Solving the structure of Lgl2, a difficult blind test of unsupervised structure determination

Ivan S. Ufimtsev[a], Lior Almagor[a,b], William I. Weis[a,b,1], and Michael Levitt[a,1]

[a]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305; and [b]Department of Molecular & Cellular Physiology, Stanford University School of Medicine, Stanford, CA 94305

In the companion paper by Ufimtsev and Levitt [Ufimtsev IS, Levitt M (2019) *Proc Natl Acad Sci USA*, 10.1073/pnas.1821512116], we presented a method for unsupervised solution of protein crystal structures and demonstrated its utility by solving several test cases of known structure in the 2.9- to 3.45-Å resolution range. Here we apply this method to solve the crystal structure of a 966-amino acid construct of human lethal giant larvae protein (Lgl2) that resisted years of structure determination efforts, at 3.2-Å resolution. The structure was determined starting with a molecular replacement (MR) model identified by unsupervised refinement of a pool of 50 candidate MR models. This initial model had 2.8-Å RMSD from the solution. The solved structure was validated by comparison with a model subsequently derived from an alternative crystal form diffracting to higher resolution. This model could phase an anomalous difference Fourier map from an Hg derivative, and a single-wavelength anomalous dispersion phased density map made from these sites aligned with the refined structure.

Lgl | unsupervised | crystal structure

Lgl proteins are essential for establishing and maintaining epithelial cell polarity (1, 2). Humans express two Lgl isoforms, Lgl1 and Lgl2; Lgl2 is a 1,020-amino acid protein containing 14 predicted WD40 repeats. Lgl proteins interact with several proteins involved in epithelial polarity, including atypical protein kinase C (aPKC), which phosphorylates a serine-riched loop in Lgl and thereby regulates its interaction with other polarity proteins as well as lipid membranes (3, 4). Budding yeast express an Lgl homolog, Sro7, whose structure and function have been characterized (5). The structure of Sro7 comprises two β-propeller barrels, each containing 7 WD40 repeats. Lgl2 and Sro7 possess distant sequence homology as predicted by the Hidden Markov model profile alignment (15% identity, 0.208 similarity) (6, 7). A subsequent Lgl2–Sro7 structural alignment revealed a smaller 10% identity (*SI Appendix*, Fig. S1) (8).

Here we describe determination of the Lgl2 crystal structure. Both conventional molecular replacement (MR) using the Sro7 model and experimental phase determination proved exceptionally challenging for these crystals. Thus, we applied our method for unsupervised solution of protein structures that had been developed using smaller proteins of known structure as test cases. We describe application of the method and validation of the solution with experimental phases.

Results

The Lgl2 construct used for crystallization spans residues 13 to 978, and excludes disordered regions at the N and C termini of the full-length protein that were predicted by a multiple sequence alignment of Lgl homologs with structured regions of Sro7. Crystals of the unphosphorylated Lgl2(residues 13 to 978) were obtained in two forms: (*i*) Crystal form 1 was in the space group P2$_1$2$_1$2$_1$ and diffracted to 3.2 Å; (*ii*) crystal form 2 was in the space group C2 and diffracted anisotropically with a maximum limit of 2.2 Å (*SI Appendix*, Table S1). More details on protein expression, purification, crystallization, data collection, and processing are provided in the accompanying paper (9).

Given the homology to Sro7, MR phasing with the Sro7 model was first attempted, but consistently failed, which was not surprising given the low sequence identity between these proteins. Experimental phasing was very problematic in both the crystal forms. Attempts to crystallize a seleno-methionine (Se-Met) incorporated protein to obtain form 1 crystals were unsuccessful, and Se-Met form 2 crystals grew rarely and gave very weak diffraction. A derivative was obtained with the mercurial thiomersal in the form 2 crystals, but these proved to be very fragile and could withstand only short soak durations, which resulted in poor incorporation of the mercury atoms and weak anomalous dispersion signals. Moreover, thiomersal-soaked crystals were nonisomorphous with the native crystals, which prevented determination of the mercury substructure and phasing by isomorphous replacement.

