membrane proteins in micellar environments compared with lipid bilayer. Fig. 1 shows schematic representations of the corresponding denatured states of membrane proteins in bilayers and micelles.

The same energy landscape that dictates folding routes also encodes functional motions. It has been suggested that the modularity in the structure of GlpG supports functional motions (1, 25). The N-terminal domain, which contains transmembrane helices 1 and 2 (TM1-2) as well as the intervening L1 loop, functions as a structural scaffold (25), whereas the C-terminal domain with its four transmembrane helices (TM3-6) includes the catalytic site (25). The C-terminal domain is apparently more flexible than the N-terminal domain; both the loop L5 (5) and the transmembrane helix TM5 (25) have been crystallized in multiple conformations. Because of this flexibility, it has been suggested that either L5 alone (5) or L5 and TM5 (25) may serve as a substrate gate for access to the catalytic site. Using free-energy landscape analysis and perturbation methods along with structural analysis, we show that there is a near-native state significantly populated under folding conditions and elucidate the state's connections to GlpG's folding mechanism and function.

Methods

Simulation and Analysis Methodology. We performed molecular dynamics simulations of a coarse-grained structure-based model (26) of GlpG based on the crystal structure with Protein Data Bank (PDB) ID code 2XOV (27). We carried out two parallel sets of simulations: one with an implicit membrane present and one without a membrane. The implicit membrane model is described in ref. 19, and the assignment of residues into the intramembrane and extramembrane residues is described in Fig. S1. We sampled at multiple temperatures above and below the corresponding folding temperatures and used umbrella sampling at each of these temperatures to sample a wide range of folded, partially folded and unfolded structures. We then used the Multistate Bennett Acceptance Ratio (MBAR) method (28) to reconstruct unbiased free-energy profiles, compute expectation values of structural order parameters, and perform perturbative calculations to test the effect of small changes to the Hamiltonian. We infer folding mechanisms by looking for low free-energy routes between the unfolded and folded states in the unbiased free-energy profiles and then performing analysis on structures sampled in the basins and saddle points along these routes. Whereas the appropriateness of various reaction coordinates for describing protein folding kinetics is vigorously discussed (29–31), here we take the pragmatic approach of comparing our inferred mechanisms to experimental data and find highly nontrivial agreement based on reaction coordinates that measure the degree of nativeness of different parts of the molecule. See the *Supporting Information* for a complete explanation of the methods.

Structure-Based Model of GlpG. The crystal structure used to define the stabilizing native interactions in our structure-based model is shown in Fig. 2. GlpG has six transmembrane helices connected by five loops. The first loop, L1, is notable because it is large and contains several small interfacial helices. Our definition of the N- and C-terminal domains of GlpG was arrived at based on the analysis of our simulation results and is therefore not imposed on the model beforehand; these two domains are found to fold semi-independently (*Results and Discussion*) using our structure-based model. Therefore, this definition arises as a direct consequence of the structure of GlpG given our way of defining its contact map. Structural bioinformatics studies have indicated that membrane proteins are stabilized by tight helix–helix interactions that are mediated by small and polar residues (32). We therefore used a 6.5-Å $C\beta$–$C\beta$ cutoff to define stabilizing native interactions, which is somewhat shorter than the cutoffs that have been applied to simulations of soluble proteins in the past. We have also selectively strengthened the local-in-sequence interactions to decouple secondary and tertiary structure formation. This modification of the model is motivated by the observation of native-like levels of secondary structure in the SDS-unfolded state of GlpG (4). See the *Supporting Information* for a precise description of the parameters used in the model.

Results and Discussion

Unfolding Always Corresponds to Loss of Tertiary Structure with Retention of Secondary Structure but Leads to a More Expanded Ensemble in the Absence of the Implicit Membrane. Experimental circular dichroism and tryptophan fluorescence measurements

