Specifications for predicting the structure of the two-domain protein ZLBT-C

Terry Oas[*,1,2,], Aulane Mpouli[2], Edward Cheng[2], and Bruce Donald[*,1,2,3]

[1]Department of Biochemistry, Duke University
[2]Department of Chemistry, Duke University
[3]Department of Computer Science, Duke University
[*]Contacts: oas@duke.edu, brd@cs.duke.edu

May, 2024

ZLBT-C is a mimic of two of the five nearly identical three-helix bundle domains from the N-terminal region of staphylococcal protein A (Figure 1)[1]. ZLBT is a biotech variant of the B domain of protein A with a lanthanide binding tag inserted between helices 2 and 3. The C domain is linked to ZLBT via the 6-residue wild-type B-C linker, KADNKF. The structures of the helical cores of both ZLBT[2] and C domains[3] have been determined, so the challenge outlined below is not to predict these structures, but rather to predict the range of structures that position the two domains relative to each other. The published description of this structure, based on experimental NMR residual dipolar coupling (RDC) data, is a continuous distribution of interdomain orientation (CDIO)(see Figure 2)[1]. Predictions could be made in the form of a 3D (Bingham) probability distribution over the space of the relative orientations of the two domains (SO(3)) or as an ensemble of population-weighted structures. In both cases, to compare predictions with experimental data, it is necessary to define domain-fixed Cartesian coordinate systems based on atomic coordinates of the core domain structures. For this reason, the predicted helical core structures must match the deposited structures within a minimal RMSD.

The predictions will be compared with NMR RDC data and small-angle X-ray scattering (SAXS) profiles. In this way, both the CDIO and the interdomain distance distribution of a prediction can be compared with experimental data. The requirements for such predictions are the following.

1. The backbone RMSD of residues 36-50, 70-85 (ZLBT helix 2/3 core, 2LR2, Model 1[2]) and 112-125, 129-143 (C helix 2/3 core, 4NPD, Alternates A[3]) of ZLBT-C should fall within 0.5 Å of each deposited structure (see Figure 1).
2. If the prediction takes the form of an ensemble, the population of each ensemble member must be given as a positive rational number and must sum to 1.0. The uncertainty should be provided for the population of each ensemble member. Coordinate files should be in PDB format.
3. If the prediction takes the form of a continuous distribution[1], the quaternion mean, variances, and covariances representing each Bingham distribution mode (if more than one) must be given, along with the relative probability of each mode. The Cartesian coordinate system of each domain used to define these quaternions should match the ones defined in the attached algorithm.
4. A graphical representation of the ZLBT-C CDIO has been published(Figure 2)[1], but not the quantitative properties described in 3) nor the interdomain distance distribution, so this remains a predictive challenge.
5. Unpublished RDC and SAXS data will be available for a ZLBT-C construct with a Gly_6 linker that replaces the wild-type linker. Predictions should be made for both the wild-type sequence and this one.
6. Predictions will be evaluated based on a comparison between the experimental and predicted CDIOs using the free energy function described in Qi, et al.[1] and the χ^2 values of the predicted vs. observed SAXS profiles.

References

[1] Y. Qi, J. W. Martin, A. W. Barb, F. Thélot, A. K. Yan, B. R. Donald, and T. G. Oas, "Continuous interdomain orientation distributions reveal components of binding thermodynamics," *Journal of Molecular Biology*, vol. 430, pp. 3412–3426, 9 2018.

[2] A. W. Barb, T. G. Ho, H. Flanagan-Steet, and J. H. Prestegard, "Lanthanide binding and IgG affinity construct: potential applications in solution NMR, MRI, and luminescence microscopy.," *Protein Sci*, vol. 21, pp. 1456–1466, Oct 2012.

[3] L. N. Deis, C. W. Pemble, Y. Qi, A. Hagarman, D. C. Richardson, J. S. Richardson, and T. G. Oas, "Multiscale conformational heterogeneity in staphylococcal protein A: Possible determinant of functional plasticity," *Structure*, vol. 22, pp. 1467–1477, 10 2014.

[4] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun, "Structural characterization of flexible proteins using small-angle x-ray scattering," *Journal of the American Chemical Society*, vol. 129, no. 17, pp. 5656–5664, 2007.

