THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

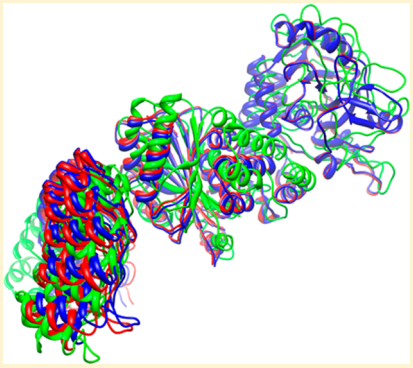
Automated Structure Refinement for a Protein Heterodimer Complex Using Limited EPR Spectroscopic Data and a Rigid-Body Docking Algorithm: A Three-Dimensional Model for an Ankyrin-CDB3 Complex

Sarah J. Edwards,[†,⊥] Christopher W. Moth,[†,⊥] Sunghoon Kim,[‡] Suzanne Brandon,[‡] Zheng Zhou,[‡] Charles E. Cobb,[‡] Eric J. Hustedt,[‡] Albert H. Beth,[‡,⊥] Jarrod A. Smith,[§,⊥] and Terry P. Lybrand*[,†,∥,⊥]

[†]Department of Chemistry, [‡]Department of Molecular Physiology & Biophysics, [§]Department of Biochemistry, [∥]Department of Pharmacology, [⊥]Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37235, United States

Supporting Information

ABSTRACT: We report here specialized functions incorporated recently in the rigid-body docking software toolkit TagDock to utilize electron paramagnetic resonance derived (EPR-derived) interresidue distance measurements and spin-label accessibility data. The TagDock package extensions include a custom methanethiosulfonate spin label rotamer library to enable explicit, all-atom spin-label side-chain modeling and scripts to evaluate spin-label surface accessibility. These software enhancements enable us to better utilize the biophysical data routinely available from various spin-labeling experiments. To illustrate the power and utility of these tools, we report the refinement of an ankyrin:CDB3 complex model that exhibits much improved agreement with the EPR distance measurements, compared to model structures published previously.



INTRODUCTION

Electron paramagnetic resonance (EPR) spectroscopy, when combined with detailed computer modeling, is a powerful and valuable tool for oligomeric protein complex structure determination and can be used to probe structural details for these protein assemblies in the many situations where X-ray crystallography or NMR spectroscopy are not practical or possible. Appropriate EPR experiments for spin-labeled protein complexes can provide distance measurements over a range of 5−80 Å,[1−4] making EPR an ideal technique to study the structures of multiprotein complexes. EPR studies also require markedly smaller sample amounts than NMR experiments or crystallization trials and can be applied routinely to samples in various physical environments, including cell membranes. EPR techniques may be particularly useful when protein complex formation involves highly flexible structures.[5]

The measurements obtained from EPR double electron−electron resonance (DEER) experiments can be analyzed in terms of Gaussian distributions[3,4,6] of the distances between the unpaired electrons in the paramagnetic spin labels as given by

$$P(R) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(R-R_0)^2/2\sigma^2} \quad (1)$$

where P is the relative probability of observing a particular distance R, R_0 is the center of the Gaussian, and σ describes the width of the Gaussian. This distance distribution reflects an ensemble of conformations and is determined by the complex interplay of global protein conformation, spin-label flexibility, and spin-label interaction with neighboring side chains. As a result, interpretation of these distance distributions based simply on side-chain C_β−C_β distances may introduce errors as large as 12−14 Å.[7,8] However, DEER distance measurements are typically more precise.[6] Thus, it should be advantageous to model spin-label side chains explicitly during three-dimensional model construction or assessment. We have previously developed the TagDock software package to treat rigid body docking problems.[9] Here, we describe an enhanced version of the TagDock software suite, which now includes an MTSSL side-chain rotamer library to facilitate explicit spin label modeling in the protein docking calculations. To illustrate the utility of these TagDock enhancements, we present results for detailed structural refinement of an important human erythrocyte cytoskeletal complex formed between ankyrin-R and the cytosolic domain of anion exchange protein AE1, or band 3 (CDB3), using EPR pair distance measurements and spin-label accessibility data[10] to filter candidate complex models.

Human erythrocytes are unique among cells in the body due to their unusual biconcave discoidal shape, high structural integrity, and elasticity. These unique properties are essential for long-term survival under the turbulent conditions often present in the

Received: October 8, 2013
Revised: April 10, 2014
Published: April 11, 2014

dx.doi.org/10.1021/jp4099705 | J. Phys. Chem. B 2014, 118, 4717−4726

circulatory system. These unique structural and mechanical properties are due to the extensive membrane skeleton, composed primarily of actin and spectrin, and the connections formed between the membrane skeleton and integral membrane proteins.[11−13] One major connection type is composed of ankyrin-R and protein 4.2, which form a complex with CDB3.[14−18] Defects in this connection complex, sometimes due to mutations in ankyrin-R, can result in spherical erythrocytes with reduced elasticity and diminished structural integrity.[19] These deformed erythrocytes are more easily ruptured, leading to a class of clinical conditions known collectively as hereditary spherocytosis, manifested by hemolytic anemia and associated problems.

A detailed structure for the ankyrin-R:CDB3 complex would help us better understand how various ankyrin or CDB3 mutations alter the skeletal membrane connections and produce spherocytosis. However, as is the case with many heterodimeric protein complexes, there is no high-resolution crystal structure yet for ankyrin-R:CDB3 and we only have high-resolution structures for individual protein components of this connection network. We reported recently a detailed model for the complex formed between the cytoplasmic domain of erythrocyte band 3 (CDB3; PDB code 1HYN) and repeats 13−24 in the membrane binding domain of ankyrin-R (ankyrin-R; PDB code 1N11), using EPR distance measurements[10] to guide a combined automated molecular docking and manual molecular model building exercise. Here we describe a refined ankyrin-CDB3 complex model, generated using the TagDock toolkit, that exhibits much improved agreement with the biophysical data while retaining the general features reported for our previous model.

The computational tools and protocols that we present here are applicable to any macromolecular dimer modeling task, when limited biophysical data such as distance measurements are available from EPR, fluorescence resonance energy transfer (FRET), solid-state NMR spectroscopy, or other biophysical techniques. Thus, the methods we describe should have broad practical utility.

METHODS

New Developments in TagDock. As described previously,[9] TagDock is a software toolkit that produces structures for macromolecular complexes by generating randomly posed docked pairs (decoys), starting from rigid structures for each monomer. The program first generates all geometrically possible docking poses and then filters all docking solutions using a penalty function that determines the agreement for each pose with available biophysical data, most typically a set of intermonomer distance measurements obtained from, e.g., EPR, NMR, or FRET experiments. In the current work, we extend TagDock to include explicit MTSSL spin labels (Figure 1) at specified amino acid positions, subject to steric constraint considerations. With this option, TagDock computes interresidue distances as Boltzmann-weighted averages of internitroxide spin-label distances determined from all spin-label conformers that can be accommodated at each site. A histogram of distance−density, in 0.1 Å bins, can be created, for qualitative comparison to experimentally determined distance distributions.