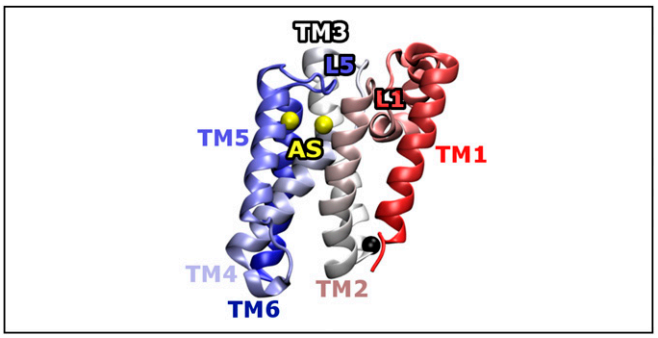

Fig. 2. Crystal structure of GlpG (PDB ID code 2XOV). A black sphere demarcates the boundary between the N- and C-terminal domains. The catalytic dyad in the active site (AS), shown in yellow and located on TM4 and TM6, is buried by TM5 and L5. The large loop L1 is made up of several interfacial helices whose axes run parallel to the membrane surface. The color of the backbone varies smoothly from red (N terminus) to white and then to blue (C terminus).

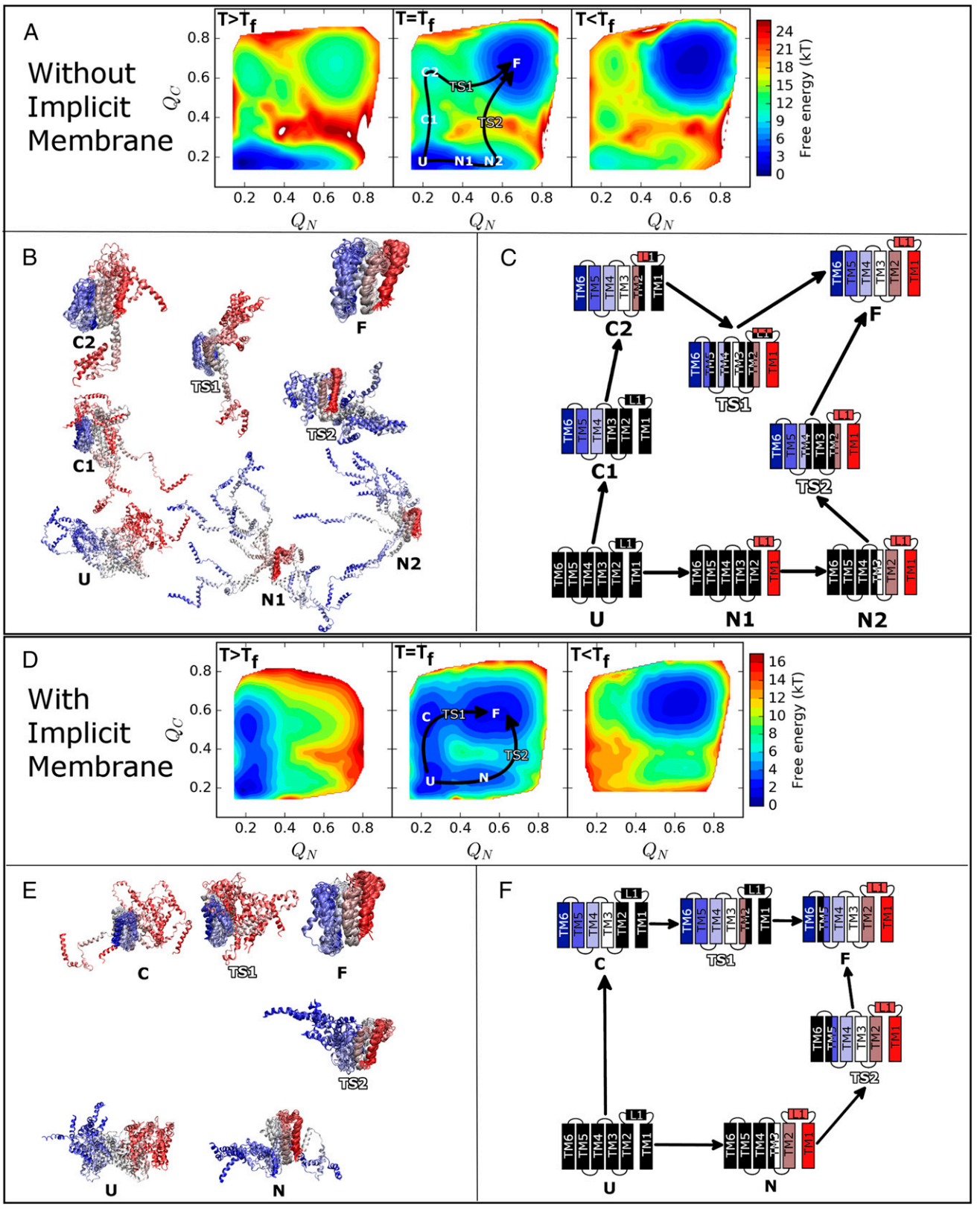

Fig. 3. Free-energy analysis and structural characterizations of GlpG without (*A*–*C*) and with (*D*–*F*) the implicit membrane. (*A* and *D*) Two-dimensional free-energy profiles above (*Left*), at (*Center*), and below (*Right*) the folding temperature (T_f) with respect to Q_N and Q_C. Q_N and Q_C measure the degree of folding within the N- and C-terminal domains, respectively. Precise definitions are given in the *Supporting Information*. Key structural states are labeled, and the inferred folding pathways are indicated with arrows. Areas shown in white are high in free energy. (*B* and *E*) Structural ensembles made up of 10 representative structures selected from low free-energy basins and transition states; folded regions in each ensemble have been aligned for clarity. (*C* and *F*) Schematic representations of the structural ensembles. Transmembrane helices and the large loop L1 are shown as fully folded (full color), partially folded (half color), or unfolded (black). The colors used in *B*, *C*, *E*, and *F* are the same as those established in Fig. 2.

indicate that unfolding of GlpG in micelles corresponds to loss of tertiary structure but retention of native levels of secondary structure (4). In the simulations, the expectation values of secondary and tertiary structure formation order parameters (see the *Supporting Information* for precise descriptions) as a function of temperature indicate that likewise, both in the absence and the presence of the implicit membrane, unfolding corresponds to loss of tertiary structure and retention of secondary structure (Fig. S2). When the implicit membrane is present, the