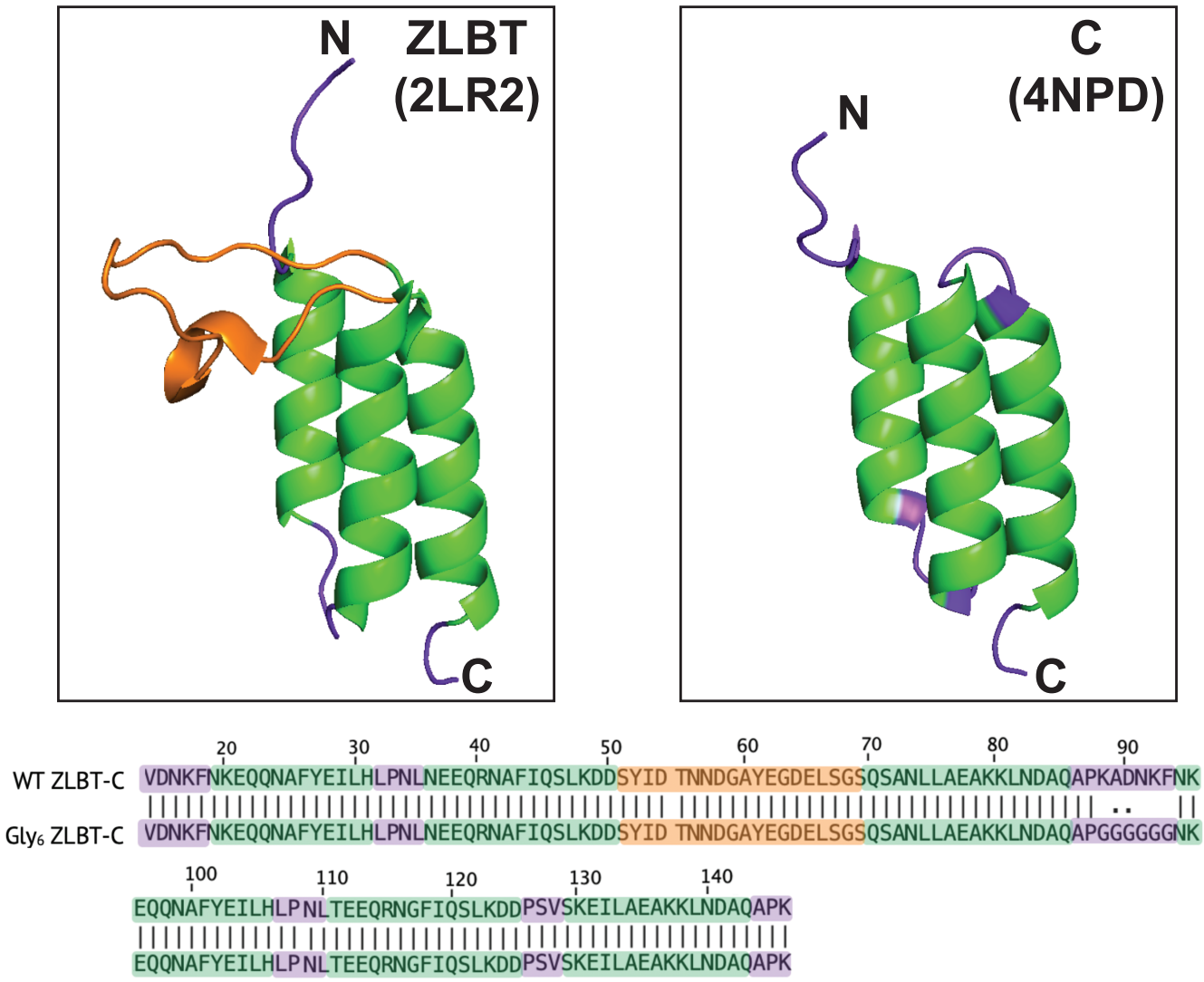

Figure 1: Structures of the ZLBT (2LR2, Model 1[2]) and C (4NPD, Alternates A[3]) domains of ZLBT-C, a two-domain protein from staphylococcal protein A[1]. In ZLBT-C, the C-terminus of the ZLBT domain is connected to the N-terminus of the C domain via a linker comprising one C-terminal residue of ZLBT (K_{88}) and five N-terminal residues of C (A_{89}-F_{93}). Color coding is as follows: Green = helical cores; Purple = termini, linker and inter-helical loops; Salmon = lanthanide binding tag (LBT). The sequences of the wild-type construct and a second with GGGGGG substituted for the KADNKF linker are shown below the structures. A_{86}, P_{87}, and N_{94} are not helical but are also not considered part of the flexible linker because they form helix caps. NOTE: The residue numbers correspond to those given in 2LR2, followed by the residues in 4NPD. To convert from the residue numbers shown in this figure to those in 4NPD, subtract 88.