A rotamer library was constructed using molecular mechanics methods in AMBER v10,[20] to determine a Boltzmann-weighted set of sterically allowed MTSSL spin-label rotamers at each site. While there is considerable precedence for rotamer libraries,[21−24] a general-purpose library for docking calculations

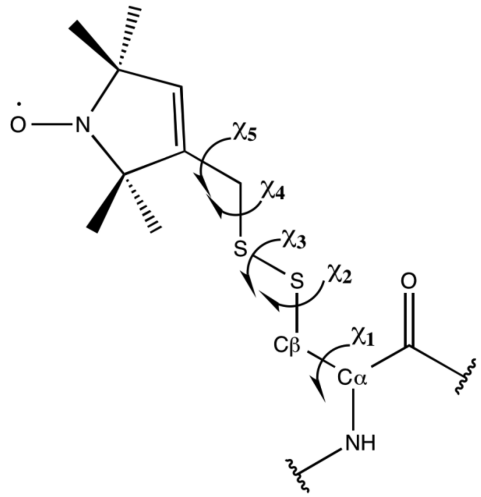


Figure 1. Methanethiosulfonate spin-label with χ angles labeled.

derived from molecular dynamics is quite distinct. We used molecular dynamics simulations to observe accessible rotamer conformations for surface-exposed MTSSL side chains in the context of a globular protein structure. We chose T4 lysozyme (T4L) as our model protein system. We modified the T4L structure (PDB entry 2OU8) by replacing native side chains at positions 5, 16, 38, 53, 65, 89, 109, 135, and 144 with MTSSL side chains. These positions were carefully selected such that (a) each position corresponds to a residue site that has been labeled successfully in previous experimental studies, so we know an MTSSL side chain can be accommodated; (b) the side chain was located at the surface of the protein, with the $C_\alpha-C_\beta$ bond vector oriented away from the center of mass of the protein; and (c) no spin-label position was close enough to any other modified position such that any two MTSSL side chains could interact directly with each other and thus bias the conformational sampling, allowing us to run simulations with all nine labels included.

We used previously published molecular mechanics force field parameters for MTSSL[25] to run these calculations. The χ_3 dihedral centered on the disulfide linkage favors values of $\pm 90°$ and is well-known to have a large energy barrier for interconversion between the two low-energy conformers. Therefore, we ran two simulations, one with χ_3 set initially at $+90°$, and another with χ_3 set at $-90°$, so as not to bias the rotamer states produced by our simulations due to the starting conformations for χ_3.

Both starting structures ($\chi_3 = +90°$; $\chi_3 = -90°$) were solvated in a truncated octahedral box of simple point charge (SPC) waters with a 12.2 Å solvent layer in all directions, producing a total system size of ~42,300 atoms. The protein atoms were then held fixed and the solvent was relaxed with 100 steps of steepest descent minimization, followed by 1900 steps of conjugate gradient minimization. Next, the solvent was held fixed and the protein was minimized and finally, the whole system was minimized without restraints.

The system was then heated to 298 K and equilibrated in a 310 ps procedure consisting of six separate molecular dynamics (MD) runs of between 10 and 80 ps each, using the Anderson temperature coupling option. At the beginning of the procedure, solute atoms were restrained to their starting positions with a force constant of 16 kcal/mol, and the MD step size was set to 0.5 fs. Restraints were iteratively decreased and the MD step size was

dx.doi.org/10.1021/jp4099705 | J. Phys. Chem. B 2014, 118, 4717−4726

iteratively increased in subsequent runs until the solute atoms were unrestrained and the step size was 1.5 fs. Constant volume dynamics were run for the first 70 ps, switching to constant pressure for the remainder of the procedure.

The equilibrated $\chi_3 = -90°$ and $\chi_3 = +90°$ systems were then each simulated for 200 ns using constant pressure MD with isotropic position scaling. A trajectory snapshot was saved every 1 ps, yielding 400,000 individual snapshots for each of the nine MTSSL side chains. The resulting 3.6 million MTSSL conformations were extracted from the trajectories for subsequent conformational analysis.

The observed values for each of the χ_1–χ_5 MTSSL dihedral angles over the two 200 ns MD simulations are depicted as polar probability plots in Figure 2. The values for the $\chi_3 = -90°$ and $\chi_3 = +90°$ simulations are overlaid in the right-hand panels, illustrating how the χ_3 conformation impacts the probability of observing each of the three low-energy states for the other side-chain dihedral angles. The left panels illustrate the overall probabilities for each dihedral value, summed over both simulations. As these data show clearly, our observed populations for χ_1–χ_5 are nearly identical to similar analysis reported by Polyhach et al.,[22] using simulations of the free MTSSL spin label only. The gauche+ population for χ_1 is more prevalent in Jeschke's simulations compared to our results, but this is the only observable difference in the two studies. This difference is presumably because their simulations were performed for an isolated MTSSL residue while we simulated the spin labels in an intact protein. A gauche+ conformation at χ_1 can incur steric clashes with the protein backbone, which is why this conformation is less prevalent in our simulations. Since these two libraries were derived using quite different simulation protocols, the overall excellent agreement between these two independent analyses lends an added measure of confidence to our rotamer libraries (i.e., the rotamer states observed do not appear to be strongly dependent on the exact conformational sampling procedures, potential functions, and so on).

Each of the 3.6 million MTSSL side-chain conformations captured during our MD simulations were classified into one of 162 possible rotamer states, defined by binning the measured values for the χ_1, χ_2, χ_4, and χ_5 dihedral angles (Figure 1) into g+ (0°–120°), g− (−120°–0°), or trans (120°–240°) conformations, and binning χ_3 into ±90° conformations. The rotamer states were then counted and ranked by frequency of occurrence. Then, for each rotamer state, each of the five χ values were averaged over all of the conformers that were classified to that state and stored as the values that define that rotamer conformation. We observed 155 of the possible 162 states, with the most frequent state representing 5.5% of the total population (~198,000 total observations), while the three least frequent rotamer states were observed only 4 times each. A cutoff was then applied such that all observed rotamer states with a probability density of at least 0.0005% were retained in our model, producing a library of the most frequently observed 146 unique rotamers (Supporting Information Table S1). We also used the same procedure to create rotamer libraries classified using only four dihedrals (χ_1–χ_4) or three dihedrals (χ_1–χ_3), which produced 51 and 18 rotamer states, respectively (Supporting Information Tables S2 and S3). The χ_4 and χ_5 dihedral angles that were not directly used in the classification of the three- and four-dihedral rotamer libraries are still reported as the corresponding averages in those libraries. These simplified rotamer libraries could be used to further enhance computational efficiency but were not used in this study.