unfolded structures largely retain native-like topologies with respect to the membrane (Fig. 3*E*), although excursions to the extramembrane regions are possible. The simulated unfolded ensemble thus resembles what is commonly understood to be the starting point for the "second stage" of membrane protein folding (6), which takes place once the helices have been inserted into the membrane by the translocon in their native orientations. The simulated unfolded ensemble in the absence of the bilayer is significantly more expanded (Fig. 3*B*). In the *Supporting Information*, we discuss a more detailed comparison of these two ensembles and experiment (Fig. S3).

Folding Can Be Initiated in Either the N- or C-Terminal Domain of GlpG. Both in the absence and presence of the implicit membrane, free-energy profiles plotted as a function of Q_N and Q_C (see the *Supporting Information* for precise descriptions), which quantify how native-like the structures are for the N- and C-terminal parts of the molecule, respectively, suggest that folding can be initiated by moving either along Q_N or Q_C [i.e., by forming native-like structure within either the N-terminal or C-terminal parts of the molecule (Fig. 3)]. Above but near the folding temperature and in the presence of the implicit membrane, the molecule populates both the fully unfolded state (U) and the C-terminal folded state (C) with TM3-6 folded. An orthogonal folding route toward the N-terminal folded state (N) is also present, although less favorable. In the absence of the implicit membrane and above the folding temperature, the molecule prefers the fully unfolded state (U) and a partially formed N-terminal structure with L1 folding onto TM1 (N1). Slightly higher in free energy in the same direction is another state with both TM1 and TM2, as well as the intervening L1, being well folded (N2). As in the case of the model with the implicit membrane, another folding route is available at higher free energy. There are also two intermediates along this route, the first with TM4-6 folded (C1) and a second which also includes folding TM2, TM3, and part of L1 onto the C-terminal part of the molecule (C2).

