ACS OMEGA

Article

http://pubs.acs.org/journal/acsodf

Structure and Dynamics of Type III Secretion Effector Protein ExoU As determined by SDSL-EPR Spectroscopy in Conjunction with De Novo Protein Folding

Axel W. Fischer,[†,‡] David M. Anderson,[⊥] Maxx H. Tessmer,[∥] Dara W. Frank,[∥] Jimmy B. Feix,[*,§] and Jens Meiler[*,†,‡]

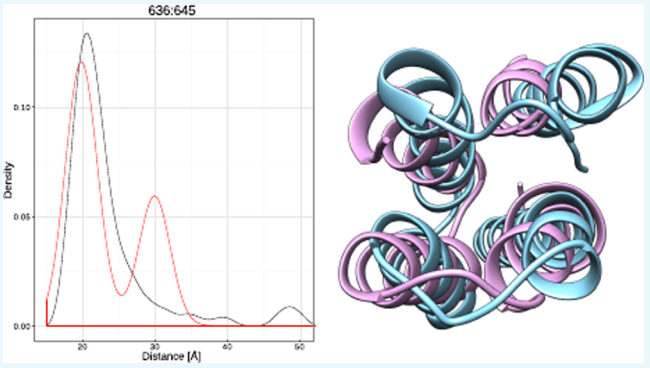
[†]Department of Chemistry and [‡]Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37232, United States

[§]Department of Biophysics and [∥]Department of Microbiology and Immunology, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, United States

[⊥]Department of Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, United States

Supporting Information

ABSTRACT: ExoU is a 74 kDa cytotoxin that undergoes substantial conformational changes as part of its function, that is, it has multiple thermodynamically stable conformations that interchange depending on its environment. Such flexible proteins pose unique challenges to structural biology: (1) not only is it often difficult to determine structures by X-ray crystallography for all biologically relevant conformations because of the flat energy landscape (2) but also experimental conditions can easily perturb the biologically relevant conformation. The first challenge can be overcome by applying orthogonal structural biology techniques that are capable of observing alternative, biologically relevant conformations. The second challenge can be addressed by determining the structure in the same biological state with two independent techniques under different experimental conditions. If both techniques converge to the same structural model, the confidence that an unperturbed biologically relevant conformation is observed increases. To this end, we determine the structure of the C-terminal domain of the effector protein, ExoU, from data obtained by electron paramagnetic resonance spectroscopy in conjunction with site-directed spin labeling and in silico de novo structure determination. Our protocol encompasses a multimodule approach, consisting of low-resolution topology sampling, clustering, and high-resolution refinement. The resulting model was compared with an ExoU model in complex with its chaperone SpcU obtained previously by X-ray crystallography. The two models converged to a minimal RMSD100 of 3.2 Å, providing evidence that the unbound structure of ExoU matches the fold observed in complex with SpcU.



INTRODUCTION

ExoU is a 74 kDa cytotoxin encoded by Gram-negative bacterium *Pseudomonas aeruginosa*.[1−4] Using the type III secretion system, ExoU is injected directly into eukaryotic cells, significantly increasing the severity of the infection.[5−7] Because of its function, ExoU needs to undergo substantial conformational changes; that is, depending on interaction partners and environment, different conformations of the protein will be thermodynamically most stable. One conformation of ExoU, in complex with its chaperone SpcU, has previously been elucidated through X-ray crystallography (Protein Data Bank (PDB) entry 3TU3).[2] This X-ray-derived model depicts ExoU as consisting of four domains. The C-terminal domain is of particular interest because it mediates the association of ExoU with the membrane,[2,8,9] that is, it is expected to undergo major conformational changes. However, all three structural models obtained through X-ray crystallography (PDB entries 3TU3,[2] 4AKX,[1] and 4QMK[8]) depict ExoU's C-terminal domain as exhibiting the same conformation, a four-helical bundle. Experiments performed by Gendrin et al. showed that even the presence of chaperone SpcU does not occlude the residues involved in lipid binding.[1] Through electron paramagnetic resonance (EPR) spectroscopy, Benson et al. provided evidence that the presence of the substrate induces conformational changes in ExoU's C-terminal domain.[10] Given the expected intrinsic flexibility of this domain, we set out to (a) confirm that the conformation of the C-terminal domain observed in the X-ray crystallography-derived model in complex with its chaperone SpcU is consistent with structural data observed for ExoU in solution,

Received: March 23, 2017
Accepted: June 15, 2017
Published: June 27, 2017

DOI: 10.1021/acsomega.7b00349
ACS Omega 2017, 2, 2977−2984

and (b) probe the structural dynamics of this domain. We chose EPR spectroscopy in conjunction with site-directed spin labeling (SDSL) in combination with computational de novo protein folding to approach these questions.

EPR spectroscopy in conjunction with SDSL provides an alternative approach to probe the structure and dynamics of a protein. Briefly, SDSL-EPR is typically employed to measure the distance between two residues. To facilitate that, two cysteine residues are introduced at the sites of interest into a cys-less variant of the protein and coupled with *S*-(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1*H*-pyrrol-3-yl)methyl methanesulfonothioate (MTSL), which carries an unpaired electron. Through the double electron−electron resonance (DEER) experiment,[11,12] the distance-dependent dipolar interaction of the two unpaired electrons can be measured and translated into a distance distribution. Because every measurement requires a distinct protein double mutant, the structural information gained from SDSL-EPR experiments is typically too sparse to unambiguously determine the protein's tertiary structure. However, in conjunction with de novo protein structure prediction methods, SDSL-EPR data could focus the sampling on conformations that are in agreement with the experimental data.