The structure was solved in crystal form 1 using the phasing method described in detail in the accompanying paper (10). We use a statistical approach to compute phase restraints that, together with the observed structure factor amplitudes F$_{obs}$, are used for structure refinement (i.e., we refine against complex-valued structure factors). As with any refinement method, our algorithm proceeds through a number of iterations. In each cycle, the algorithm builds a fixed number of trial electron density maps (here 50 maps) generated by our density modification pipeline based on the current model. The pipeline is seeded with phases of the model and random amplitudes scaled by $\exp(-B_0 s^2)$, where B_0 and s are the overall B factor and the magnitude of the reciprocal

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Fig. 1. Statistics of the Lgl2 models progressively constructed in the solution process started with the best MR model found automatically by our unsupervised refinement pipeline in the pool of 50 MR solutions. Blue, pink, and brown bars represent the relative number of atoms, residues, and sequence identities (structurally aligned identical amino acids) with respect to the final solution. Black line denotes RMSD from the solution plotted on a different scale. The "auto_x" models were built in Coot and refined by our unsupervised refinement code for ∼60 macrocycles each. Structure "auto_9" had phases of good quality, sufficient for standard manual processing. This model was finalized manually with Coot and phenix.refine in 40 build–refine iterations. The dashed lines represent summarized statistics of these runs.

vector, thus producing a different density map each time. It proceeds through 30 iterations of automatic density modification with introduction of F_{obs} and low-resolution phases of the model. Fifty trial models are then automatically built into these electron density maps by Buccaneer (11, 12), refined by Refmac (13), and ranked according to their R-free factors (14). Next, we combine the 20 best trial models by averaging their figure-of-merit-weighted (FOM-weighted) (15, 16) density maps, and use these to derive new phase estimates for restraints in the next refinement macrocycle. The model is then refined with the new complex-value structure factor restraints and the previous trial models are discarded.

Building many trial structures at each refinement macrocycle is computationally expensive, but is a crucial step in our algorithm: It enables the refinement process to navigate over plateaus and escape local minima in the R-free hypersurface. This step is also highly parallelizable and therefore benefits from the availability of compute clusters. Unlike many state-of-the-art approaches that directly optimize maximum likelihood target functions (17–19), our method, thanks to the use of additional phase restraints, deforms the model toward more interpretable parts of the density map. It is therefore driven, to a large extent, by interpretability of the density map rather than by F_{obs}, although, as discussed above, the R-free values are used indirectly to rank the trial structures used to derive the phase restraints.

We started the solution process by building a single search model with Modeler (20) based on an Sro7–Lgl2 sequence alignment computed by HHpred (6) and the Sro7 template. After removing loops with more than 4 amino acids from the model, the N-terminal and C-terminal barrels including side chains were placed independently by Phaser MR (21). Using the placed model and the structure of Sro7 as multiple templates, we built a nearly full-length model of Lgl2 by Modeler, and processed the model by our unsupervised refinement pipeline. Not surprisingly, this attempt failed to improve the initial model in any way.

Fig. 2. (*Left*) The initial (cyan) and final (silver) structures of Lgl2 protein. (*Right*) Evolution of R-free of the parent model. The parent model was manually rebuilt at every ∼60th macrocycle. The total number of trial structures built in this run is 27,500.
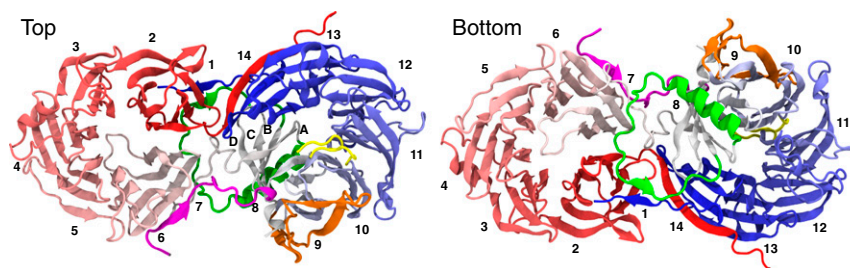
Fig. 3. Structure of human Lgl2 crystal form 1. The N-terminal barrel is colored in red-to-white and the C-terminal barrel is colored in white-to-blue colors. Large insertion elements in the C-terminal barrel are colored in green (8A–8B), magenta (8D–9A), orange (9C–9D), and yellow (10D–11A).