Optimal Energy–Entropy Compensation for the Modular Structure Results in a Multistep Folding Pathway That Backtracks During the Rate-Limiting Step Without the Implicit Membrane but Does Not Backtrack in the Membrane with Its Accompanying Topological Constraints. After initiating folding through either the N- or C-terminal domains, GlpG must fold the other half of the molecule to arrive at the folded state. In the membrane, this completion of folding occurs in a straightforward manner, with both pathways (U→C→TS1→F and U→N→TS2→F) being approximately equal in free energy (Fig. 3*D*). Without the implicit membrane energy term to constrain the topology, however, folding becomes more complex. Although initiating folding via the N-terminal domain (U→N1→N2) is more favorable than initiating folding via the C-terminal (U→C1→C2), starting along this route is ultimately not productive as the molecule later encounters a relatively high free-energy barrier (TS2) associated with organizing the large and unconstrained C-terminal domain. Folding does not proceed by propagating the folding "front" through the interface between the N- and C-terminal domains because there are relatively few contacts on the interface. Instead, the high free-energy barrier to folding is lowered somewhat through simultaneous organization of TM4-6 (a decrease in energy) at the same time as breaking the interface between L1/TM2 and TM3 (an increase in entropy), which was formed in N2. Breaking the interface between L1/TM2 and TM3 is an example of "backtracking" (i.e., the required unfolding of natively folded substructures while proceeding from the unfolded state to the folded state). By making optimal use of energy-entropy compensation, GlpG is able to reduce the free-energy barrier between a partially folded state and the completely folded state because there are multiple sites for nucleating folding. Once both domains are independently folded in TS2, a saddle point in the free-energy surface is reached and folding can proceed downhill to the folded state (F). This effect is also operative when folding is initiated in the C-terminal direction (U→C1→C2). Proceeding initially uphill in free energy, GlpG arrives at C2 where TM2-6 and parts of the loop L1 are folded. Because L1 is quite large, however, there exists a high entropic barrier to consolidate folding of TM1. Again, a compromise is made by simultaneously forming the interface TM1-TM2 and contacts within L1 along with releasing of L1 from its position docked against L3 and breaking the interface between TM2 and the C-terminal domain (TM3/L3/TM4). Finally, folding can proceed downhill toward the folded state by reforming the interface between TM2 and the C-terminal domain and reinserting L1. Note that the presence of high-energy intermediates and multiple folding pathways are compatible with the apparent two-state behavior observed in the micelle-mediated folding experiments. Folding is cooperative in experiments and in our simulations, but free-energy landscape analysis allows us to resolve high free-energy intermediate states and multiple pathways that would not necessarily be apparent from the initial experimental data alone. With the simulation-derived structural model for the parallel

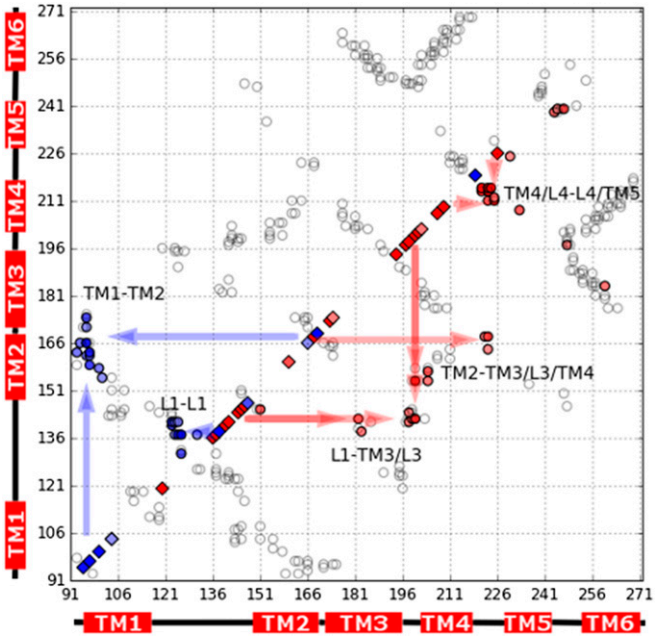

Fig. 4. Contact map of GlpG showing the C2→TS1 structural transition. The axes are labeled with residue indices. Contacts that change their occupancy by more than 20% when going from C2 to TS1 are shown in blue filled circles (gained in TS1; upper diagonal) and red filled circles (lost in TS1; lower diagonal). All other native contacts satisfying $|i - j| > 4$ are shown as empty circles. Positive (blue) and negative (red) experimental φ-values satisfying $|\varphi| > 0.2$ are plotted along the diagonal as filled diamonds. Arrows illustrate the proposed connections between the experimental φ-values and the contacts that are either lost or gained in the simulated structural ensembles. Text labels indicate the interfaces that are either formed or broken during the transition. Note that the positive φ-value at position 219 (the only significantly positive φ-value in the C-terminal domain) is derived from a mutation that actually accelerates folding and unfolding, like those that lead to the negative φ-values, but is formally positive because the mutation slightly stabilizes (rather than destabilizes) the native state.