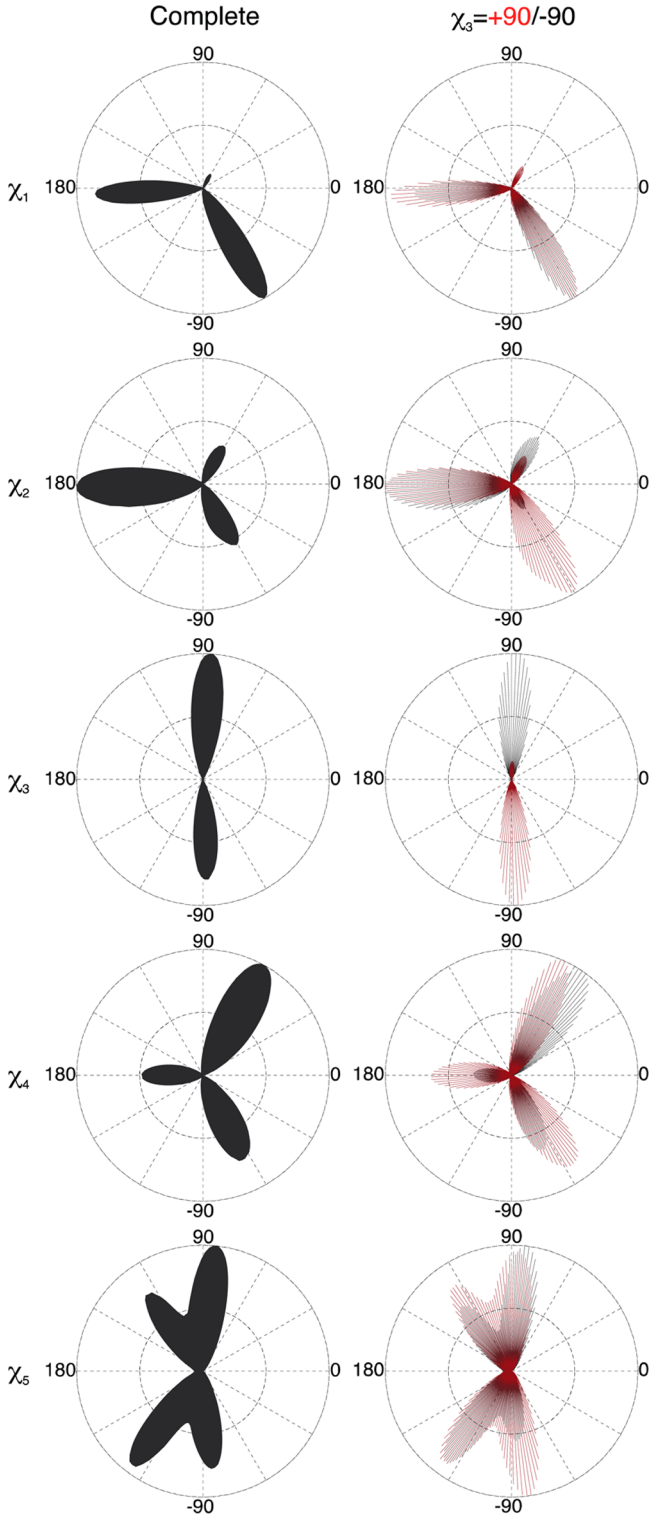


Figure 2. Polar probability plots for χ_1–χ_5 rotamers of the MTSSL spin label, determined from molecular dynamics simulations. On the right side, probabilities for our simulations with χ_3 initialized at +90° (red) and −90° (black) are superimposed. On the left side, the two are summed to give the probability for each dihedral taken from all snapshots of both simulations.

The rotamer library is input to TagDock as a simple text file, with each line specifying one of our observed MTSSL rotamers by listing its five χ dihedral angles and the probability density for that rotamer. In the present work, 146 rotamers, ranging in

dx.doi.org/10.1021/jp4099705 | J. Phys. Chem. B 2014, 118, 4717–4726

probability from 5.5% to 0.0005%, are input in this manner. Each MTSSL rotamer is inserted at each spin-label position in both isolated protein monomers prior to the initial docking step. TagDock then filters each rotamer candidate by applying one of four user-selectable algorithms. The user can elect to simply incorporate all 146 rotamer states at each position (i.e., no bump-checking), or else employ a bump-check algorithm to eliminate rotamers that form unfavorable steric interactions with either backbone (N, C, CA, O) atoms and/or neighboring side-chain heavy atoms. The bump-check algorithm itself is simple and efficient: when the distance between a pair of heavy atoms (i.e., non-hydrogen atoms) is less than 85% of the sum of their van der Waals radii,[22,26] a steric clash is noted and the spin-label rotamer candidate is eliminated on that basis. Rotamers that pass a backbone bump check, but fail when side-chain clashes are considered, can often be accommodated with minor rearrangements of neighboring protein side chains, and we use the SCWRL 4.0[27] software package to make these adjustments objectively and quickly. SCWRL 4.0 leverages a backbone-dependent rotamer library to globally sample a protein's energetically accessible side-chain conformations. Though SCWRL 4.0 is most frequently utilized for *in-silico* mutagenesis experiments, it can also be used to repack side chains in response to user-defined, occlusive quasi-atom hard spheres. SCWRL 4.0 takes fixed protein backbone atom coordinates and generates new side-chain heavy atom coordinates, which globally minimize potential energy as calculated from a combination of an anisotropic hydrogen bonding term and van der Waals contacts. Sophisticated computational algorithms in SCWRL 4.0 efficiently manage the geometric and combinatorial challenges of rapid side-chain sampling. Even with large proteins, SCWRL 4.0 typically returns repacked structures within seconds.

For each MTSSL rotamer state we consider, we generate a PDB file that contains fixed protein backbone coordinates and the MTSSL side-chain atoms defined as a collection of hard spheres. We then use SCWRL to globally repack all remaining protein side chains around this hard-sphere site. The resulting repacked protein structure is again bump-checked by TagDock, and all MTSSL rotamer states that pass this second side-chain atom bump-check step are added to the set of accessible rotamers associated with that spin-label position. Using this protocol to iterate over all MTSSL rotamer states that fail the initial side-chain bump check, TagDock identifies the largest plausible set of MTTSL rotamer conformations at each spin-label position. All MTSSL rotamer conformations selected for each monomer partner are retained in all subsequent docking trials, with no further adjustment or reevaluation during the docking pose calculations.

For each decoy pose generated, interresidue distances are calculated for all accepted spin-label rotamer states at each position, using the N−O bond vector midpoint for each spin-label residue. This calculation yields a probability-weighted distance distribution for each residue pair (i.e., a Boltzmann-weighted distance distribution, since our rotamer state populations are derived from long MD trajectories). We emphasize that these distance distributions will not necessarily reproduce the fine detail of an experimental DEER distribution profile, as we do not at present include local backbone dynamics in these calculations, but backbone motion will make a significant contribution to the distribution profile in some cases. We have shown in previous studies for a large number of test examples[9] that we can almost always obtain good docking results without inclusion of backbone dynamics, provided we generate a sufficiently large ensemble of docking poses during the coarse-docking phase so that the intermonomer distances populate the full distance distribution ranges observed in the DEER experiments.

Scoring Function. Structure refinement protocols that utilize distance data require a metric that can efficiently and objectively evaluate structures, both for ranking the relative quality of structures (i.e., how well do the structures fit the experimental data) and for automatic filtering of large numbers of structures (i.e., data reduction via automatic elimination of structures that are inconsistent with the experimental data). Each experimental distance restraint is typically represented as a distance distribution or possibly a distance range, rather than a discrete distance value, so we need a scoring function that is designed specifically to handle these types of distance restraint data. The distance data generated in NMR NOE experiments likewise are represented as distance ranges, since the NMR experiments also sample an ensemble of conformations that each have unique values for any specific interatomic distance. The harmonic penalty function we used previously for filtering structures is not well suited to these types of distance data sets, since a harmonic function "emphasizes" the median distance value, while penalizing distances that deviate from the median but still fall within the experimental distance range. Therefore, we have implemented a flat-bottomed harmonic scoring function, similar to those used widely by the NMR structure refinement community.[20,28−30] Any distance that falls within the allowable range specified by the "flat-bottomed" region of this function incurs no penalty. This scoring function has the form

$$\text{score} = \sum_{i=1}^{N} \nu_i \tag{2}$$

where N is the number of individual restraints and ν_i is the individual penalty score,