To find the initial structure that eventually led to the solution, we diversified the Sro7−Lgl2 sequence alignment by manually realigning certain segments that were poorly aligned by the automatic procedure, which produced 10 different sequence alignments in total. In one case, a poorly aligned segment was intentionally misaligned with insertions and deletions to prevent any alignment for homology modeling. In another case, insertions in the segment were removed to provide a better-aligned substructure. Fifty search models were built by Modeler (10 alignments, five models per alignment), and each barrel was placed independently by Phaser. This procedure did not yield a single good MR solution, but instead produced a set of models with translation function z-scores in the 6.0 to 8.0 range. Fifty nearly full-length Lgl2 structures were built by Modeler using the placed models and the structure of Sro7 as multiple templates. All of the structures were processed by our unsupervised refinement code for 80 macrocycles in parallel on the Xstream cluster at Stanford University. A total of 200,000 trial structures was built in this screening run (50 models × 80 macrocycles × 50 trial structures per macrocycle). Forty-nine of the models showed no progress in terms of R-free improvement, while one demonstrated good R-free dynamics and was chosen for further processing.

This selected structure contained essentially the full sequence and had a 2.8-Å Cα-RMSD from the subsequently solved structure. Interestingly, this initial RMSD was very close to the 2.9-Å "convergence radius" of our method estimated previously for a similar resolution range (10). Given the relatively large initial RMSD of the model, it is not surprising that only 12% of the side chains fitted the electron density map reasonably well, and only these were left in the model. The rest of the side chains were manually converted to alanines (the second brown bar in Fig. 1).

The structure was refined for ∼60 macrocycles with our unsupervised refinement code and then manually rebuilt from the best autogenerated trial structure (hand-picked based on criteria of R-free, chain integrity, etc.) in Coot (22, 23). The procedure was repeated nine times and produced nine models, referred to as auto_1 to auto_9 in Fig. 1, until we reached an R-free value of 0.34 for the auto_9 model. Fig. 2 shows the evolution of the R-free value in all of the nine runs concatenated together. As usual, 50 trial structures were built by Buccaneer and refined by Refmac in each macrocycle, so that a total of 27,500 trial structures were built in the entire run. Finally, since the manually rebuilt starting structures were built sometimes quite aggressively (up to the 0.5σ level in the $2mF_{obs}−DF_c$ density map), four times in the solution process the structures contained too many errors for the unsupervised pipeline to correct. The resulting refined structures did not contain any useful signal to build the next starting models. Such runs were repeated with different starting models built in a more conservative way. These unsuccessful runs are not shown in Figs. 1 and 2.

The phases of the ninth model were of sufficient quality to allow us to stop using our computationally expensive refinement tool and switch to phenix.refine (24, 25), which was now able to consistently produce difference maps that enabled manual model building. The model was finalized in 40 manual build−refine iterations to produce the final model with an R-free of 0.25.

In agreement with the Sro7−Lgl multiple sequence alignments and WD40 motif assignment carried out in expasy.org, the refined crystal form 1 Lgl2 model comprises 14 WD40 repeats folded into two seven-bladed β-propeller structures (Fig. 3) in a similar topology and has overall similarity to Sro7. Comparing Lgl2 and Sro7, the N-terminal β-propellers are more similar between the structures than the C-terminal β-propellers (RMSD 1.75 Å and 2.44 Å for the N and C propellers, respectively), whose blades are substantially twisted with respect to each other (9). In addition, Lgl2 has unique structural features, particularly in regions outside the β-propeller framework. Nevertheless, the relative orientation between the β-propellers is similar in these two proteins. Thus, the framework important for the structure and rigidity of the protein core is conserved between yeast and humans, whereas the variable loop regions likely enable protein−protein interactions specific to each protein.

