Biophysical Journal
Article

Biophysical Society

Residue coevolution and mutational landscape for OmpR and NarL response regulator subfamilies

Mayu Shibata,[1,2] Xingcheng Lin,[3,4] José N. Onuchic,[2,5] Kei Yura,[1,6,7] and Ryan R. Cheng[8,*]
[1]Graduate School of Humanities and Sciences, Ochanomizu University, Bunkyo, Tokyo, Japan; [2]Center for Theoretical Biological Physics, Rice University, Houston Texas; [3]Department of Physics, North Carolina State University, Raleigh, North Carolina; [4]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina; [5]Department of Physics and Astronomy, Chemistry, and Biosciences, Rice University, Houston, Texas; [6]Center for Interdisciplinary AI and Data Science, Ochanomizu University, Bunkyo, Tokyo, Japan; [7]Graduate School of Advanced Science and Engineering, Waseda University, Shinjuku, Tokyo, Japan; and [8]Department of Chemistry, University of Kentucky, Lexington, Kentucky

ABSTRACT DNA-binding response regulators (DBRRs) are a broad class of proteins that operate in tandem with their partner kinase proteins to form two-component signal transduction systems in bacteria. Typical DBRRs are composed of two domains where the conserved N-terminal domain accepts transduced signals and the evolutionarily diverse C-terminal domain binds to DNA. These domains are assumed to be functionally independent, and hence recombination of the two domains should yield novel DBRRs of arbitrary input/output response, which can be used as biosensors. This idea has been proved to be successful in some cases; yet, the error rate is not trivial. Improvement of the success rate of this technique requires a deeper understanding of the linker-domain and inter-domain residue interactions, which have not yet been thoroughly examined. Here, we studied residue coevolution of DBRRs of the two main subfamilies (OmpR and NarL) using large collections of bacterial amino acid sequences to extensively investigate the evolutionary signatures of linker-domain and inter-domain residue interactions. Coevolutionary analysis uncovered evolutionarily selected linker-domain and inter-domain residue interactions of known experimental structures, as well as previously unknown inter-domain residue interactions. We examined the possibility of these inter-domain residue interactions as contacts that stabilize an inactive conformation of the DBRR where DNA binding is inhibited for both subfamilies. The newly gained insights on linker-domain/inter-domain residue interactions and shared inactivation mechanisms improve the understanding of the functional mechanism of DBRRs, providing clues to efficiently create functional DBRR-based biosensors. Additionally, we show the feasibility of applying coevolutionary landscape models to predict the functionality of domain-swapped DBRR proteins. The presented result demonstrates that sequence information can be used to filter out bioengineered DBRR proteins that are predicted to be nonfunctional due to a high negative predictive value.

SIGNIFICANCE We extensively explored amino acid coevolution of the bacterial DNA-binding response regulator (DBRR) subfamilies at full protein length scale. The full-length coevolutionary analysis revealed the evolutionarily selected residue interactions between the linker and the domains. The mutational landscape from our coevolutionary analysis can be applied to predict the functionality of domain-swapped DBRR proteins, particularly for screening nonfunctional DBRR variants. This result will contribute to streamlining the development of novel biosensors. Additionally, we present the potential inactivation mechanism of DBRRs and their commonality across the subfamilies. Our result not only addresses biologically intriguing questions on inter-domain communication but also offers useful insights for effective domain rewiring of DBRRs.

INTRODUCTION

DNA-binding response regulators (DBRRs) function in tandem with their histidine kinase (HK) partner proteins as the primary sensory response mechanism in bacteria, called two-component system (TCS) (1). DBRRs function downstream of their cognate HK partner typically by regulating transcription through its interactions with the bacterial genome. A typical DBRR protein is composed of two domains connected by a linker: a conserved N-terminal receiver (REC) domain and a structurally diverse C-terminal DNA-binding effector (EFF) domain (2,3). This diversity further classifies DBRR family into subfamilies, including

Submitted November 10, 2023, and accepted for publication January 24, 2024.