The computational protein structure prediction pipeline employed in this article is based on the de novo method BCL::Fold,[13] which was specifically developed to predict the tertiary structure of large proteins. To facilitate this objective, the secondary structure elements (SSEs) of the protein are predicted using machine learning methods. Conformations of the predicted SSEs exhibiting idealized dihedral angles are subsequently arranged in the three-dimensional space by a Monte Carlo Metropolis (MCM) algorithm. The intermediary and final models are evaluated using knowledge-based potentials that assign a pseudoenergy score to each model.[14] Although this method has been successful at de novo sampling the tertiary structure of large proteins, distinguishing between accurate and inaccurate models based on their pseudoenergy score alone remains a challenge.[15] However, it was demonstrated that incorporation of limited experimental data significantly mitigates problems in model discrimination.[16−19] The Rosetta method[20,21] was used to add atomic detail and energy-optimize the final models.

In this article, we discuss the structure and dynamics of ExoU and provide a benchmark evaluating the influence of SDSL-EPR data on protein structure prediction. In the Results and Discussion section, we detail the computational protein structure prediction pipeline, describe the available SDSL-EPR data, compare the prediction results to those from the X-ray-derived model of ExoU, and evaluate the influence of SDSL-EPR data on de novo protein structure prediction. In the Methods section, we describe the experimental approach used to obtain the SDSL-EPR data.

RESULTS AND DISCUSSION

Here, we report the results of the de novo protein structure prediction with and without the inclusion of SDSL-EPR data. The results were evaluated for of sampling accuracy and discrimination of inaccurate models as described above. We begin with an analysis of improvements in sampling accuracy when SDSL-EPR data is incorporated into the protein structure prediction algorithm. The influence of SDSL-EPR data on the discrimination of inaccurate models is then considered. This section starts with an outline of the benchmarking procedure that was used to evaluate the influence of SDSL-EPR data on de novo protein structure prediction accuracy. This outline is followed by sections providing a detailed description of the protein structure prediction protocol, an analysis of the available SDSL-EPR data for the C-terminal domain of ExoU, and a description of the algorithm used to translate the SDSL-EPR data into structural restraints that are usable by the prediction algorithm. This section is concluded by an evaluation of the predicted tertiary structure, its agreement with the X-ray-derived model and a discussion of its consistency with the SDSL-EPR data.

Summary of the Available SDSL-EPR Data for the C-Terminal Domain of ExoU. For the C-terminal domain of ExoU, seven intradomain SDSL-EPR distance measurements were available (see Table S1 and Figures S1 and 1B for details).

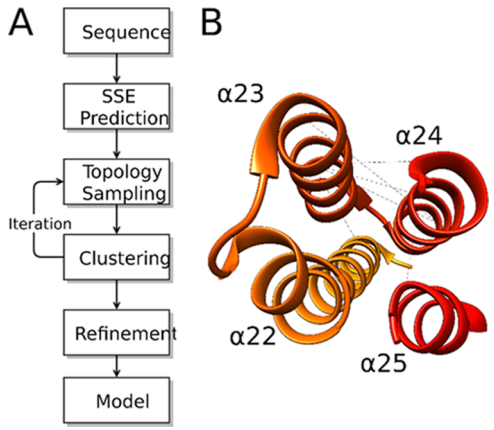


Figure 1. Protein structure prediction pipeline and SDSL-EPR data for the c-terminal domain of ExoU. (A) De novo protein structure prediction pipeline for the C-terminal domain of ExoU employed a hierarchical approach consisting of modules for secondary structure prediction, low-resolution topology sampling, and high-resolution refinement. (B) Seven intradomain SDSL-EPR measurements were available (shown as dashed lines) for the C-terminal domain of ExoU.

The X-ray-derived model is in good agreement with the restraints derived from the SDSL-EPR measurements, as indicated by an average agreement score of −0.88 (Table S1, see the following section for details regarding quantifying the agreement of models with the SDSL-EPR data). Of the seven restraints, four are between α-helices 23 and 24, one is between α-helices 22 and 23, one is between α-helix 23 and the loop region connecting α-helices 24 and 25, and one is between α-helix 24 and the loop connecting α-helices 24 and 25. As shown in Figure 1B, the SDSL-EPR restraints well-describe the relative positions of α-helices 23 and 24.

Summary of the Benchmark Setup To Evaluate the Influence of SDSL-EPR Data on De Novo Structure Prediction. The influence of SDSL-EPR data on de novo protein structure prediction was evaluated by performing two independent structure prediction runs, one with incorporated SDSL-EPR data and one in the absence of SDSL-EPR data, for the C-terminal domain of the effector protein, ExoU. The protocols for both prediction runs were predominately identical, only differing in the scoring function that was extended by a scoring term quantifying the agreement of the model with the SDSL-EPR data for one prediction run (see the following sections for details). For each prediction run, about 100 000 low-resolution models and about 50 000 high-

DOI: 10.1021/acsomega.7b00349
ACS Omega 2017, 2, 2977−2984

resolution full-atom models were sampled and subsequently analyzed under the aspects of sampling accuracy and discrimination of inaccurate models (see the following sections for details). A previously published X-ray-derived model of ExoU (PDB entry 3TU3)[2] was used as a reference structure for evaluating the sampling accuracy and model discrimination, the reported RMSD values are between the sampled models and the X-ray-derived model of ExoU. However, no information about the X-ray-derived model was used in the protein structure prediction protocol.

Protein Structure Prediction Protocol. The protein structure prediction protocol (see Figure 1A) consisted of two modules: a module for low-resolution sampling of possible topologies and a module for the construction of loop regions and high-resolution refinement of the resulting model. The two modules were connected through a data aggregation step using filtering and clustering. The low-resolution topology sampling was performed iteratively: upon conclusion of the first iteration of the low-resolution topology sampling, the most favorable models by pseudoenergy score and agreement with the SDSL-EPR data (if applicable) were selected as start models for a second round of optimization using the topology sampling module.