The higher-resolution form 2 Lgl2 crystal structure was solved by using the refined form 1 structure as a search model. The refined form 2 crystal structure showed a high structural similarity to the form 1 crystal structure (RMSD = 0.72 Å, 97.9% structure overlap). In addition, the higher resolution of this crystal form revealed additional structural details, including water molecules and other small molecules, as well as an alpha-helix at the Lgl2 N terminus (residues 14 to 20) that was missing in the form 1 model (Fig. 4). Finally, a phosphorylated version of Lgl2, showing no observable changes to the unphosphorylated protein, was solved from a third crystal form and was highly similar to the form 2 solved structure (9).
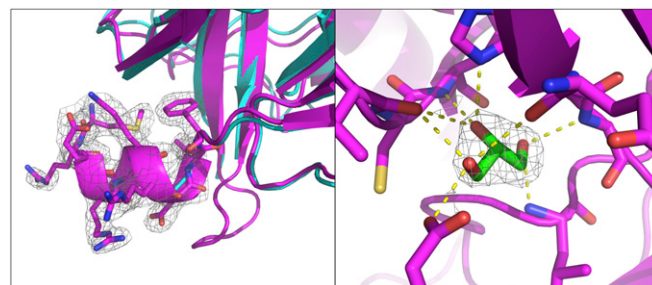

Fig. 4. Structural features in Lgl2 crystal form 2 not present in the crystal form 1 structure. (*Left*) Helix α1 appears only in the crystal form 2 (magenta) and not the crystal form 1 map. (*Right*) A glycerol molecule built in the native crystal form 2 model.
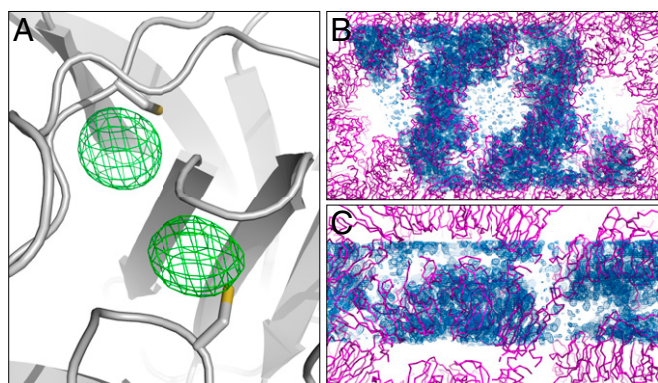
Fig. 5. (*A*) Anomalous difference map of the thiomersal-soaked Lgl2 crystal form 2 data, phased using the refined unphosphorylated crystal form 2 native model. The difference map, contoured at 5σ, clearly points to the locations of the bound mercury atoms adjacent to C933 and C456. Additional peaks could be also found near C889, C384, C389, and C626. (*B* and *C*) Orthogonal views of the solvent-flattened SAD derived electron density map generated using the detected Hg sites. Due to the low anomalous signal, the map quality is low (figure of merit = 0.223); nevertheless, the map correctly aligns with the refined structure.

As a final validation of the structure, the refined Lgl2 model was used to find Hg sites in the previously unsolved thiomersal-soaked unphosphorylated form 2 crystals. Locating the anomalous atom substructure ab initio using Patterson methods often fails due to weak anomalous signal. As shown in Fig. 5*A*, anomalous difference Fourier maps of the thiomersal data made using the refined native model phases show clear positive peaks located near surface-exposed cysteine residues. These peaks, positioned around these chemically reactive and accessible residues, are likely derived from interacting heavy atom compounds and thus could be used to locate and refine the mercury substructure. Using these positions, single-wavelength anomalous dispersion (SAD) phases were generated. The unbiased solvent-flattened electron density map generated using these SAD phases correctly enveloped the refined structure (Fig. 5 *B* and *C*), but was otherwise not interpretable for model building. Nonetheless, these results add to the validity of the MR-solved Lgl2 native structures.