*Correspondence: ryan.r.cheng@uky.edu

Editor: Rebecca Wade.
https://doi.org/10.1016/j.bpj.2024.01.028

the two most abundant classes: OmpR subfamily (approximately 48% of DBRRs calculated from https://www.ncbi.nlm.nih.gov/Complete_Genomes/RRcensus.html (2)) with winged helix-turn-helix EFF domain and NarL subfamily (28%) with helix-turn-helix EFF domain (4) (Fig. 1 *A*). Despite the variations, the typical functional mechanism can be summarized as follows. DBRRs assume both active and inactive states (5) (Fig. 1 *B*). When the proteins are not phosphorylated, this equilibrium shifts toward the inactive state where many DBRRs form a closed state through residue interactions between the REC and EFF domains (6,7). Upon phosphoryl transfer from the HK, the DBRR shifts toward the active state typically with an extended conformation that homodimerizes through the REC domain, enabling it to then bind to DNA (8).

DBRRs have gained the attention of bioengineering researchers as abundant building blocks for novel biosensors. Theoretically, novel DBRR-based biosensors can be made by connecting the REC and EFF domains of different DBRRs to make the proteins with arbitrary input/output combinations. In particular, a recent work has successfully demonstrated a systematic procedure for creating novel functional DBRRs by domain swapping using domains from the same DBRR subfamily (9). However, the error rate of the swapping experiments is not trivial. One factor that can improve the efficiency and success rate of the swapping experiments is the better understanding of residue interactions of DBRRs, specifically linker-domain and inter-domain interactions, which have not yet been extensively investigated for these two subfamilies. Knowledge on these interactions would also shed light on the functional conformational modes of DBRR proteins, which can uncover more of their functional mechanisms.

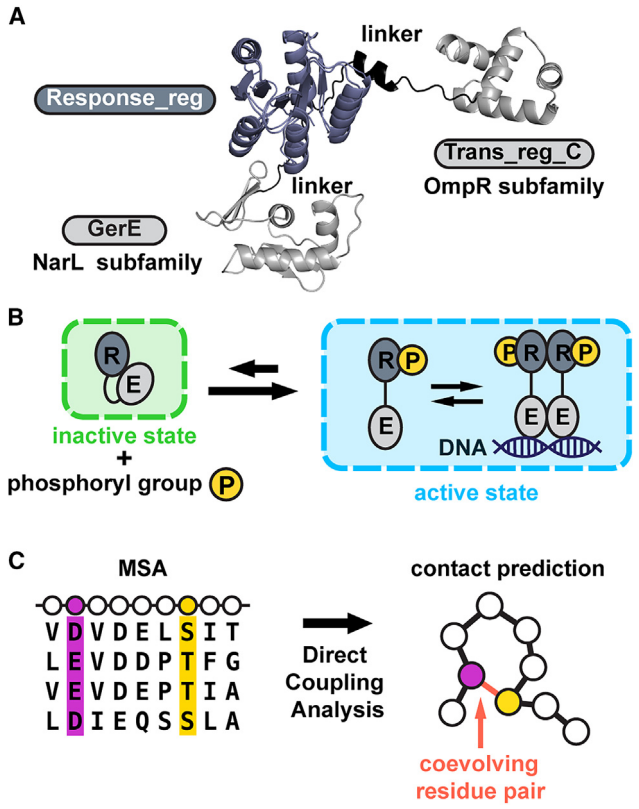

FIGURE 1 Cartoon illustrating amino acid coevolution within the dominant subfamilies of DBRRs. (*A*) The domain architectures of OmpR and NarL subfamilies consist of an REC domain and an EFF domain. The REC domain (Pfam (28): Response_reg) is common among all DBRR proteins, whereas the EFF domains are distinct for each subfamily: for OmpR (Pfam: Trans_reg_C) and NarL (Pfam: GerE). The 3D structures of "PDB: 7LZA" (OmpR subfamily) (29) and "PDB: 4ZMR" (NarL subfamily) (30) are also shown. (*B*) Basic functional mechanism of DBRRs of the two subfamilies. R, REC domain; E, EFF domain. (*C*) DCA quantifies amino acid coevolution. Circles show positions of amino acid residues. Highly coevolving residue pairs reflect the maintenance of their interactions over the course of natural selection.