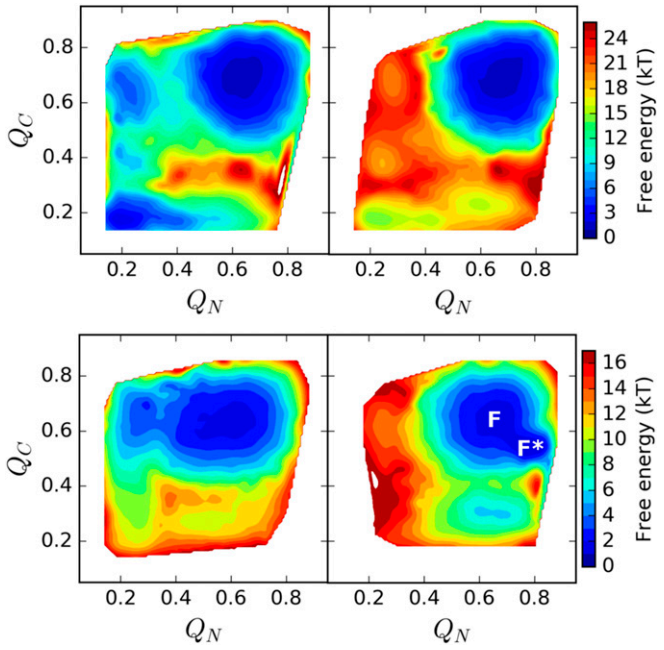

Fig. 5. Two-dimensional free-energy profiles of GlpG without (*Upper*) and with (*Lower*) the implicit membrane below the folding temperature, and with an N-terminal domain destabilized (*Left*) and stabilized (*Right*) by 10%. A near-native state (F*) is highly populated and accessible from the folded state (F) when the N-terminal is stabilized and the implicit membrane is present.

pathways, it should be possible to design experiments that probe this aspect of GlpG folding.

Of the two putative folding pathways, the latter one, initiated through folding the C-terminal domain, has the lower free-energy transition state (TS1) and should be dominant. TS1 differs somewhat from the transition state ensemble with an unfolded C-terminal domain and N-terminal folding nucleus, which was inferred without the aid of modeling and based on the distribution of experimentally measured canonical ($0 < \varphi < 1$) φ-values in GlpG (4). However, in that study, thermodynamically destabilizing mutations that accelerated folding and unfolding were found throughout TM3-5 and in L1. The resulting φ-values are negative. Destabilizing mutations that slow folding, leading to positive φ-values, were found largely on the interface TM1-TM2 but also in L1. Fig. 4 shows the difference between the average contact maps of TS1 and C2 as well as the connections between the contacts that are gained and lost in going from C2 to TS1 and the experimentally measured φ-values. Mutations that destabilize the interface between L1/TM2 and the C-terminal domain accelerate folding because formation of TS1 involves breaking those contacts. Mutations that destabilize the interface TM1-TM2 will slow folding because formation of TS1 also involves forming that interface. Mutations that primarily affect contacts within the C-terminal domain result in near-zero φ-values, because those contacts are largely preserved in the C2→TS1 transition. Thus, we see that the dominant mechanism predicted by simulations in the absence of the membrane (U→C1→C2→TS1→F) provides a detailed structural explanation of the previously puzzling preponderance of negative φ-values measured in the C-terminal domain of GlpG. On a topologically unconstrained but perfectly funneled landscape, folding is complicated by GlpG's modular structure and the high entropic cost of organizing helical bundles from their unconstrained partially folded states. Nonnative frustrated interactions need not be invoked to explain the presence of a large number of negative φ-values in GlpG.

A recent single-molecule force spectroscopy study in bicelles and micelles also found evidence for structural modularity in GlpG unfolding (3). The authors found that the unfolding of GlpG at high force was cooperative. The authors were also able to characterize two transiently populated metastable states. The authors' structural interpretation of the unfolding via intermediates closely corresponds to the reverse of one of our folding pathways (F→TS2→N→U) in the presence of the bilayer, whereas the structural decomposition of GlpG into domains given in the supplementary information of the study's article corresponds more or less exactly to the reverse of one of our dominant folding pathways (F→C2→C1→U) in the absence of the bilayer. These encouraging correspondences (see the *Supporting Information* for a more detailed discussion) suggest that further computational and experimental work should allow us to create a unified picture of SDS- and force-induced unfolding of GlpG in micelles and bilayers.