In the first module (Procedures S1 and S2 for details), low-resolution topology sampling, the secondary structure of the protein was predicted using PsiPred[22,23] and Jufo9D.[24] The resulting SSEs were subsequently arranged in the three-dimensional space using the de novo protein structure prediction algorithm, BCL::Fold.[13] BCL::Fold employs an MCM algorithm to sample the possible topologies arising from the predicted SSEs. The BCL::Fold prediction consists of six stages: five assembly stages and one refinement stage. In each MC step, a randomly chosen perturbation (mutate) is applied to the current protein model. The assembly and refinement stages differed in the mutates applied by the MCM algorithm. Whereas the mutates during the assemble stages apply topology-changing perturbations like large-scale translations of SSEs or swapping of SSEs, the mutates during the refinement stage apply only small-scale perturbations like rotating helices around their main axes. After each application of a mutate, the resulting protein model is evaluated using a scoring function, E. Scoring function E is the weighted sum of various knowledge-based scoring terms E_i and assigns a pseudoenergy score s to a protein model m by computing $s = E(m) = \sum_i w_i \cdot E_i(m)$.[14] The scoring terms, E_i, each evaluate different properties of the protein model such as steric interferences, residue−residue interactions, SSE−SSE packing, or residue exposure.[14] Depending on the score difference between the current model and last accepted model, a Metropolis criterion either accepts or rejects the new model.[13,14] The Metropolis criterion in conjunction with simulated annealing is used to prevent sampling trajectories from getting trapped in local pseudoenergy minima. Subsequently, the MCM algorithm resumes with the latest accepted model. The MCM optimization of each assembly and refinement stages lasted for a maximum of 4000 MC steps with the optimization terminating early if no improvement in the pseudoenergy score was achieved for 800 MC steps in a row. In total, this module resulted in 50 000 models.

Upon conclusion of the first topology sampling module, the models were ranked according to their pseudoenergy score. The best 10% of the models (∼10 000 models) were selected for clustering using a k-means implementation in R,[25] with the root-mean-square distance (RMSD) between the backbone C_α-coordinates being the metric for quantifying the dissimilarity between models (see Procedure S3 for details). The number of clusters was dynamically adjusted to maximize the average Silhouette width,[26,27] which quantifies how tight the grouping of the data points in each cluster is. Briefly, the Silhouette of a data point n is computed as $s(n) = (b(n) - a(n))/\max\{a(n),b(n)\}$, where $a(n)$ is the average dissimilarity between n and all other data points in the same cluster and $b(n)$ is the lowest average dissimilarity between n and a data point in any other cluster. The average Silhouette width of the clustering is computed accordingly as $\sum_{n=1}^{k} s(n)/k$. The Silhouette width ranges from −1 to 1 with a higher value indicating a good matching of the data elements to their respective clusters and poor matching to other clusters, therefore indicating that clustering represents the underlying data well. For structure prediction with and without SDSL-EPR data, the clustering resulted in four and seven clusters, respectively. The cluster medoids and the most favorable model by the pseudoenergy score were chosen for another round of optimization using the topology sampling module. The protocol of this optimization matched the protocol described above but used the selected models as start models. Upon conclusion of the second round of optimization, the same clustering protocol was applied and resulted in three and seven clusters, respectively. The cluster medoids and the most favorable model by the pseudoenergy score were selected as start models for the second module, that is, the construction of loop regions and high-resolution refinement of the models using Rosetta.

In the second module (see Procedure S4 for details), the Rosetta software suite[21,28] was used to construct loop regions, add side-chain coordinates, and perform a high-resolution refinement of the provided protein models. The cyclic coordinate descent (CCD) algorithm[29] was employed for construction of loop regions, and the SDSL-EPR data was incorporated into CCD and the subsequent refinement using the motion-on-a-cone (CONE) model[30] (see the following sections for details) following previously published protocols.[20] The weight of the score quantifying the agreement of the model with the SDSL-EPR data was set to 40 to ensure that the score accounts for approximately 40% of the total pseudoenergy score. For each of the provided start models, 500 full-atom models were sampled using this protocol, resulting in about 20 000 models.

Incorporating SDSL-EPR Data into Computational Protein Structure Prediction. To use SDSL-EPR spectroscopy for distance measurements in a protein, a spin label carrying a free electron needs to be introduced at the two sites of interest. The distance between the two spin-labeling sites is then determined indirectly by measuring the dipolar interaction between the two free electrons, which is inversely proportional to their cubed distance.[11,12] The indirect nature of this measurement poses challenges for using the observed data in a protein structure prediction algorithm. First, even if the backbone of the protein is inflexible, the proteins in the sample for the measurement will exhibit different conformations of the spin label, resulting in a distribution of distances rather than one observed distance. Second, depending on the type of spin label and its conformation, the distance between the free electron and the backbone of the spin-labeled residue can be rather large, adding uncertainty to the measurement. For example, for spin label MTSL, the Euclidean distance between