Here, we thoroughly examine these residue interactions by analyzing their amino acid coevolutionary signatures using direct coupling analysis (DCA). DCA is a statistical method that is able to quantify the amino acid coevolution between pairs of residues within a protein (10,11) (Fig. 1 *C*), i.e., the observed correlations between the amino acid residue types at different sites developed mainly to maintain functionally important interaction over natural selection. A high degree of coevolutionary coupling between residue sites is often used as a proxy for inferring their spatial proximity in a three-dimensional (3D) structure of a folded protein or complex. This approach has been used successfully to predict residue contacts within a folded protein or protein complex (12–16). A number of previous studies have applied coevolutionary analysis to TCS proteins. The REC domain of RRs is often used as a model protein system to develop and evaluate coevolution-based analyses (17,18), many of which take advantage of the subfamily-specific homodimer interfaces (13,19). The coevolution of dimerization and histidine phosphotransfer (DHp) domain of the HKs was studied to understand intra-protein domain association of hybrid HK (20). The coevolution between HKs and RRs has been investigated as well to predict the specific HK-RR complex structure (21) and interacting HK-RR pairs (22), as well as how mutations affect HK-RR interaction specificity (23,24) and the functional activity of TCS (25) using sequence coevolutionary landscapes. Mutational landscapes inferred using coevolutionary information have been used to study the effect of mutations on the fold and function of proteins outside TCS as well (26,27).

We specifically focus on the sequence diversity of the two major classes of DBRR proteins, OmpR-like and NarL-like DBRRs, to identify coevolving residue pairs of DBRRs for full-length proteins that include REC domain, linker, and the EFF domains using DCA. We find evolutionarily selected linker-domain residue interactions, which agree with residue contacts in known experimental structures. Additionally, we create a mutational landscape from our

inferred coevolutionary couplings to predict the functionality of bioengineered DBRR proteins that were created via domain swap. We further explore residue coevolution between the REC and EFF domains to find the inter-domain residue interactions. Our analyses identified coevolution in the REC-EFF residue interactions of known inactive monomer structures. Additionally, integrating molecular dynamics (MD) simulation and structural analyses, we find previously unknown inter-domain residue interactions that are associated with potential inactive monomer conformations. In these conformations, DNA recognition helix is partly covered by REC domain, making the DNA interface inaccessible in both subfamilies. Our results suggest that OmpR and NarL subfamily DBRRs are implemented with the same inactivation mechanism through REC-EFF residue interactions.

MATERIALS AND METHODS

Multiple sequence alignment

Bacterial proteins of OmpR and NarL subfamilies were collected from the Pfam database (ver. 35.0) in InterPro (31) as those consist of the following domain architecture: PF00072 (Response_reg)-PF00486 (Trans_reg_C) for OmpR subfamily and PF00072 (Response_reg)-PF00196 (GerE) for NarL subfamily. Full-length Multiple Sequence Alignments (MSAs) and two-domain MSAs (without the linker) were prepared separately. To construct full-length MSAs, full-length amino acid sequences were retrieved from UniProt (32). After removing sequences with nonstandard amino acids, 1000 sequences were randomly sampled as MSA seeds. Seed sequences were first aligned by MAFFT (33) then excessively gapped columns (gap fraction >50%) were removed. HMM profiles of the seed MSAs were computed by HMMER (34). The amino acid sequences prepared above were aligned to the HMM profile using HMMER. To improve the quality of the MSAs, largely gapped rows (gap fraction >10%) and columns (gap fraction >50%) were further removed. REC domain positions of OmpR and NarL subfamily MSAs were mapped based on the alignment between amino acid sequences of Protein Data Bank (PDB) (35) entries "PDB: 7LZA" (OmpR subfamily) and "PDB: 4ZMR" (NarL subfamily) computed by MAFFT. Instead of building MSAs from UniProt sequences, two-domain MSAs were prepared directly from the protein domain MSAs provided by Pfam. Amino acid sequences with the fraction of gaps exceeding 20% as well as sequences with nonstandard amino acids were removed. Then, sequences of REC (PF00072) and EFF (PF00196 or PF00486) domains were concatenated to yield two-domain MSAs.

The sampling bias correction by sequence similarity reweighting (threshold of 0.8) (36,37) left 35,618.68 (full-length MSA) and 30,728.84 (two-domain MSA) effective sequences in OmpR subfamily, and 25,452.28 (full-length MSA) and 24,302.67 (two-domain MSA) effective sequences in NarL subfamily.