TM5 Is Loosely Bound Even Under Folding Conditions. GlpG is an intramembrane protease of the rhomboid serine protease class (33). GlpG cleaves specific transmembrane substrates using a catalytic dyad that is buried within the lipid bilayer (25). Fig. S4 shows two crystal structures of GlpG, one in a "closed" conformation, the one used to construct our structure-based energy landscape, and the other in an "open" conformation, where L5 and TM5 have bent away from the rest of the structure to expose partially the catalytic dyad. It has been suggested that TM5 functions as a substrate gate that opens for full-sized substrates to gain access to the catalytic site (25).

Preferential stabilization of the contacts within the N-terminal domain by 10% suffices to populate a near-native state (F*) under folding conditions in the presence of the implicit membrane (Fig. 5), according to our perturbation calculations. Structural analysis of this state revealed a heterogeneous ensemble of near-native conformations with a common feature: TM5 was unbound from TM4 and TM6, thereby exposing the catalytic dyad. In this state, deviations from the closed crystal structure occur most significantly in TM5 and the connecting loops L4 and L5 (Fig. 6). Whether or not TM5 must undergo significant conformational rearrangements in order for full-sized substrates to access the proteolytic site is a matter of some controversy (5, 25, 34). Our model suggests that the conformation of TM5 is highly dynamic even under folding conditions, which is consistent with the experimental observation that tethering TM5 to TM2 eliminates enzymatic

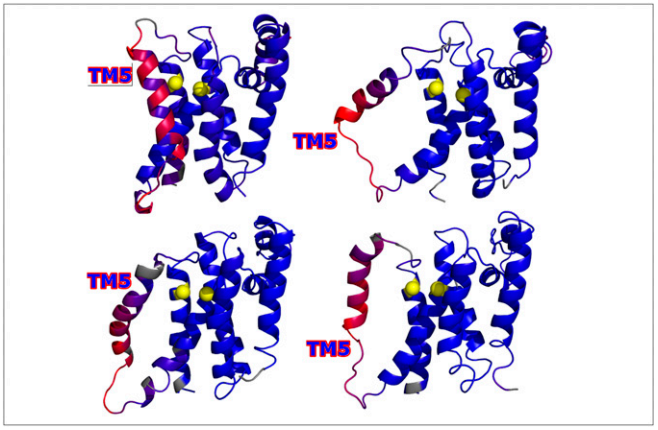

Fig. 6. Representative structures from a near-native state (F*) (Fig. 5) sampled while simulating with the implicit membrane present. The structures were all aligned to the closed crystal structure (PDB ID code 2XOV) and colored according to the individual residue rmsd values. Blue indicates low rmsd (high similarity to the crystal structure), and red indicates high rmsd. The catalytic dyad is shown using yellow spheres. High rmsd values are localized to the C-terminal half of the molecule and to TM5 in particular. Movement of TM5 exposes the catalytic dyad, thereby allowing substrate access. This state is highly populated under folding conditions when strengthening the contacts in the N-terminal half of the molecule by 10% relative to the contacts in the C-terminal half of the molecule.

activity (1). The fact that stabilizing the N-terminal part of the molecule increases the population of this state agrees with the experimental observation that destabilizing L1 reduces enzymatic activity (1, 25), highlighting the role of the N-terminal part of the molecule as a structural scaffold. Whereas TM5 is mobile in F*, F* differs, crucially, from TS2 in the implicit membrane (Fig. 3 *D–F*) by TM6 remaining bound to TM4. The tight association between TM4 and TM6 is mediated by GXXXAXXG and GXXXGXXXA motifs, which stabilize the C-terminal domain and protect against unfolding during GlpG's functional motions.