DOI: 10.1021/acsomega.7b00349
ACS Omega 2017, 2, 2977−2984

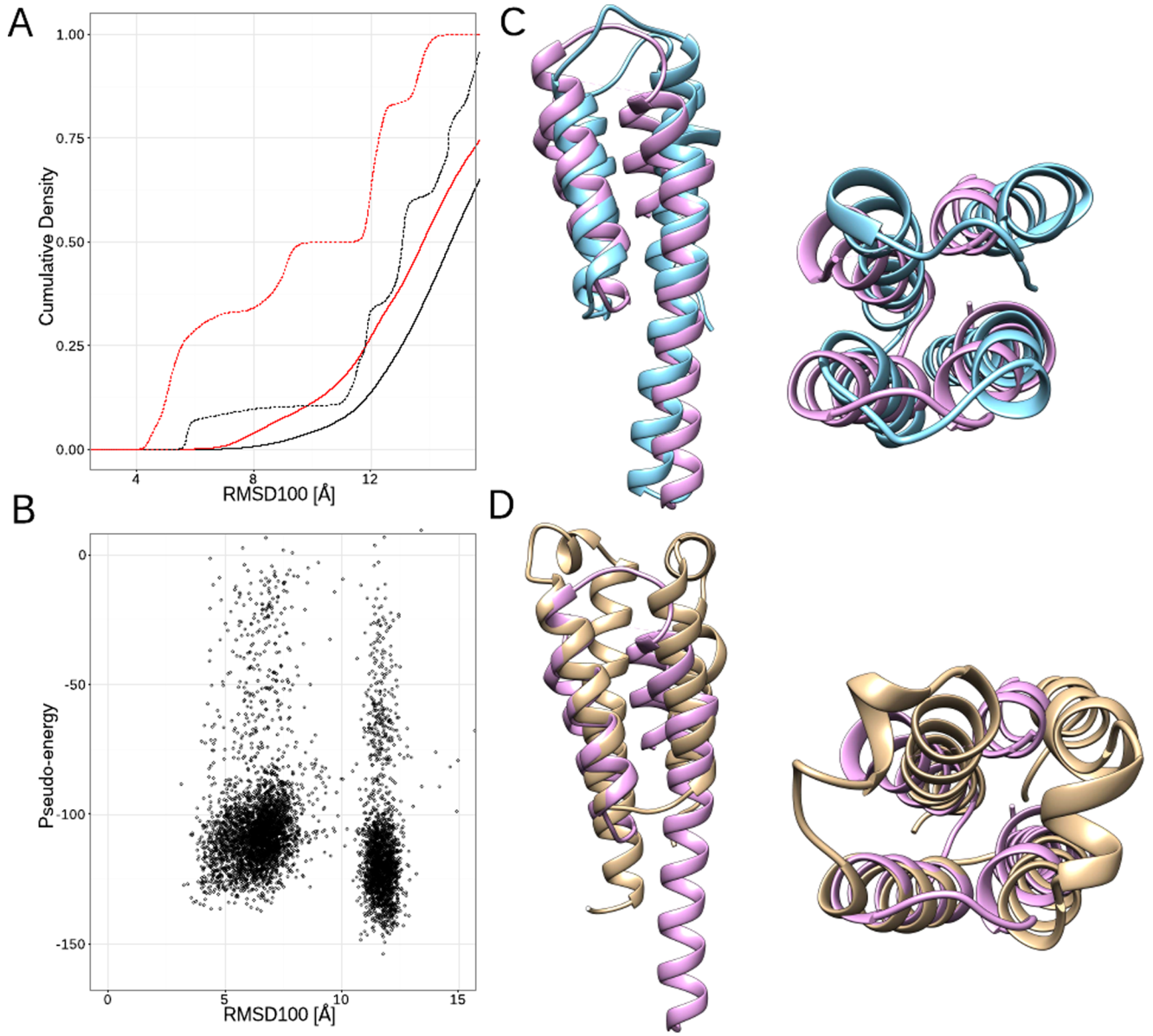


Figure 2. Prediction results for the C-terminal domain of ExoU. (A) Comparison of the sampling densities between prediction with (red) and without (black) SDSL-EPR data. Results are shown for the first (solid line) and second (dashed line) iterations of the low-resolution topology sampling. (B) Sampled models are shown as black dots with their pseudoenergy score and RMSD100 relative to the X-ray-derived model. (C) Most accurate model predicted (blue) superimposed with the X-ray-derived model (purple, PDB entry 3TU3) from top and side views. (D) Alternative model (beige) predicted by the prediction pipeline superimposed with the X-ray-derived model (purple, PDB entry 3TU3).

the C_β-atom of the spin-labeled residue and the spin label's free electron can be up to 8.5 Å.[30]

To use the distances measured in the SDSL-EPR experiment within a protein structure prediction algorithm, a function to quantify the agreement between the experimental data and a protein model needs to be defined. This function needs to capture both aforementioned properties of the SDSL-EPR measurement, the flexibility of the spin label and the indirectness of the measurement. Previously, different approaches to define such a function have been published. The motion-on-a-cone (CONE) model[18,30] uses a knowledge-based approach to account for these factors. This implicit approach provides a rapid way to estimate the probability of observing a certain C_β–C_β distance (D_{BB}) given a measured spin–spin distance (D_{SL}). The agreement score based on the CONE model is defined on the basis of the difference between D_{BB} and D_{SL}, which can range from −12 to 12 Å. The value of the scoring function ranges from 0.0, which means no agreement, to −1.0, which means best possible agreement. This approach has been successfully used for de novo prediction of membrane proteins[18] and soluble proteins that exist in multiple relevant states.[17]

Because of its significantly faster computation time, we employed the CONE model[30] to translate the measured spin–spin distances into structural restraints for the de novo protein structure prediction algorithm. For the structure prediction algorithm, the weight of the CONE-based score quantifying the agreement of the protein model with the SDSL-EPR data was set to 40, which ensured that this score accounted for about 40% of the total score, a contribution percentage for limited experimental data that provided the best prediction results in previous studies.[15] Additionally, we added a quadratic potential

DOI: 10.1021/acsomega.7b00349
ACS Omega 2017, 2, 2977−2984

Table 1. Prediction Results for the C-Terminal Domain of ExoU with and without the SDSL-EPR Data[a]

	low-resolution I			low-resolution II			high resolution		
setup	μ_{10} [Å]	τ_5 [%]	e	μ_{10} [Å]	τ_5 [%]	e	μ_{10} [Å]	τ_5 [%]	e
no data	6.0	0.0	0.6	4.9	<0.1	0.1	4.6	<0.1	0.1
SDSL-EPR data	5.1	<0.1	2.8	3.9	1.2	2.5	3.2	0.7	1.2

[a]Incorporation of SDSL-EPR data results in improved sampling accuracy and model discrimination, as shown by improvements in the average RMSD100 over the 10 most accurate models sampled (μ_{10}), in the percentage of models with an RMSD100 less than 5 Å relative to the X-ray-derived model (τ_5), and in the enrichment (e).