Structural data, residue contacts, and functional residues

All the 3D structures in this study were obtained from PDB. Among the full-domain PDB entries listed by Pfam, structures with resolution worse than 3.2 Å were discarded. The residue numbers in the following text are reported using the reindexed residue numbers where the first residue with coordinates was counted as the first residue. Each residue number was mapped to an MSA position based on the alignments of protein sequences to HMM profiles generated by HMMER. Phosphorylated structures were identified by the presence of beryllium trifluoride (BeF_3^-), a phosphoryl-group mimic, at the highly conserved aspartic acid site. Residue pairs were identified as contacts when their Cα-Cα distance was less than 10 Å (full-length DCA) or Cα-Cα distance was no more than 12 Å (two-domain DCA) in at least one 3D structure. A subset of residue pairs forming contacts in both monomer and dimer structures (not necessarily in the same structure) is referred to as monomeric-and-dimeric contacts. Chain A of "PDB: 7LZA" (OmpR subfamily) and "PDB: 4ZMR" (NarL subfamily) were used as representative structures of OmpR and NarL subfamilies, respectively, in the following analyses. The secondary structures of the representative structures were assigned by DSSP (38,39) after missing heavy atoms were modeled by MODELLER (40). Residues that form homodimeric contacts, HK interface contacts, and DNA interface contacts were identified when at least one heavy atom of a residue was found within 3.5 Å of its target in at least one known experimental structure (other DBRR protein, HK partner, and DNA, respectively).

DCA

The strength of the amino acid coevolution (direct couplings) between all residue pairs were quantified from MSA data using DCA. In DCA, an amino acid sequence $\vec{\sigma}$ of length L from an MSA $\{\vec{\sigma}^{(l)}\}_{l=1,\ldots,M}$ with M sequences was assumed as a sample extracted from a probabilistic model of the MSA. Thus, a probability distribution $P(\vec{\sigma})$ derived from the model must reproduce empirical observations of the MSA: single-site frequency $f_i(\sigma_i)$ and pairwise frequency $f_{ij}(\sigma_i, \sigma_j)$ of position i and j. The most generalized form of such a probabilistic distribution was given by Potts-model probability distribution shown as

$$P(\vec{\sigma}) = \frac{1}{Z} exp\left(\sum\nolimits_{1 \leqq i \leqq L} h_i(\sigma_i) + \sum\nolimits_{1 \leqq i < j \leqq L} J_{ij}(\sigma_i, \sigma_j)\right), \tag{1}$$

where h_i and J_{ij} represent single-site fields and pairwise couplings, respectively. The latter is associated with the direct coevolution between positions i and j of the MSA. Parameters of the probability distributions were optimized in a pseudo-likelihood maximization approach using MATLAB (The MathWorks) codes from Ekeberg et al. (36,37). The strength of direct couplings between the positions was given in the form of Frobenius norm

$$FN_{ij} = \sqrt{\sum_{1 \leq \sigma_i, \sigma_j \leq q} J_{ij}(\sigma_i, \sigma_j)^2}, \tag{2}$$

where $1, \ldots, q$ denote all the observed states (amino acids and gap) of the position in the MSA. Average product correction was applied to the raw Frobenius norms expressed in the following formula:

$$\mathrm{FN}_{ij}^{APC} = \mathrm{FN}_{ij} - FN_{\cdot j} FN_{i\cdot} / FN_{\cdot\cdot}, \tag{3}$$

where $FN_{\cdot j}$, $FN_{i\cdot}$, and $FN_{\cdot\cdot}$ represent mean over position i, j, and both i and j, respectively.

The output scores FN_{ij}^{APC} were standardized as

$$Z_{ij}^{FN\text{-}APC} = \left(FN_{ij}^{APC} - \overline{FN^{APC}}\right) \Big/ s, \tag{4}$$

where $\overline{FN^{APC}}$ and s show the mean and standard deviation of FN_{ij}^{APC} for every i and j considered. $Z_{ij}^{FN\text{-}APC}$ is referred as coevolutionary score and coevolving residue pairs identified by DCA are referred as DCA pairs in the following text. In the two-domain model (REC and EFF domain only), intra-domain DCA pairs of which separations were less than five, i.e., $|i - j| < 5$, were eliminated to focus on mid- and long-range residue coevolution.

Mutational favorability of domain-swapped proteins

Experimental datasets of domain-swapped proteins were obtained from references (9,41,42). In these experiments, domain-swapped DBRR proteins were created by concatenating the N-terminal region from one DBRR (DBRR 1) with the C-terminal region of another DBRR (DBRR 2) such that the resultant protein consisted of an REC domain, linker, and EFF domain (in that order). However, it should be noted that the MSA positions of the concatenated DBRR segments may overlap, i.e., sometimes contain the same MSA positions. Consequently, any computation of the mutational change with respect to a wild-type reference requires rules for selection of that reference (discussed further below). The pairwise coevolutionary couplings of a residue pair J_{ij} were computed by DCA. The difference in the direct couplings of position i (amino acid a) and j (amino acid b) of the wild-type protein upon domain swapping $\Delta J_{ij(a\to a',b\to b')}$ was computed as