$$\nu_i = \begin{cases} \left(\left(r_{i_{\text{exp}}} - \sigma\right) - r_{i_{\text{decoy}}}\right)^2 & \text{if } \left[\left(r_{i_{\text{exp}}} - \sigma_i - h_{range}\right) < r_{i_{\text{decoy}}} < \left(r_{i_{\text{exp}}} - \sigma_i\right)\right] \\ 0 & \text{if } \left[\left(r_{i_{\text{exp}}} - \sigma_i\right) \le r_{i_{\text{decoy}}} \le \left(r_{i_{\text{exp}}} + \sigma_i\right)\right] \\ \left(\left(r_{i_{\text{exp}}} + \sigma\right) - r_{i_{\text{decoy}}}\right)^2 & \text{if } \left[\left(r_{i_{\text{exp}}} + \sigma_i\right) < r_{i_{\text{decoy}}} < \left(r_{i_{\text{exp}}} + \sigma_i + h_{range}\right)\right] \end{cases} \tag{3}$$

where $r_{i_{\text{exp}}}$ is an experimental distance, $r_{i_{\text{decoy}}}$ is the corresponding distance in the decoy being scored, σ is the associated width of the distance distribution, and h_{range} is the width of the range over which the scoring function is harmonic. The default value we use for h_{range} is 4 Å, but it is a user-tunable parameter. Outside of the range defined in eq 3, $(r_{i_{\text{exp}}} - \sigma_i - h_{\text{range}}) < r_{i_{\text{decoy}}} < (r_{i_{\text{exp}}} + \sigma_i + h_{\text{range}})$, the scoring function becomes linear with a slope equal to that at the point $(r_{i_{\text{exp}}} + \sigma_i + h_{\text{range}})$ or $(r_{i_{\text{exp}}} - \sigma_i - h_{\text{range}})$, illustrated in Figure 3. The utility of this function for structure selection and/or refinement using experimental distance range data is well-documented.[28−30]

We have also implemented a feature to compute nitroxide spin-label surface accessibility at labeled positions. This option is accomplished with a simple script that exports a PDB structure file and then uses an external program to calculate solvent accessibility. We use the MSMS program[31] for this operation, but the user can specify any suitable surface area calculation software for this calculation via simple modification of the control script. As with inter-residue distance measurements, residue surface accessibility data are not used as restraints during the docking procedure, but rather used in the filtering step to determine

dx.doi.org/10.1021/jp4099705 | J. Phys. Chem. B 2014, 118, 4717−4726

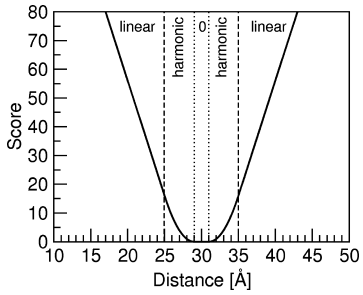


Figure 3. Illustration of the flat-bottomed harmonic scoring function with $r_{i_{exp}}$ = 30 Å, σ = 1 Å, and h_{range} = 4 Å. The flat-bottomed harmonic score is calculated as 0 for r within σ of $r_{i_{exp}}$, harmonic between σ and $\sigma + h_{range}$ from $r_{i_{exp}}$, and linear outside this region. The value of 1 Å for σ is similar to the experimental values of σ found in Supporting Informaton Table S4.

which docking poses are consistent with the experimental accessibility data.

RESULTS

Using the published experimental distances and distributions[32] (Table S4, Supporting Information), we first generated structures for the CDB3 homodimer. We followed the computational protocol reported in previous test calculations.[9] Ideally, we would begin docking calculations with the crystal structures for the independent protein monomers, but there are neither X-ray nor solution-phase structures for CDB3 monomer. Since the individual CDB3 monomer chains in the homodimer crystal have similar, but unique, conformations, designated P and Q, a docking calculation that begins with these unique structures would clearly be biased and would not provide a meaningful assessment of program performance. We therefore performed a 10 ns molecular dynamics simulation with a Generalized Born continuum solvent model using AMBER.[20] We applied weak positional restraints to maintain backbone atoms near the crystallographic coordinates for P and Q monomers, while allowing side-chain atoms to move freely, and then used the final configurations for P and Q monomers from the dynamics trajectory to begin docking calculations. First, a coarse-resolution search was performed. The second molecule of the dimer complex was randomly and repeatedly rotated and translated onto the surface of a large virtual sphere centered on the first molecule. Next, the shortest intermonomer C_β distance was calculated, and the second molecule was translated toward the first molecule by the distance, with no specific tests for unreasonable physical overlaps at this stage. We generated 5×10^5 random docking poses and then selected the 200 poses that best satisfied the experimental distance measurements. We next performed higher resolution docking searches, using these 200 poses as starting structures. The high-resolution docking phase entailed a series of more focused docking attempts, restrained by progressively smaller translation and rotation angle values, all driven by a Monte Carlo optimization procedure. During the first phase of high-resolution docking, translation moves were restricted to a range [0.0−3.0 Å] and rotation about each axis to angle values from the range of 0.0−15.0°. These search ranges were tightened, or "focused" during subsequent iterations until the translation moves were restricted to the range of 0.0−1.0 Å and rotations to the range of 0.0−1.0°. We retained the 200 top docking poses, i.e., those poses that gave the best agreement with the experimental distance measurements, for final analysis.

We performed two sets of test calculations: one with simple C_β−C_β distance filtering reported previously,[9] and a second calculation using explicit all-atom spin labels and our new rotamer library (together with the SCWRL option as described in Methods). In both cases, we generated 500,000 decoys using individual monomers from the homodimer crystal structure and performed focused refinement on the best 200 decoys, scoring with the flat-bottom harmonic function. We then plotted the penalty score for each of the 200 refined docking poses versus the C_α root mean square deviation (RMSD) for that pose relative to the homodimer crystal structure, as shown in Figure 4. The plot

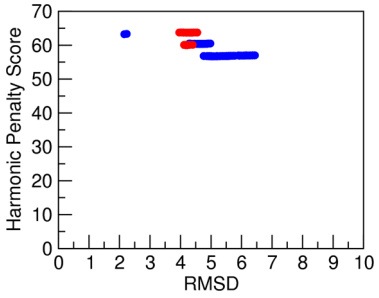


Figure 4. C_α RMSD with respect to the crystal structure of CDB3 for the top 200 TagDock results from the rigid C_β algorithm previously published[9] (red) and incorporating the MTSSL rotamer library (blue). In both cases, we use the flat-bottom harmonic scoring function. We see similar degeneracy for both algorithms; however, the incorporation of the spin-label rotamer library yields a third, lower RMSD cluster. In both cases we used all experimental distances and distributions from Zhou et al.[32]

shows that both filtering strategies yield two sets of docking poses that cluster tightly around two different penalty score values (we should note that there is no statistically significant difference between the two scores for either filtering method). However, the filtering calculation based on explicit spin-label side-chain modeling also produced a third penalty score cluster with smaller RMSD values relative to the crystal structure. Specifically, filtering based on simple C_β distances yields an optimal pose with a 4.0 Å RMSD relative to the crystal structure, while the optimal pose we generate when we use explicit spin-label side chains and the rotamer library for conformational sampling has a 2.2 Å RMSD relative to the crystal structure, as shown in Figure 5. We should note that the C_β filtering results reported here are based on the new flat-bottom harmonic scoring function, so the results differ slightly from those reported previously.[9]