function to penalize the models with $D_{SL}-D_{BB}$ values outside of the range of the CONE model.[18]

De Novo Prediction Results Confirm the Correctness of the X-ray-Derived Model. The X-ray crystallography model of the ExoU/SpcU (PDB entry 3TU3)[2] structure is of high quality (resolution = 1.9 Å, R_{free} = 0.225, R_{work} = 0.191). The C-terminal domain makes few crystal lattice contacts that are overall unlikely to perturb its confirmation: V57, L55, and G82 of SpcU appear to form a hydrophobic pocket for α-helix 23 and SpcU S51 and R83 potentially hydrogen-bond to ExoU residues N657 and E636, respectively. Otherwise, SpcU does not appear to influence the structure of the C-terminal four-helix bundle. Hence, we started with the hypothesis that de novo structure prediction in conjunction with SDSL-EPR will ultimately be consistent with this conformation. The de novo prediction of the C-terminal domain of ExoU resulted in two dissimilar topologies (Figure 2B,C). Whereas one topology is represented by models exhibiting a structural dissimilarity to the X-ray-derived model as low as 3.2 Å, the other topology is structurally very dissimilar with an RMSD100 of about 12 Å relative to the X-ray-derived model. Notably, both topologies have comparable agreements with the SDSL-EPR data. The approach described in this study is orthogonal to the procedures used for obtaining the X-ray-derived model. Although there is not enough experimental data to rule out either of the two topologies, the partial convergence of the de novo method on the topology of the X-ray-derived model reassures its correctness. The topology that is structurally dissimilar to the X-ray-derived model arrives at a more favorable pseudoenergy score than the structurally similar topology (Figure 2B,C). However, this does not necessarily mean that the alternative topology is energetically more stable but could also be an artifact caused by inaccuracies of the free energy approximations. Artifacts like this have been observed in previous studies and might be eliminated by obtaining additional distance measurements.[17,18]

We were also interested in examining the experimental bimodal distance distributions observed for A629−A645 and Q649−S672 (see Figure S1 for details). To evaluate the agreement of the X-ray-derived model of ExoU with the determined distance distributions, we performed explicit simulation of the distance distribution for the double mutant A629C−A645C, as described in the Methods section. The double mutant Q649C−S672C was not evaluated because residue S672 was not resolved in the X-ray-derived model and modeling the missing coordinates would introduce additional bias. For double mutant A629−A645, explicit simulation of the spin labels did not result in a bimodal distribution but in a distinct peak at around 25.5 Å (Figure S2). For comparison, EPR spectroscopy determined two peaks: 19.3 ± 1.6 and 24.1 ± 1.9 Å (Figure S1). Taking the accuracy limit of the X-ray-derived model and the fixed backbone during the explicit simulation into account, we conclude that the X-ray-derived model is in agreement with the measured mean distance of 24.1 Å. Additional simulations were performed for the remaining double mutants that had both spin-labeling sites resolved in the X-ray-derived model. The simulated peaks matched the experimentally determined peaks well (Figure S2) given the accuracy limit of the X-ray-derived model and the fixed backbone during the simulation. For one double mutant, E636−N657, the simulation resulted in two peaks: one peak at around 18.5 Å, which agrees with the experimentally determined peak at 20.0 ± 3.6 Å, and one peak at around 13.5 Å, which would be too short to detect through the DEER experiment.

Incorporating SDSL-EPR Data Increases the Probability of Sampling Accurate Models. De novo sampling of conformations through an MCM algorithm is a statistical process. The similarity of the sampled models to the X-ray-derived model corresponds to a normal distribution. To evaluate if incorporation of SDSL-EPR data increases the probability of sampling accurate models, the shifts between the distributions resulting from de novo sampling with and without SDSL-EPR data can be compared. However, the more important aspect is the accuracy of the most accurate models alongside the percentage of accurate models. To quantify improvements in sampling accuracy, we compared the average RMSD100 values of the 10 most accurate models, μ_{10}, for the two prediction runs. We chose to compare the RMSD100 averages over 10 models instead of 1 to mitigate the effect of statistical outliers. Moreover, the percentage of models with an RSDM100 of less than 5 Å relative to the X-ray-derived model, τ_5, was compared. In addition, we also investigated whether incorporation of SDSL-EPR data results in increased clustering of the sampled models, as would be expected because the incorporated restraints should exclude conformations that significantly violate the EPR-derived restraints.

The results of the iterative protocol clearly demonstrate that incorporation of even a small number of SDSL-EPR distance restraints significantly increases the probability of sampling accurate models. Additionally, the most accurate models sampled arrive at an accuracy not observed for de novo protein structure prediction in the absence of SDSL-EPR data. This is demonstrated by changes of the μ_{10} values, which improve from 6.0 to 5.1 Å with the inclusion of the SDSL-EPR restraints for the first iteration of the low-resolution topology sampling (Figure 2A and Table 1). The τ_5 values for the first iteration were too low to be compared in a meaningful way. Another notable effect of incorporating the SDSL-EPR data was an increased clustering of the sampled models, which is likely caused by the exclusion of models that are not in agreement with the experimental data. The improved clustering is demonstrated by improvements of the average Silhouette width (see the Methods section for details), which was 0.21 for the first iteration of low-resolution topology sampling without the SDSL-EPR data and improved to 0.57 when experimental

DOI: 10.1021/acsomega.7b00349
ACS Omega 2017, 2, 2977−2984

data was included. Because of improved clustering, a more accurate set of models could be selected for the second iteration of the low-resolution topology when including the SDSL-EPR data. This is demonstrated by more favorable μ_{10} and τ_5 values after the second iteration, 3.9 Å and 1.2% as compared to 4.9 Å and less than 0.1%, respectively, when no experimental data was used. This pattern propagated to the high-resolution refinement step. For prediction with the SDSL-EPR data, the μ_{10} and τ_5 values arrived at 3.2 Å and 0.7%, whereas they were 4.6 Å and less than 0.1% for the prediction without using experimental data (Table 1).