Discussion

The ability of our method to perform large-scale deformations of a structural model without human intervention raises concerns about the danger of overrefinement. Indeed, when we were experimenting with refinement of incomplete models, sometimes we observed a tendency to overrefine incorrectly placed models: The models would fit the electron density well, yet no interpretable $2mF_{obs}-DF_c$ or difference density map would be observed outside the model region. We therefore paid careful attention to the interpretability of the density maps refined in the unsupervised runs. Several times early in the solution process, we had to roll back an entire unsupervised refinement run (60 macrocycles), because it produced models that we could not improve manually. Restarting the run with a different initial model always helped. This problem was not observed at late refinement stages (auto_8 and auto_9 in Fig. 1), when incorrect parts of the model showed up as large negative peaks in the difference density map and generally did not affect other regions of the map.

Subsequent validations of the solved structure confirmed the correctness of the model. First, the refined crystal form 1 model was used to obtain the higher-resolution crystal form 2 model, and the refined crystal form 2 model was successfully used to solve the structure of phosphorylated Lgl2 crystallized in the third crystal form, which was not part of this study (9). The two structures overlapped with RMSD of 0.72 Å, and 97.9% of the residues are present in both models. Second, the crystal form 2 model was used to find the location of mercury peaks in the crystals (crystal form 2) soaked in thiomersal and revealed unambiguous peaks above 5σ in the anomalous difference density map. The solution was validated in a blind way, because (*i*) the SAD dataset was not available when the crystal form 1 was solved and, (*ii*) while we had access to the 2.2-Å crystal form 2 dataset, it was not used in any way.

The difficulty in obtaining crystallographic phases for Lgl2 using conventional methods provided a serendipitous opportunity for applying and experimentally validating our method for unsupervised determination of macromolecular crystal phases. Having solved the Lgl2 structure with data from three different crystal forms, along with the opportunity to compare the results with an experimentally phased structure, validated the correctness of the solution provided by the method. With the growing abundance of computational power in the form of graphical processing units, this method will be useful for other challenging crystallographic projects.

Methods

Protein Crystallization and Data Collection. Lgl2 crystals were grown by vapor diffusion at 22 °C. The unphosphorylated Lgl2(13 to 978) protein was crystalized in two conditions: Condition A, 19.8% PEG 3350, 0.29 M Na$_2$SO$_4$, 0.1 M bis-Tris propane pH 7.5, 3% methanol, yielded mostly plates with a P422 lattice that were not suitable for data collection, and occasionally square pyramid shaped crystals (crystal form 1) in space group P2$_1$2$_1$2$_1$. Condition B, 18 to 21% PEG 2000 MME, 80 mM to 100 mM SPG (2:7:7 succinic acid:sodium dihydrogen phosphate:glycine) pH 6.0, yielded crystals (crystal form 2) in space group C2. Harvested crystals were soaked in cryoprotectant solutions (25% ethylene glycol and 10% glycerol in crystallization solutions for crystal forms 1 and 2, respectively) before being frozen in liquid nitrogen for data collection. For SAD data collection, unphosphorylated Lgl2(13-978) form 2 crystals were cross-linked with glutaraldehyde (25% solution) introduced by vapor diffusion, and then soaked in 1 mM thiomersal for 60 min before being washed with cryosolutions and frozen. Diffraction data were collected under cryogenic conditions at the Stanford Synchrotron Radiation Laboratory (SSRL) and the Advanced Photon Source. Unit cell parameters, data collection, and refinement statistics are shown in *SI Appendix*, Table S1.