We next generated structures for the ankyrin-CDB3 complex, starting with crystal structures for the isolated ankyrin-R fragment (repeats 13−24)[33] (PDB code 1N11) and CDB3[34] (PDB code 1HYN), and used both the side-chain C_β approximation and explicit spin-label side-chain modeling for distance measurements. We used the experimental distances and distributions published previously[10] to filter results from both calculations with our new flat-bottomed harmonic scoring function. In Figure 6, we show the docking pose closest to the geometric mean of the top 200 candidates for each filtering protocol, as well as the structure closest to the geometric mean of the top 30 poses generated previously with RosettaDock and manual modeling building.[10] The top 30 models reported previously have scores ranging from 285 to 470, while the geometric mean score is 53 for the C_β−C_β method and 22 for the

dx.doi.org/10.1021/jp4099705 | J. Phys. Chem. B 2014, 118, 4717−4726

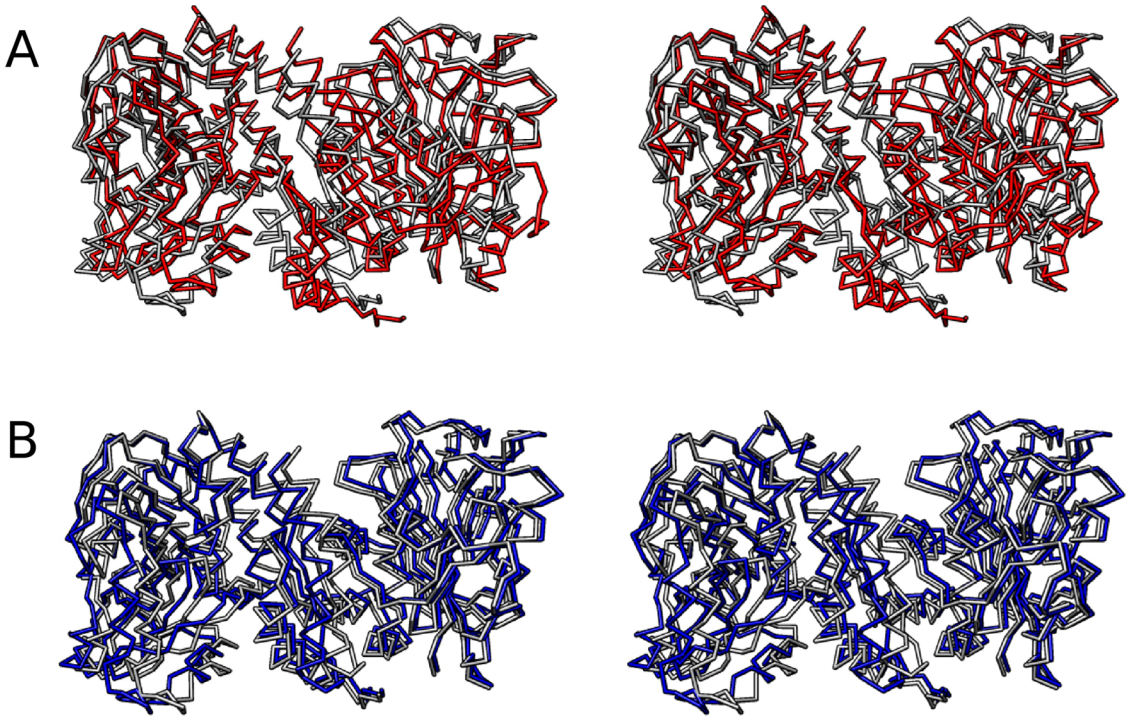


Figure 5. C_α backbone stereoplots for TagDock results shown in Figure 4, overlaid on the crystal structure of the CDB3 homodimer (gray, PDB 1HYN). Panel A displays the best structure (red) from the C_β distance filtering method (C_α RMSD ~ 4.0 Å). Panel B shows the best structure (blue) from the explicit spin-label rotamer sampling method (C_α RMSD ~ 2.2 Å).

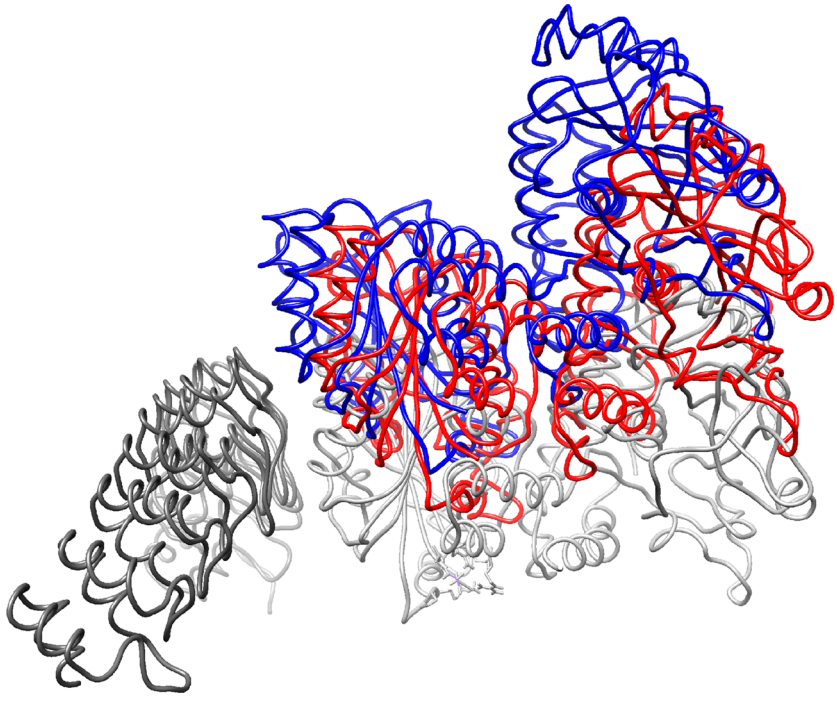


Figure 6. CDB3 dimers docked to ankyrin. Three structures are compared, with the ankyrin subunits (dark gray) superimposed. In gray, we show the docking result that is closest to the mean from a previously published RosettaDock calculation after applying a nonautomated filtering procedure.[10] In red, we show the closest to mean TagDock result from the rigid C_β algorithm. In blue, we show the TagDock result using the MTSSL rotamer library. In all cases, we used the 20 experimental distances reported previously[10] to filter our docking pose solutions.

explicit spin-label rotamer method, where lower scores indicate better agreement with the experimental distance data.