In conclusion, incorporation of limited experimental data from SDSL-EPR spectroscopy into de novo protein structure prediction results in excluding models that violate the restraints. Although the experimental data in this test case are too sparse to unambiguously determine the tertiary structure of the C-terminal domain of ExoU, the probability of sampling accurate models is significantly improved. This was demonstrated by improvements of μ_{10} and τ_5 values, as well as improvements of the average Silhouette width of the clusters, which indicates an increased clustering of the sampled models.

Incorporation of the SDSL-EPR Data Improves Discrimination of Inaccurate Models. Distinguishing between accurate and inaccurate models resulting from de novo protein structure prediction is typically hindered by the reduced resolution of the sampled conformations that result from the relatively coarse-grained approaches used to approximate a model's free energy. This was demonstrated by the prediction results in the absence of SDSL-EPR data. Although moderately accurate models with an RMSD100 of 4.7 Å relative to the X-ray-derived model could be sampled during the first iteration of the low-resolution topology search, the employed scoring function was unable to correctly distinguish between accurate and inaccurate models, as indicted by an enrichment value of 0.6 (Table 1). Accurate models were sampled with low probability, resulting in an accordingly low density. As a consequence, accurate models could not be detected through clustering. This was demonstrated by the Silhouette scores that were 0.21 when the SDSL-EPR data were used and 0.57 otherwise, indicating a broader range of conformations considered favorable in the absence of SDSL-EPR data. In general, incorporation of the SDSL-EPR data significantly improved the scoring function's ability to distinguish between accurate and inaccurate models, which can be shown by comparing the enrichment values that arrived at 0.6, 0.1, and 0.1 for the two iterations of the low-resolution topology search and the high-resolution refinement in the absence of experimental data, respectively, but upon incorporation of the SDSL-EPR data improved to 2.8, 2.5, and 1.2, respectively (Table 1).

Limitations in Conformation Sampling and Model Discrimination Remain. The most accurate full-atom model sampled by the presented pipeline arrives at an RMSD100 of 3.2 Å (Figure 2 and Table 1) relative to the X-ray-derived model (PDB entry 3TU3), which was reported at a resolution of 1.9 Å. Therefore, the most accurate model sampled is not within the accuracy limit of the experimentally determined reference structure. Assuming that the X-ray-derived model correctly and accurately represents the protein's major population in the equilibrium, the most accurate model does not capture the protein's tertiary structure at atomic detail. This may be attributed in part to necessary simplifications when sampling conformations. Neither the low-resolution topology sampling nor the high-resolution refinement exhaustively searches the conformational space, and the most accurate model sampled could indeed be the most accurate model possible when using these methods in a de novo approach. For future studies, it will be worth investigating if this pipeline should be augmented with molecular dynamics simulations.

Although the discrimination of inaccurate models could be improved substantially through incorporation of SDSL-EPR data, as was demonstrated by the improvements of the enrichment values (Table 1), it is still not possible to reliably select the most accurate models. The models with the most favorable score cluster around RMSD100 values between 7 and 13 Å (Figure 2B). However, the difference in pseudoenergy between the models with the most favorable pseudoenergy and the models with the most favorable RMSD100 relative to the X-ray-derived model accounts for less than 15% of the most favorable score. This indicates that the discrimination problem could be resolved through additional SDSL-EPR distance measurements. In this initial study, the C-terminal domain of ExoU was predicted using only seven EPR-derived restraints, which in conjunction with the low-resolution translation of experimental distances into structural restraints is not sufficient to remove ambiguity from the prediction. Nonetheless, significant improvements were made even with this modest set of distance measurements, providing a valuable benchmark for further studies evaluating the impact of a more comprehensive set of constraints on de novo structure prediction.

CONCLUSIONS

Using EPR spectroscopy in conjunction with de novo protein structure prediction provided an orthogonal approach to probe the structure of ExoU. The prediction converged on a conformation that is topologically identical and structurally similar (RMSD100 of 3.2 Å) to the X-ray-derived model in complex with its chaperone SpcU (PDB entry 3TU3). This result confirms that the fold of the ExoU C-terminal domain in solution matches the fold when in complex with its chaperon SpcU. From a different perspective, we established a protocol to predict a model of a soluble protein from limited SDSL-EPR data using a combined approach consisting of BCL::Fold, R, and Rosetta. This approach can be applied to all soluble proteins.

METHODS

In this section, we detail the experimental methods used to obtain the SDSL-EPR data and the computational methods to explicitly simulate EPR-derived distance distributions in silico. This section is concluded by a description of the quality metrics used to evaluate the protein structure prediction results.

DEER Spectroscopy and Determination of Distance Distributions. Four-pulse DEER data were collected on a Bruker E-580 pulse EPR spectrometer (Bruker Biospin) operating at Q-band (34 GHz), equipped with an EN5107D2 resonator and a 10 W microwave amplifier. Selected MTSL-labeled double-cysteine mutants of ExoU were prepared in 20 mM 3-[*N*-morpholino]propanesulfonic acid, 145 mM NaCl, pH 7.2, using perdeuterated water and containing 25% (v/v) perdeuterated glycerol as cryoprotectant. Samples containing a final protein concentration of approximately 0.1 mM in a volume of 12 μL were flash-frozen in liquid N_2 and immediately placed in the resonator where sample temperature was

DOI: 10.1021/acsomega.7b00349
ACS Omega 2017, 2, 2977−2984

maintained at 80 K using an Oxford cryostat. Data were background-corrected and analyzed by model-free Tikhonov regularization using DeerAnalysis2011.[31]

Explicit Simulation of EPR-Derived Distance Distributions. To further evaluate the agreement of the X-ray-derived model with the SDSL-EPR data, explicit simulation of the spin label distance distribution was performed for double mutants that had both spin-labeling sites resolved in the X-ray-derived model. For the explicit simulation, the endogenous residues at the spin-labeling sites were replaced with R1A, which is a cysteine residue spin-labeled with MTSL, using Rosetta's application for a fixed backbone design, "fixbb". The resulting model was subsequently energy-optimized using Rosetta's "relax" application. The relaxation was constrained to the start coordinates to avoid introducing bias through Rosetta's scoring function. Constraining to start coordinates limited backbone perturbations to less than 0.1 Å. Per double mutant, 1000 independent trajectories were simulated and the spin−spin distances observed in each trajectory were extracted to determine the spin−spin distance distribution.