Conclusions

Experiments that probe membrane protein folding on the single-residue (4, 35) and the single-molecule (3) levels begin to allow us to determine the mechanisms by which membrane proteins fold and function. Nevertheless, many details of these processes remain hidden to even the most sensitive experiments. Using mixed micelles provides powerful tools for investigating membrane protein biophysics because of the relative simplicity and general applicability of these micelles, but the structure of the denatured state and its effect on folding mechanisms needs to be better understood. Thus far, studies of how residual structure in the denatured state affects folding have focused on soluble proteins and have used atomistic simulations (36), NMR and other types of spectroscopy (37), or combinations of the two (38). The question of residual structure is certainly no less important for membrane proteins, but the membrane environment poses challenges to both NMR and atomistic simulations. In this work, we used a coarse-grained energy landscape model to explore two limiting models of the folding of an intramembrane protease, GlpG: one limit in which the helices largely remain embedded in the membrane with their proper orientations, as is expected for the denatured state in lipid bilayers, and another limit where no constraints are placed on the alignment of helices in the unfolded state, this being taken as a model for the SDS-denatured state in micelles. Despite the simplicity of these models, on their basis, we have been able to propose a solution to the major puzzle in the experimental study of GlpG's folding mechanism, characterize a near-native state with potential functional significance, and show how these phenomena are related to GlpG's modular structure and topological constraints on the motions of partially folded states. The modular architecture of GlpG supports functional motions, including a highly mobile TM5, and leads to backtracking during the rate-limiting step of folding when the entropic cost of organizing helical bundles is high, as is the case in the absence of a bilayer. By providing a structurally detailed resolution of the φ-value puzzle, our analysis gives strong support to the notion that GlpG folding in mixed micelles proceeds by assembling helices with native levels of secondary structure from a state with few other constraints, as guided by a funneled, minimally frustrated landscape.

ACKNOWLEDGMENTS. We thank Sin Urban for ongoing constructive discussions about GlpG. K.L.L. and N.P.S. acknowledge support from the Novo Nordisk Foundation. K.L.-L., N.P.S., and D.E.O. were supported by Danish Research Council Grant DFF-4090-00220 and Carlsberg Foundation Grant CF14-0287. H.H.T. and P.G.W. were supported by National Institute of General Medical Sciences Grant R01 GM44557 and the D. R. Bullard-Welch Chair at Rice University (Grant C-0016). Computational resources were supported in part by the Data Analysis and Visualization Cyberinfrastructure funded by the National Science Foundation Grant OCI-0959097.