We also generated CDB3-ankyrin complex structures using subsets of the DEER distance data, to assess solution "convergence" as a function of total interresidue distances used in the filtering step. Each spin-label position in one partner that had distances measured to multiple positions in the other partner became a restraint set; e.g., CDB3 70−ANK 707 and CDB3 70−ANK 722 is one distance subset of the full data set. Table 1 shows the C_α RMSD for the structure closest to the geometric mean of

dx.doi.org/10.1021/jp4099705 | J. Phys. Chem. B 2014, 118, 4717−4726

Table 1. Pairwise C_α RMSD Results (Å) for Subsets of Our Full Distance Restraint Data Set, with and without Solvent Accessible Surface Area (SASA) Filtering[a]

position	no. of restraints	distance	distance + SA-SA
CDB3 70	2	17.3 (18.5)	18.5 (14.2)
CDB3 130	5	2.9 (17.8)	6.2 (13.4)
CDB3 166	4	19.9 (15.7)	18.8 (14.0)
CDB3 254	3	16.4 (15.2)	10 (13.6)
CDB3 302	2	24.4 (23.4)	9.9 (14.0)
ANK 509	2	23.9 (19.2)	18.5 (16.8)
ANK 524	3	27.3 (18.9)	10.8 (11.4)
ANK 608	3	3.7 (15.5)	3.9 (14.9)
ANK 623	4	20.1 (18.1)	11.1 (14.3)
ANK 707	2	12.5 (18.5)	3 (15.3)
ANK 722	2	16 (19.8)	9 (13.0)

[a]The closest structure to the mean of each ensemble is compared to the reference structure obtained by refining the CDB3-ankyrin structure with all available restraints. The average pairwise RMSD for each ensemble is shown in parentheses. For reference, the pairwise C_α RMSD for the ensemble generated when all experimental distances are used is ~0.4 Å. In most cases, reducing the total number of distances used for the filtering step yields docking poses that deviate significantly from the results obtained when all 20 distances are used for filtering, and the structural variation among the top 200 docking poses is considerable. When SASA data[10] are included in the filtering step (column 4), the docking poses obtained are generally much closer to the solution generated with the full 20-distance restraint set, illustrating the potential utility of these data to help select the best docking models.

the top 200 refined TagDock poses computed with various distance data subsets, compared to the structure closest to the geometric mean of the top 200 refined TagDock poses generated when all experimental distances are used for solution filtering. In some cases, a distance subset contains as few as two interresidue distances for filtering, so it is not surprising that the C_α RMSD values for these subsets are quite large. For the limited-distance data subsets only, we also computed spin-label solvent accessibility for each docking pose and compared these results to the experimental accessibility measurements[10] as an additional filter, to test if solvent accessible surface area data can be used effectively to improve the filtering process. When we used the solvent accessibility filter together with limited-distance data subsets to eliminate docking poses that were inconsistent with the experimental data, we observed improved C_α RMSD values for the mean structures in most cases, as well as dramatically smaller pairwise C_α RMSD values for all remaining poses that passed both the distance and accessibility filters (i.e., the docking pose solution set is more tightly clustered about a mean structure), suggesting solution convergence within the limits of the available experimental data.

DISCUSSION

We emphasize that the goal of our refinement procedure is not to precisely match a given distance, nor should it be. The results obtained in the EPR experiments are interpreted as Gaussian distributions of the distances from an ensemble of structures, and this fact greatly influenced our choice of a flat-bottomed harmonic scoring function. This function does not "penalize" any structure that falls within the Gaussian distribution, unlike a simple harmonic function that favors the mean value of the distribution. We do not expect a single structure to simultaneously match the center of the Gaussians for all 20 distances measured for the ankyrin-CDB3 complex, but we do expect that a small ensemble of structures will reproduce the experimental distance distributions, as has been reported in a previous study of a synaptotagmin-SNARE complex by Lai et al.[35] In our previous study with a large database of protein heterodimer complexes, we rarely observed any docking models with low penalty scores that exhibit any notable distance violations.[9] In only a few cases have we obtained a docking solution with a good penalty score, i.e., excellent overall agreement with experimental distance data, which nonetheless displayed one pair distance that deviated significantly from the experimental value. For one of these cases, the pair-distance discrepancy was due to an experimental error, and the experimental result was subsequently retracted. In the remaining cases, we strongly suspect the isolated pair-distance discrepancies are likely due to our rigid backbone docking protocol. Proteins are inherently flexible, and at a minimum, minor structural fluctuations are to be expected. Our earlier studies with a large data set of protein heterodimer complexes suggests that we can generally obtain good docking models without explicit inclusion of backbone flexibility.[9] However, in cases where large conformational changes occur during complex formation, it will likely be important to include the impact of local backbone flexibility on spin-label distance distributions to achieve better agreement with EPR DEER experiments. It was not necessary to include backbone flexibility for the CDB3 homodimer docking calculations reported here because of the computational protocol we used to generate the isolated CDB3 monomer structures. The large size and inherent flexibility of the MTSSL spin labels can also produce local structural fluctuations that could easily alter simulated distances by several angstroms, and this was our primary motivation to incorporate an MTSSL rotamer library for explicit side-chain conformational sampling in the TagDock toolkit.

Our results for the CDB3 homodimer show clearly that explicit consideration of MTSSL side-chain conformations during the solution filtering step can yield much better docking poses than a simple C_β–C_β distance assessment, in this case, CDB3 homodimer models that are considerably closer to the reference crystal structure. We emphasize that the MTSSL rotamer library must be coupled with an efficient method to repack protein side chains adjacent to the spin-label site, and we use the SCRWL application for this task. At some residue positions, we do not identify any allowable spin-label side-chain rotamers until we have performed extensive repacking for all neighboring amino acid side chains. Failure to identify and consider all possible spin-label rotamers at each labeled site will almost certainly reduce the number of candidate docking poses that pass the experimental distance distribution filtering step, and lead to the inappropriate exclusion of viable docking pose solutions.

The TagDock toolkit is not intrinsically limited to experimental distance measurements for solution filtering.[9] Other metrics, such as side-chain solvent accessible surface area, can be used for the filtering step as well. Our ankyrin:CDB3 docking results, using limited distance measurements combined with spin-label solvent accessibility measurements, illustrate the potential power and effectiveness of combined filtering metrics. When solvent accessibility data are used together with limited distance measurements for filtering, we achieve in many cases a docking solution ensemble that is closer to the result obtained with the full distance restraint data set than we could achieve with limited distance data alone. This result is significant, because

dx.doi.org/10.1021/jp4099705 | J. Phys. Chem. B 2014, 118, 4717−4726

distance measurement based on site-directed spin labeling is a labor-intensive procedure, and any strategy that might yield comparably good docking results with fewer experimental measurements has obvious appeal. Of course, solvent accessibility is a useful filtering metric only when the specific residue positions are impacted by the protein docking interface. Otherwise, the solvent accessibility profiles provide little or no additional useful information.

The refined ankyrin-R:CBD3 complex reported here exhibits much better agreement with the DEER distance and spin-label accessibility measurements, compared to our earlier model.[10] These improved results are directly attributable to the TagDock toolkit design and implementation. Since TagDock calculations are fully automated and require no user intervention for model generation or scoring, we avoid potential biases that might be introduced with any manual model building or model selection procedures, unlike our previously published work. Furthermore, since TagDock exhaustively considers all geometrically possible docking poses during the initial docking step, we are far more likely to identify all solutions that satisfy the experimental restraints. As reported previously,[9] solutions sets typically converge to a well-defined consensus structure as additional experimental restraints are added in the filtering step, providing a convenient metric to determine the optimal solution (within the limits of the experimental data used for the filtering process).

An alternate model reported previously for the ankyrin-CDB3 complex[36] places CDB3 more directly along the concave surface region of the ankyrin-R fragment. While our automated docking protocol sampled this region extensively, it did not identify any docking poses in the region that provide good agreement with the DEER distances or solvent accessibility measurements. We should note that our previous RosettaDock calculations,[10] when performed without distance restraints, did generate docking poses in the ankyrin-R concave surface region that scored well, but these poses failed to satisfy the DEER distance restraints.