Structure Solution and Refinement. Diffraction data were integrated by XDS (26) and scaled by Aimless (27). Due to the anisotropic diffraction, the unphosphorylated Lgl2 and pLgl2 crystal form 2 data were subjected to the STARANISO Server (Global Phasing Ltd.) (staraniso.globalphasing.org/cgi-bin/staraniso.cgi) to perform an anisotropic cutoff and to apply an anisotropic correction to the data. The Lgl2 model obtained as described in *Results* for the crystal form 1 data was further refined using phenix.refine program (24). For the native crystal form 2 data solution, the refined form 1 Lgl2 model was used as a search model for MR in phenix.phaser (24). Structure refinement was done using phenix.refine and Buster (28). Refinement statistics are provided in *SI Appendix*, Table S1.

Structure Analysis. Structure 3D alignments were performed using Click server for topology-independent comparison of bimolecular 3D structures (cospi.iiserpune.ac.in/click).

1. Humbert PO, Russell SM, Smith L, Richardson HE (2015) The Scribble–Dlg–Lgl module in cell polarity regulation. *Cell Polarity 1* (Springer, New York), pp 65–111.
2. Allam AH, Charnley M, Russell SM (2018) Context-specific mechanisms of cell polarity regulation. *J Mol Biol* 430:3457–3471.
3. Zhu J, et al. (2014) Phosphorylation-dependent interaction between tumor suppressors Dlg and Lgl. *Cell Res* 24:451–463.
4. Hong Y (2018) aPKC: The kinase that phosphorylates cell polarity. *F1000Res* 7:F1000 Faculty Rev-903.
5. Hattendorf DA, Andreeva A, Gangar A, Brennwald PJ, Weis WI (2007) Structure of the yeast polarity protein Sro7 reveals a SNARE regulatory mechanism. *Nature* 446:567–571.
6. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33:W244–W248.
7. Remmert M, Biegert A, Hauser A, Söding J (2011) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9:173–175.
8. Russell RB, Barton GJ (1992) Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins* 14:309–323.
9. Almagor L, Ufimtsev IS, Ayer A, Li J, Weis WI (2019) Structural insights into the aPKC regulatory switch mechanism of the human cell polarity protein lethal giant larvae 2. *Proc Natl Acad Sci USA* 116:10804–10812.
10. Ufimtsev IS, Levitt M (2019) Unsupervised determination of protein crystal structures. *Proc Natl Acad Sci USA* 116:10813–10818.
11. Cowtan K (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* 62:1002–1011.
12. Cowtan K (2008) Fitting molecular fragments into electron density. *Acta Crystallogr D Biol Crystallogr* 64:83–89.
13. Murshudov GN, et al. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* 67:355–367.
14. Brünger AT (1992) Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355:472–475.
15. Read RJ (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr A* 42:140–149.
16. Winn MD, et al. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67:235–242.
17. Adams PD, Pannu NS, Read RJ, Brünger AT (1997) Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proc Natl Acad Sci USA* 94:5018–5023.
18. Terwilliger TC (2000) Maximum-likelihood density modification. *Acta Crystallogr D Biol Crystallogr* 56:965–972.
19. Pannu NS, Read RJ (1996) Improved structure refinement through maximum likelihood. *Acta Crystallogr A* 52:659–668.
20. Eswar N, et al. (2006) Comparative protein structure modeling using Modeller. *Curr Protoc Bioinf* 15:5.6.1–5.6.30.
21. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Cryst* 40:658–674.
22. Emsley P, Cowtan K (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60:2126–2132.
23. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66:486–501.
24. Adams PD, et al. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66:213–221.
25. Afonine PV, et al. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* 68:352–367.
26. Takimura T, et al. (2010) Structures of the PKC-iota kinase domain in its ATP-bound and apo forms reveal defined structures of residues 533-551 in the C-terminal tail and their roles in ATP binding. *Acta Crystallogr D Biol Crystallogr* 66:577–583.
27. Evans PR, Murshudov GN (2013) How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* 69:1204–1214.
28. Blanc E, et al. (2004) Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr D Biol Crystallogr* 60:2210–2221.

BIOPHYSICS AND COMPUTATIONAL BIOLOGY