$$\Delta J_{ij(a\to a',b\to b')} = \sum_{ref}\left(J_{ij(a',b')} - J^{ref}_{ij(a,b)}\right) \Big/ m, \quad (5)$$

where a' and b' represent amino acids of position i and j of the domain-swapped proteins, respectively, and the sum is taken over the number of wild-type reference proteins m. The reference wild-type proteins were determined as follows with respect to the MSA positions of DBRR 1 and/or DBRR 2: 1) if both positions i and j of the domain-swapped protein were encoded in the same wild-type protein (DBRR 1 and 2), then the respective wild-type protein was used as the reference (hence, $m = 1$ and $\Delta J_{ij(a\to a',b\to b')} = 0$); 2) if the resultant domain-swapped protein is constructed such that the i-th MSA position belongs to DBRR 1 and containing the j-th MSA positions of both DBRR 1 and DBRR 2, then the wild-type reference was chosen to be DBRR 1 (hence, $m = 1$); and 3) if the resultant domain-swapped protein is constructed such that both the i-th and j-th MSA positions of the concatenated segments were found in both DBRR 1 and DBRR 2. Then these two proteins were used as the wild-type reference (hence, $m = 2$). Mutational favorability

$$F = \sum \Delta J_{ij(a\to a',b\to b')}, \quad (6)$$

was computed by taking the sum of $\Delta J_{ij(a\to a',b\to b')}$ of all the DCA pairs that coincided with residue contacts.

MD simulation

Novel monomer structures were generated by MD simulation with GROMACS (43) 5.0.4 in the Structure-based Model (SBM) approach (see Supporting Material for more details on the SBM approach), using the representative structures of the subfamilies as templates. After missing atoms were filled using MODELLER, the template structures were processed using SMOG (44) (ver. 2.1) to generate topology files including coarse-grained Cα models (45) and potentials. Each DCA pair was incorporated to the topology files with a potential V_{ij}, which combined Gaussian potential to r^{-12} repulsive term of Lennard-Jones potential (46)

$$V_{ij} = A\left(\left(1 + (1/A)\left(\sigma_{NC}/r_{ij}\right)^{12}\right)\right.$$
$$\left.\left(1 - exp\left(-\left(r_{ij} - r_0\right)^2 \Big/ \left(2\sigma^2\right)\right)\right) - 1\right), \quad (7)$$

where r_0 (equilibrium distance), σ (decay), and A (amplitude) are the parameters determining the shape of the Gaussian potential well, and σ_{NC} defines the exclusion volume. This potential stabilized the residue pairs to be at the equilibrium distance. The amplitude parameter A was fixed at 300ε, where ε is the unit of energy of the structure-based model, to constrain the model to generate structures consistent with the desired DCA pairs as contacts. The structures of the domains were fixed by strengthening dihedral angle parameters k^1_d and k^3_d of the dihedral angle potential V_d

$$V_d = k^1_d(1 - cos(\varphi - \varphi_0)) + k^3_d(1 - cos(3(\varphi - \varphi_0))), \quad (8)$$

by a factor of 10 from the given values in the topology files prepared by SMOG (φ_0 shows the angle between the planes defined by residues i, j, and k and residues j, k, and l). Moreover, all dihedral and pair constraints for intra-linker or linker-domain residue interactions were removed to assure the flexibility of the linker. The simulation consisted of five stages. In the first four stages, r_0 and σ of Gaussian potentials of DCA pairs were progressively altered (Table S1). After confirming that all the incorporated DCA pairs reached equilibrium distance, the MD simulation was extended for a longer duration (fifth stage) starting with the final structures of the previous four-stage simulation under the same condition as the fourth stage. See Supporting Material for more details on coarse-grained simulation. A coarse-grained structure was periodically sampled from the trajectory to create an ensemble of 2000 structures. Side chains were added to the obtained Cα structures using MODELLER for structural analysis.