CONCLUSION

Using a combination of EPR distance measurements and rigid docking simulations, we have defined a fast and efficient method for determining and refining the structure of a complex between two proteins based on EPR distance data. For our specific test cases, we studied the CDB3 homodimer and the complex between CDB3 and ankyrin-R. We show that a small distance restraint data set such as that used in this study is sufficient to reliably dock two rigid proteins. Furthermore, sampling spin-label conformations with our new custom MTSSL rotamer library improves agreement with the crystal structure of the CDB3 homodimer complex, compared to calculations based on a simple C_β–C_β distance approximation. In cases where one or more of the individual protein partners undergoes significant conformational changes, this strategy will likely not be sufficient to refine an atomic resolution model. In these cases, a simple rigid body docking strategy to generate initial model candidates will likely be inadequate, and efficient strategies for flexible protein docking calculations will be needed. We are currently evaluating several computational schemes for this purpose, using structurally well-characterized protein–protein complexes as test systems. Local protein backbone conformational sampling will be crucial for these situations, and we are testing several modified MD simulation methods that are more computationally efficient than traditional equilibrium MD simulations, such as the accelerated MD method of Hamelberg et al.[37]

There are of course many good protein docking programs available now, as we reviewed in a previous publication.[9] However, there are only a few protein docking programs or tools that model spin-label side chains explicitly when utilizing EPR data. RosettaEPR[38] has this capability, as does the MtsslSuite[39] available in the PyMol graphics package.[40] However, the spin-label rotamer libraries employed in these tools are derived primarily from existing crystal structures. At the moment, there are not enough crystal structures with MTSSL side chains to derive rotamer libraries that yield the same statistical confidence levels as, e.g., the SCRWL amino acid rotamer libraries.[27] As discussed above, this was our primary motivation to develop a Boltzmann-weighted MTSSL rotamer library from extensive simulation data. The MMM software package[22] is the only other tool that utilizes an MTSSL rotamer library similar to ours and that can incorporate explicit spin-label rotamer sampling along with sampling of neighboring side-chain rotamers, in protein docking calculations. However, TagDock differs substantially from all other protein docking programs in that it incorporates automated distance-difference matrix analysis to analyze results and suggest additional experimental measurements.[9,41] This capability, coupled with the computational performance (TagDock can generate $\sim 1 \times 10^6$ docking poses in 1 h on a single-processor Linux workstation), makes TagDock particularly useful for rapid docking model generation and assessment.[9] For these reasons, we promote the package primarily as a tool to facilitate experiment planning, rather than a replacement for other capable protein docking programs. Of course, in cases where the rigid body docking protocol is reasonable (i.e., cases where the individual monomers do not undergo significant conformational changes during complex formation), TagDock is also an effective and efficient protein docking program.[9] The software extensions reported here that enable us to better utilize EPR distance data further enhance TagDock's utility as an experiment planning tool.

The results presented in this work and in an earlier publication[10] for the ankyrin-CDB3 complex demonstrate clearly that detailed, chemically reasonable three-dimensional structural models can be generated for protein–protein complexes using a small number of accurate, long-range distance measurements and a relatively simple, inexpensive computational refinement strategy, provided there is sufficient information to define the internal three-dimensional structures for the individual protein partners used as starting structures for the automated docking calculations. The experimental data used in this study for solution filtering (DEER distance and spin-label solvent accessibility measurements) can be generated for many other protein complexes, and the TagDock toolkit can be used for any oligomer docking project where reliable three-dimensional structures are available for the individual monomers, so the procedures and strategy presented here should be widely applicable for many biomolecular oligomer structure refinement problems.

ASSOCIATED CONTENT

Supporting Information

Tables listing full statistics for all MTSSL side-chain rotamer conformations. This material is available free of charge via the Internet at http://pubs.acs.org.

dx.doi.org/10.1021/jp4099705 | J. Phys. Chem. B 2014, 118, 4717–4726

AUTHOR INFORMATION

Corresponding Author

*E-mail: terry.p.lybrand@vanderbilt.edu. Tel.: (615)-343-1247. Fax: (615)-936-2211.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported in part by National Institutes of Health Grant P01 GM080513 (T.P.L.). The TagDock software, user's manual, and examples are available for download at http://tagdock.vueinnovations.com.

ABBREVIATIONS

CDB3: cytoplasmic domain of erythrocyte band 3; EPR: electron paramagnetic resonance; DEER: double electron−electron resonance; MTSSL: methanethiosulfonate spin label; FRET: fluorescence resonance energy transfer

REFERENCES

(1) Borbat, P. P.; Davis, J. H.; Butcher, S. E.; Freed, J. H. Measurement of Large Distances in Biomolecules Using Double-quantum Filtered Refocused Electron Spin-echoes. *J. Am. Chem. Soc.* **2004**, *126*, 7746−7747.

(2) Borbat, P. P.; Georgieva, E. R.; Freed, J. H. Improved Sensitivity for Long-Distance Measurements in Biomolecules: Five-Pulse Double Electron-Electron Resonance. *J. Phys. Chem. Lett.* **2013**, *4*, 170−175.

(3) Hustedt, E. J.; Beth, A. H. Nitroxide Spin-Spin Interactions: Applications to Protein Structure and Dynamics. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 129−153.

(4) Hustedt, E. J.; Smirnov, A. I.; Laub, C. F.; Cobb, C. E.; Beth, A. H. Molecular Distances from Dipolar Coupled Spin-Labels: The Global Analysis of Multifrequency Continuous Wave Electron Paramagnetic Resonance Data. *Biophys. J.* **1997**, *72*, 1861−1877.

(5) Wen, P.-C.; Verhalen, B.; Wilkens, S.; McHaourab, H. S.; Tajkhorshid, E. On the Origin of Large Flexibility of P-glycoprotein in the Inward-Facing State. *J. Biol. Chem.* **2013**, *288*, 19211−19220.

(6) Brandon, S.; Beth, A. H.; Hustedt, E. J. The Global Analysis of DEER Data. *J. Magn. Reson.* **2012**, *218*, 93−104.

(7) Borbat, P. P.; McHaourab, H. S.; Freed, J. H. Protein Structure Determination Using Long-Distance Constraints from Double-Quantum Coherence ESR: Study of T4 Lysozyme. *J. Am. Chem. Soc.* **2002**, *124*, 5304−5314.

(8) Hustedt, E. J.; Stein, R. A.; Sethaphong, L.; Brandon, S.; Zhou, Z.; DeSensi, S. C. Dipolar Coupling Between Nitroxide Spin Labels: The Development and Application of a Tether-in-a-Cone Model. *Biophys. J.* **2006**, *90*, 340−356.

(9) Smith, J. A.; Edwards, S. J.; Moth, C. W.; Lybrand, T. P. TagDock: An Efficient Rigid Body Docking Algorithim for Oligomeric Protein Complex Model Construction and Experiment Planning. *Biochemistry* **2013**, *52*, 5577−5584.

(10) Kim, S.; Brandon, S.; Zhou, Z.; Cobb, C. E.; Edwards, S. J.; Moth, C. W.; Parry, C. S.; Smith, J. A.; Lybrand, T. P.; Hustedt, E. J.; et al. Determination of Structural Models of the Complex Between the Cytoplasmic Domain of Erythrocyte Band 3 and Ankyrin-R Repeats 13−24. *J. Biol. Chem.* **2011**, *286*, 20746−20757.

(11) Bennett, V.; Baines, A. J. Spectrin and Ankyrin-Based Pathways: Metazoan Inventions for Integrating Cells into Tissues. *Physiol. Rev.* **2001**, *81*, 1353−1392.