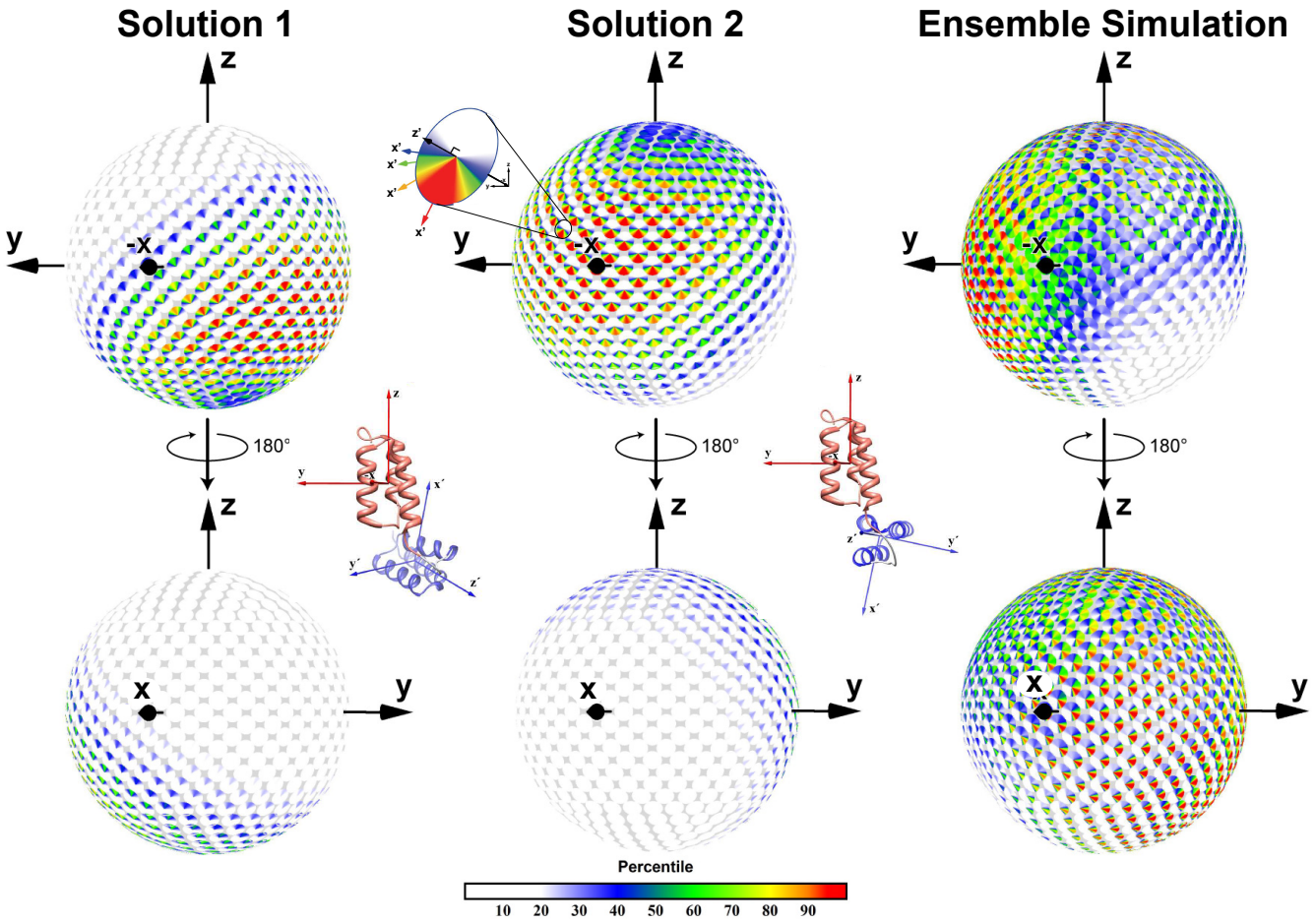

Figure 2: Disk-on-Sphere representations of two degenerate solutions to the observed NMR RDC values for ZLBT-C. The two solutions are equally probable because there is a bipolar degeneracy of the RDC observations[1]. Color-coded probabilities of the C domain's x'-axis orientation are depicted as disks whose position on the sphere corresponding to the ZLBT coordinate frame (x,y,z) represents the z'-axis of the C domain. The ribbon drawing structures represent the most probable interdomain orientation for each solution. The Ensemble Simulation is a kernalized representation of a ZLBT-C ensemble using the RanCh program[4]. The RanCh simulation suggests that Solution 2 is the correct solution because the high-probability orientations of Solution 1 are predicted to be infeasible in the RanCh simulation, likely due to steric clashes. However, predictions will be compared with both solutions, using the free energy function described by Qi, et al.[1].