1. Baker RP, Young K, Feng L, Shi Y, Urban S (2007) Enzymatic analysis of a rhomboid intramembrane protease implicates transmembrane helix 5 as the lateral substrate gate. *Proc Natl Acad Sci USA* 104(20):8257–8262.
2. Baker RP, Urban S (2012) Architectural and thermodynamic principles underlying intramembrane protease function. *Nat Chem Biol* 8(9):759–768.
3. Min D, Jefferson RE, Bowie JU, Yoon TY (2015) Mapping the energy landscape for second-stage folding of a single membrane protein. *Nat Chem Biol* 11(12):981–987.
4. Paslawski W, et al. (2015) Cooperative folding of a polytopic α-helical membrane protein involves a compact N-terminal nucleus and nonnative loops. *Proc Natl Acad Sci USA* 112(26):7978–7983.
5. Zoll S, et al. (2014) Substrate binding and specificity of rhomboid intramembrane protease revealed by substrate-peptide complex structures. *EMBO J* 33(20):2408–2421.
6. Popot JL, Engelman DM (1990) Membrane protein folding and oligomerization: The two-stage model. *Biochemistry* 29(17):4031–4037.
7. Pohlschröder M, Prinz WA, Hartmann E, Beckwith J (1997) Protein translocation in the three domains of life: Variations on a theme. *Cell* 91(5):563–566.
8. Zhang B, Miller TF, 3rd (2012) Long-timescale dynamics and regulation of Sec-facilitated protein translocation. *Cell Reports* 2(4):927–937.
9. Dill KA (1990) Dominant forces in protein folding. *Biochemistry* 29(31):7133–7155.
10. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. *Adv Protein Chem* 14:1–63.
11. Papoian GA, Ulander J, Wolynes PG (2003) Role of water mediated interactions in protein-protein recognition landscapes. *J Am Chem Soc* 125(30):9170–9178.
12. Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. *Science* 268(5214):1144–1149.
13. Schlebach JP, Woodall NB, Bowie JU, Park C (2014) Bacteriorhodopsin folds through a poorly organized transition state. *J Am Chem Soc* 136(47):16574–16581.
14. Curnow P, et al. (2011) Stable folding core in the folding transition state of an alpha-helical integral membrane protein. *Proc Natl Acad Sci USA* 108(34):14133–14138.
15. Otzen DE (2011) Mapping the folding pathway of the transmembrane protein DsbB by protein engineering. *Protein Eng Des Sel* 24(1-2):139–149.
16. Oliveberg M, Wolynes PG (2005) The experimental survey of protein-folding energy landscapes. *Q Rev Biophys* 38(3):245–288.
17. Fersht AR, Sato S (2004) Phi-value analysis and the nature of protein-folding transition states. *Proc Natl Acad Sci USA* 101(21):7976–7981.
18. Itzhaki LS, Otzen DE, Fersht AR (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. *J Mol Biol* 254(2):260–288.
19. Kim BL, Schafer NP, Wolynes PG (2014) Predictive energy landscapes for folding α-helical transmembrane proteins. *Proc Natl Acad Sci USA* 111(30):11031–11036.
20. Truong HH, Kim BL, Schafer NP, Wolynes PG (2015) Predictive energy landscapes for folding membrane protein assemblies. *J Chem Phys* 143(24):243101.
21. Onuchic JN, Wolynes PG (2004) Theory of protein folding. *Curr Opin Struct Biol* 14(1):70–75.
22. Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA* 84(21):7524–7528.
23. Go N (1983) Theoretical studies of protein folding. *Annu Rev Biophys Bioeng* 12:183–210.
24. Levy Y, Wolynes PG, Onuchic JN (2004) Protein topology determines binding mechanism. *Proc Natl Acad Sci USA* 101(2):511–516.
25. Wu Z, et al. (2006) Structural analysis of a rhomboid family intramembrane protease reveals a gating mechanism for substrate entry. *Nat Struct Mol Biol* 13(12):1084–1091.
26. Eastwood MP, Wolynes PG (2001) Role of explicitly cooperative interactions in protein folding funnels: A simulation study. *J Chem Phys* 114(10):4702–4716.
27. Vinothkumar KR, et al. (2010) The structural basis for catalysis and substrate specificity of a rhomboid protease. *EMBO J* 29(22):3797–3809.
28. Shirts MR, Chodera JD (2008) Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys* 129(12):124105.
29. Cho SS, Levy Y, Wolynes PG (2006) P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc Natl Acad Sci USA* 103(3):586–591.
30. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES (1998) On the t ransition coordinate for protein folding. *J Chem Phys* 108(1):334–350.
31. Zheng W, Best RB (2015) Reduction of All-Atom Protein Folding Dynamics to One-Dimensional Diffusion. *J Phys Chem B* 119(49):15247–15255.
32. Eilers M, Patel AB, Liu W, Smith SO (2002) Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys J* 82(5):2720–2736.
33. Vinothkumar KR, Freeman M (2013) Intramembrane proteolysis by rhomboids: Catalytic mechanisms and regulatory principles. *Curr Opin Struct Biol* 23(6):851–858.
34. Wang Y, Zhang Y, Ha Y (2006) Crystal structure of a rhomboid family intramembrane protease. *Nature* 444(7116):179–180.
35. Hong H, Blois TM, Cao Z, Bowie JU (2010) Method to measure strong protein-protein interactions in lipid bilayers using a steric trap. *Proc Natl Acad Sci USA* 107(46):19802–19807.
36. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. *Science* 334(6055):517–520.
37. Mayor U, Grossmann JG, Foster NW, Freund SMV, Fersht AR (2003) The denatured state of Engrailed Homeodomain under denaturing and native conditions. *J Mol Biol* 333(5): 977–991.
38. Lindorff-Larsen K, et al. (2004) Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *J Am Chem Soc* 126(10):3291–3299.
39. Davtyan A, et al. (2012) AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J Phys Chem B* 116(29):8494–8503.
40. Plimpton S (1995) Fast Parallel Algorithms for Short-Range Molecular-Dynamics. *J Comput Phys* 117(1):1–19.
41. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. *J Mol Graph* 14(1):33–38, 27–28.
42. DeLano WL, Lam JW (2005) PyMOL: A communications tool for computational models. *Abstr Pap Am Chem S* 230:U1371–U1372.
43. Tusnády GE, Dosztányi Z, Simon I (2005) TMDET: Web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics* 21(7):1276–1277.
44. Krishnamani V, Hegde BG, Langen R, Lanyi JK (2012) Secondary and tertiary structure of bacteriorhodopsin in the SDS denatured state. *Biochemistry* 51(6):1051–1060.