Quantitative Evaluation of the Protein Structure Prediction Results. The accuracy of the structure prediction results was evaluated under two aspects: the sampling accuracy, which is the structural similarity between the sampled models and the experimentally determined reference structure, and the discrimination of inaccurate models, which is how well the employed scoring function could distinguish between the accurate and inaccurate models. For quantifying the sampling accuracy, the protein size-normalized RMSD (RMSD100)[32] was used, which can be computed as $\mathrm{RMSD100} = \mathrm{RMSD}/\ln\sqrt{(l/100)}$, with RMSD being the RMSD between the C_α-coordinates of the two structures and l being the number of residues in the superimposition. To quantify the model discrimination, the enrichment metric[15] was used, which can be computed as $e = \frac{\#TP}{\#P} \times 10$. The sets TP and P are both subsets of the set of all sampled models. Set P contains the 10% of the models with the lowest RMSD100 relative to the experimentally determined reference structure. Set TP is computed from sets P and PS, which contains the 10% of the models with the most favorable pseudoenergy score, as $TP = P \cap PS$. Therefore, set TP contains the 10% most accurate models that are at the same time among the 10% of the models with the most favorable pseudoenergy score. Accordingly, the enrichment ranges from 0 to 10 and an enrichment value of 1.0 indicates that the selection by the employed scoring function is purely random and discrimination of inaccurate models does not take place. Enrichment values greater than 1.0 indicate that the scoring function is able to distinguish between accurate and inaccurate models, whereas enrichment values less than 1.0 indicate that the scoring function is selecting against accurate models. An enrichment value of 1.0 indicates that 10% of the most accurate models can be identified by the scoring function.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.7b00349.

Table of agreement between X-ray-derived model of ExoU and SDSL-EPR data, experimentally determined and simulated EPR spectra for ExoU, and protocol capture for de novo prediction of ExoU from EPR data (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: jfeix@mcw.edu. Phone: +1 (414) 955-4037. Fax: +1 (414) 955-6512 (J.B.F).

*E-mail: jens@meilerlab.org. Phone: +1 (615) 936-5662. Fax: +1 (615) 936-2211 (J.M.).

ORCID

Axel W. Fischer: 0000-0003-0237-7365

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This project was supported by NIH grants GM114234 (to J.B.F.) and AI104922 (to D.W.F.). Work in the Meiler laboratory is further supported through NIH (R01 GM080403 and R01 GM099842) and NSF (CHE 1305874). We want to thank the staff of the Advanced Computing Center for Research & Education (ACCRE) as well as Roy Hoffman and William C. Riner of the Center for Structural Biology at Vanderbilt University for computational support. Parts of the data analysis were performed using *R* in conjunction with the ggplot2[33] and cluster[25] packages. Figures depicting protein models were created using Chimera,[34] and composite figures were created using Inkscape.

■ REFERENCES

(1) Gendrin, C.; Contreras-Martel, C.; Bouillot, S.; Elsen, S.; Lemaire, D.; Skoufias, D. A.; Huber, P.; Attree, I.; Dessen, A. Structural Basis of Cytotoxicity Mediated by the Type III Secretion Toxin ExoU from *Pseudomonas aeruginosa*. *PLoS Pathog.* **2012**, *8*, No. e1002637.

(2) Halavaty, A. S.; Borek, D.; Tyson, G. H.; Veesenmeyer, J. L.; Shuvalova, L.; Minasov, G.; Otwinowski, Z.; Hauser, A. R.; Anderson, W. F. Structure of the Type III Secretion Effector Protein ExoU in Complex with Its Chaperone SpcU. *PLoS One* **2012**, *7*, No. e49388.

(3) Phillips, R. M.; Six, D. A.; Dennis, E. A.; Ghosh, P. In Vivo Phospholipase Activity of the *Pseudomonas aeruginosa* Cytotoxin ExoU and Protection of Mammalian Cells with Phospholipase A2 Inhibitors. *J. Biol. Chem.* **2003**, *278*, 41326−41332.

(4) Sato, H.; et al. The mechanism of action of the *Pseudomonas aeruginosa*-encoded type III cytotoxin, ExoU. *EMBO J.* **2003**, *22*, 2959−2969.

(5) Shaver, C. M.; Hauser, A. R. Relative Contributions of *Pseudomonas aeruginosa* ExoU, ExoS, and ExoT to Virulence in the Lung. *Infect. Immun.* **2004**, *72*, 6969−6977.

(6) Roy-Burman, A.; Savel, R. H.; Racine, S.; Swanson, B. L.; Revadigar, N. S.; Fujimoto, J.; Sawa, T.; Frank, D. W.; Wiener-Kronish, J. P. Type III Protein Secretion Is Associated with Death in Lower Respiratory and Systemic *Pseudomonas aeruginosa* Infections. *J. Infect. Dis.* **2001**, *183*, 1767−1774.

(7) Allewelt, M.; Coleman, F. T.; Grout, M.; Priebe, G. P.; Pier, G. B. Acquisition of Expression of the *Pseudomonas aeruginosa* ExoU Cytotoxin Leads to Increased Bacterial Virulence in a Murine Model of Acute Pneumonia and Systemic Spread. *Infect. Immun.* **2000**, *68*, 3998−4004.