(12) Nilsson, K. R.; Bennett, V. Ankyrin-Based Patterning of Membrane Microdomains: New Insights into a Novel Class of Cardiovascular Diseases. *J. Cardiovas. Pharmacol.* **2009**, *54*, 106−115.

(13) Cunha, S. R.; Mohler, P. J. Ankyrin-Based Cellular Pathways for Cardiac Ion Channel and Transporter Targeting and Regulation. *Semin. Cell Dev. Biol.* **2011**, *22*, 166−170.

(14) Che, A.; Morrison, I. E.; Pan, R.; Cherry, R. J. Restriction by Ankyrin of Band 3 Rotational Mobility in Human Erythrocyte Membranes and Reconstituted Lipid Vesicles. *Biochemistry* **1997**, *36*, 9588−9595.

(15) Thevenin, B. J.-M.; Low, P. S. Kinetics and Regulation of the Ankyrin-Band 3 Interaction of the Human Red Blood Cell Membrane. *J. Biol. Chem.* **1990**, *265*, 16166−16172.

(16) Blackman, S. M.; Piston, D. W.; Beth, A. H. Oligomeric State of Human Erythrocyte Band 3 Measured by Fluorescence Resonance Energy Homotransfer. *Biophys. J.* **1998**, *75*, 1117−1130.

(17) Casey, J. R.; Reithmeier, R. A. Analysis of the Oligomeric State of Band 3, the Anion Transport Protein of the Human Erythrocyte Membrane, by Size Exclusion High Performance Liquid Chromatography: Oligomeric Stability and Origin of Heterogeneity. *J. Biol. Chem.* **1991**, *266*, 15726−15737.

(18) Hanspal, M.; Golan, D. E.; Smockova, Y.; Yi, S. J.; Cho, M. R.; Liu, S. C.; Palek, J. Temporal Synthesis of Band 3 Oligomers During Terminal Maturation of Mouse Erythroblasts. Dimers and Tetramers Exist in the Membrane as Preformed Stable Species. *Blood* **1998**, *92*, 329−338.

(19) Delaunay, J. Molecular Basis of Red Cell Membrane Disorders. *Acta Haematol.* **2002**, *108*, 210−218.

(20) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; et al. *AMBER 10*; University of California: San Francisco, CA, USA, 2008.

(21) Dunbrack, R. L.; Cohen, F. E. Bayesian Statistical Analysis of Protein Side-Chain Rotamer Preferences. *Protein Sci.* **1997**, *6*, 1661−1681.

(22) Polyhach, Y.; Bordignon, E.; Jeschke, G. Rotamer Libraries of Spin Labelled Cysteines for Protein Studies. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2356−2366.

(23) Roux, B.; Islam, S. M. Restrained-Ensemble Molecular Dynamics Simulations Based on Distance Histograms from Double Electron−Electron Resonance Spectroscopy. *J. Phys. Chem. B* **2013**, *117*, 4733−4739.

(24) Islam, S. M.; Stein, R. A.; McHaourab, H. S.; Roux, B. Structural Refinement from Restrained-Ensemble Simulations Based on EPR/DEER Data: Application to T4 Lysozyme. *J. Phys. Chem. B* **2013**, *117*, 4740−4754.

(25) DeSensi, S. C.; Rangel, D. P.; Beth, A. H.; Lybrand, T. P.; Hustedt, E. J. Simulation of Nitroxide Electron Paramagnetic Resonance Spectra from Brownian Trajectories and Molecular Dynamics Simulations. *Biophys. J.* **2008**, *94*, 3798−3809.

(26) Bondi, A. van der Walls Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.

(27) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L. Improved Prediction of Protein Side-Chain Conformations with SCWRL4. *Proteins* **2009**, *77*, 778−795.

(28) Bruenger, A.; Karplus, M. Molecular Dynamics Simulations with Experimental Restraints. *Acc. Chem. Res.* **1991**, *24*, 54−61.

(29) Nilges, M.; Clore, G. M.; Gronenborn, A. M. Determination of Three-Dimensional Structures of Proteins from Interproton Distance Data by Dynamical Simulated Annealing from a Random Array of Atoms: Circumventing Problems Associated with Folding. *FEBS Lett.* **1988**, *239*, 129−136.

(30) Moore, J.; Case, D.; Chazin, W.; Gippert, G. Three-Dimensional Solution Structure of Plastocyanin from the Green Alga Scenedesmus Obliquus. *Science* **1988**, *240*, 314−317.

(31) Sanner, M. F.; Olsen, A. J.; Spehner, J.-C. Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers* **1996**, *38*, 305−320.

(32) Zhou, Z.; DeSensi, S. C.; Stein, R. A.; Brandon, S.; Dixit, M.; McArdle, E. J.; Warren, E. M.; Kroh, H. K.; Song, L.; Cobb, C. E.; et al. Solution Structure of the Cytoplasmic Domain of Erythrocyte Membrane Band 3 Determined by Site-Directed Spin Labeling. *Biochemistry* **2005**, *44*, 15115−15128.

(33) Michaely, P.; Tomchick, D. R.; Machius, M.; Anderson, R. G. W. Crystal Structure of a 12 ANK Repeat Stack from Human Ankyrin R. *EMBO J.* **2002**, *21*, 6387−6396.
(34) Zhang, D.; Kiyatkin, A.; Bolin, J. T.; Low, P. S. Crystallographic Structure and Functional Interpretation of the Cytoplasmic Domain of Erythrocyte Membrane Band 3. *Blood* **2000**, *96*, 2925−2933.
(35) Lai, A. L.; Huang, H.; Herrick, D. Z.; Epp, N.; Cafiso, D. S. Synaptotagmin 1 and SNAREs Form a Complex That Is Structurally Heterogeneous. *J. Mol. Biol.* **2011**, *405*, 696−706.
(36) Grey, J. L.; Kodippili, G. C.; Simon, K.; Low, P. S. Identification of Contact Sites Between Ankyrin and Band 3 in the Human Erythrocyte Membrane. *Biochemistry* **2012**, *51*, 6838−6846.
(37) Hamelberg, D.; de Oliveira, C. A. F.; McCammon, J. A. Sampling of Slow Diffusive Conformational Transitions with Accelerated Molecular Dynamics. *J. Chem. Phys.* **2007**, *127*, 155102.
(38) Alexander, N. S.; Stein, R. A.; Koteiche, H. A.; Kaufmann, K. W.; McHaourab, H. S.; Meiler, J. RosettaEPR: Rotamer Library for Spin Label Structure and Dynamics. *PLoS One* **2013**, *8*, e72851.
(39) Hageleukin, G.; Abdullin, D.; Ward, R.; Schiemann, O. In Silico Spin Labeling, Trilateration and Distance-Constrained Rigid Body Docking in PyMol. *Mol. Phys.* **2013**, *111*, 2757−2766.
(40) *The PyMol Molecular Graphics System*, Version 1.7.0.0; Schrödinger: 2014.
(41) Akke, M.; Forsén, S.; Chazin, W. J. Solution Structure of $(Cad^{2+})_1$-calbinden D_{9k} Reveals Details of the Stepwise Structural Changes along the Apo→$(Ca^{2+})_1^{II}$→$(Ca^{2+})_2^{I,II}$ Binding Pathway. *J. Mol. Biol.* **1995**, *252*, 102−121.