(8) Tyson, G. H.; Halavaty, A. S.; Kim, H.; Geissler, B.; Agard, M.; Satchell, K. J.; Cho, W.; Anderson, W. F.; Hauser, A. R. A Novel Phosphatidylinositol 4,5-Bisphosphate Binding Domain Mediates Plasma Membrane Localization of ExoU and Other Patatin-like Phospholipases. *J. Biol. Chem.* **2015**, *290*, 2919−2937.

(9) Tessmer, M. H.; Anderson, D. M.; Buchaklian, A.; Frank, D. W.; Feix, J. B. Cooperative substrate-cofactor interactions and membrane

DOI: 10.1021/acsomega.7b00349
ACS Omega 2017, 2, 2977−2984

localization of the bacterial PLA2 enzyme, ExoU. *J. Biol. Chem.* **2017**, 3411−3419.
(10) Benson, M. A.; Komas, S. M.; Schmalzer, K. M.; Casey, M. S.; Frank, D. W.; Feix, J. B. Induced Conformational Changes in the Activation of the *Pseudomonas aeruginosa* type III Toxin, ExoU. *Biophys. J.* **2011**, *100*, 1335−1343.
(11) Jeschke, G. DEER Distance Measurements on Proteins. *Annu. Rev. Phys. Chem.* **2012**, *63*, 419−446.
(12) de Vera, I. M. S.; Blackburn, M. E.; Galiano, L.; Fanucci, G. E. Pulsed EPR distance measurements in soluble proteins by Site-Directed Spin Labeling (SDSL). *Curr. Protoc. Protein Sci.* **2013**, 17.17.1−17.17.29.
(13) Karakaş, M.; Woetzel, N.; Staritzbichler, R.; Alexander, N.; Weiner, B. E.; Meiler, J. BCL::Fold - De Novo Prediction of Complex and Large Protein Topologies by Assembly of Secondary Structure Elements. *PLoS One* **2012**, *7*, No. e49240.
(14) Woetzel, N.; Karakaş, M.; Staritzbichler, R.; Müller, R.; Weiner, B. E.; Meiler, J. BCL::Score-Knowledge Based Energy Potentials for Ranking Protein Models Represented by Idealized Secondary Structure Elements. *PLoS One* **2012**, *7*, No. e49242.
(15) Fischer, A. W.; Heinze, S.; Putnam, D. K.; Li, B.; Pino, J. C.; Xia, Y.; Lopez, C. F.; Meiler, J. CASP11 − An Evaluation of a Modular BCL::Fold-Based Protein Structure Prediction Pipeline. *PLoS One* **2016**, *11*, No. e0152517.
(16) Hofmann, T.; Fischer, A. W.; Meiler, J.; Kalkhof, S. Protein structure prediction guided by crosslinking restraints − A systematic evaluation of the impact of the crosslinking spacer length. *Methods* **2015**, *89*, 79−90.
(17) Fischer, A. W.; Bordignon, E.; Bleicken, S.; García-Sáez, A. J.; Jeschke, G.; Meiler, J. Pushing the size limit of de novo structure ensemble prediction guided by sparse SDSL-EPR restraints to 200 residues: The monomeric and homodimeric forms of BAX. *J. Struct. Biol.* **2016**, *195*, 62−71.
(18) Fischer, A. W.; Alexander, N. S.; Woetzel, N.; Karakas, M.; Weiner, B. E.; Meiler, J. BCL::MP-Fold: Membrane protein structure prediction guided by EPR restraints. *Proteins: Struct., Funct., Bioinf.* **2015**, *83*, 1947−1962.
(19) Putnam, D. K.; Weiner, B. E.; Woetzel, N.; Lowe, E. W.; Meiler, J. BCL::SAXS: GPU accelerated Debye method for computation of small angle X-ray scattering profiles. *Proteins: Struct., Funct., Bioinf.* **2015**, *83*, 1500−1512.
(20) Hirst, S.; Alexander, N.; Mchaourab, H. S.; Meiler, J. ROSETTAEPR: An Integrated Tool for Protein Structure Determination From Sparse EPR Data. *Biophys. J.* **2011**, *100*, No. 216a.
(21) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, Ca; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y. E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **2011**, *487*, 545−574.
(22) McGuffin, L. J.; Bryson, K.; Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **2000**, *16*, 404−405.
(23) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.
(24) Leman, J. K.; Mueller, R.; Karakas, M.; Woetzel, N.; Meiler, J. Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 1127−1140.
(25) Maechler, M.; Rousseeuw, P.; Struyf, A.; Hubert, M.; Hornik, K. *Cluster: Cluster Analysis Basics and Extensions*, 2015, 3−81
(26) Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53−65.
(27) de Amorim, R. C.; Hennig, C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Inf. Sci.* **2015**, *324*, 126−145.
(28) Kaufmann, K. W.; Lemmon, G. H.; DeLuca, S. L.; Sheehan, J. H.; Meiler, J. Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry* **2010**, *49*, 2987−2998.
(29) Canutescu, A. A.; Dunbrack, R. L. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **2003**, *12*, 963−972.
(30) Alexander, N.; Al-Mestarihi, A.; Bortolus, M.; Mchaourab, H.; Meiler, J. De Novo High-Resolution Protein Structure Determination from Sparse Spin-Labeling EPR Data. *Structure* **2008**, *16*, 181−195.
(31) Jeschke, G.; Chechik, V.; Ionita, P.; Godt, A.; Zimmermann, H.; Banham, J.; Timmel, C. R.; Hilger, D.; Jung, H. DeerAnalysis2006 - a comprehensive software package for analyzing pulsed ELDOR data. *Appl. Magn. Reson.* **2006**, *30*, 473−498.
(32) Carugo, O.; Pongor, S. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.* **2001**, *10*, 1470−1473.
(33) Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, 2009; pp 27−41.
(34) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

DOI: 10.1021/acsomega.7b00349
ACS Omega 2017, 2, 2977−2984