Structural analysis

Residue contacts and local conformational frustration of the sampled structures were computed by R package frustratometeR (47). After extracting the samples satisfying all the incorporated DCA pair constraints, principal-component analysis (PCA) was performed on internal positional vectors between the intra-monomer Cα atom pairs. The converted vectors along the first three principal components were subjected to k-means clustering ($n_{\text{clusters}} = 3$). The residue contacts observed in more than half of the samples were extracted and compared among the three clusters of each subfamily. Two clusters were merged when more than 95% of the contacts in both clusters were shared. Structural similarity search was performed using Dali server (48) against PDB. Residue flexibility was predicted using MEDUSA (49) from an amino acid sequence. As references, residue contacts, and local conformational frustration of chain A of structures "PDB: 7LZ9" (29) (inactive monomer) and "PDB: 7LZA" (active monomer) of OmpR subfamily, and chain A of "PDB: 4HYE" (50) (inactive monomer) and "PDB: 4ZMR" (active dimer) in NarL subfamily were also computed after missing residues were filled in using MODELLER. The solvent-accessible surface area (SASA) of the side chain of an amino acid residue was calculated using GetArea server (51).

RESULTS AND DISCUSSION

Full-length coevolutionary profiles detected strong residue coevolution of linker-domain residue interactions

Residue pairs with DCA score $Z^{FN-APC}_{ij} > 1.964$ (OmpR subfamily) and 1.487 (NarL subfamily) were identified as highly coevolving residue pairs (herein referred to as DCA pairs). These cutoffs were defined such that residue pairs with a DCA score higher than the cutoff were reliably observed to be structural contacts, with a positive predictive value (PPV) greater than 0.95 against union sets of known residue contacts (Cα-Cα distance <10 Å) for each subfamily (Fig. S1). A PPV of 0.95 was chosen to select the most strongly coevolving residue pairs in this part of the study. The DCA score filtering left 711 OmpR DCA subfamily pairs and 908 NarL subfamily DCA pairs. The top coevolving residue pairs identified by DCA in fact captured the unique contacts for each subfamily (Fig. 2 *A*), in particular

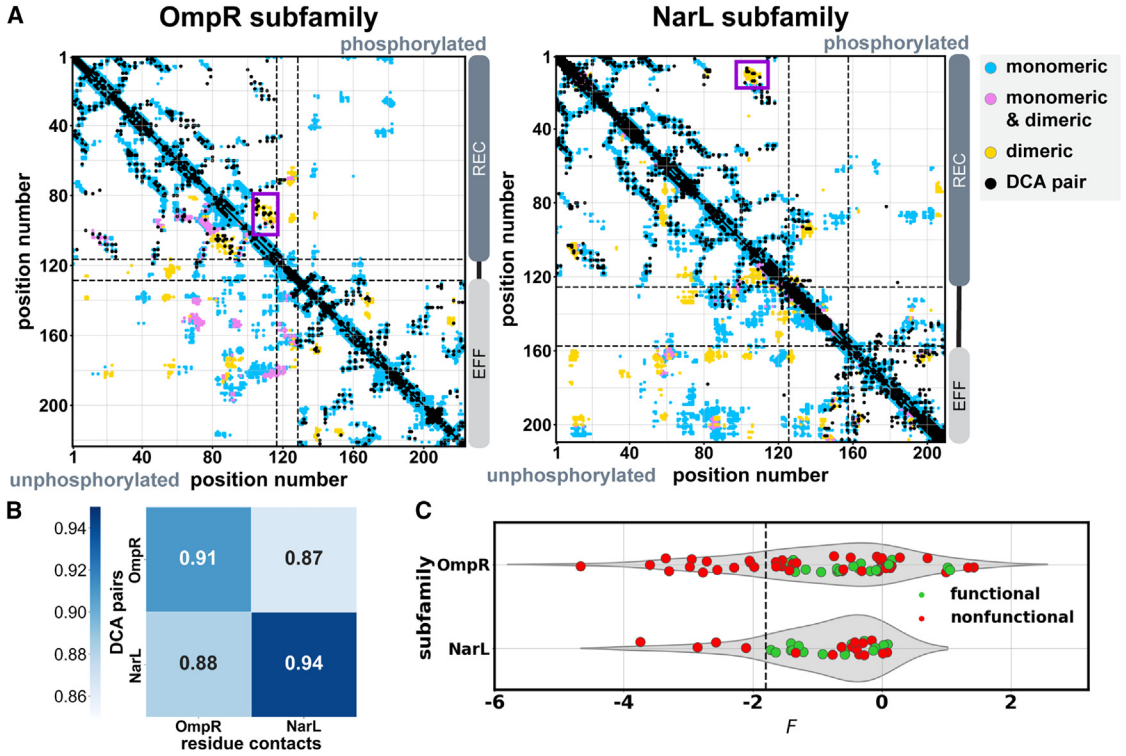

FIGURE 2 Residue contacts, DCA pairs, and overall mutational favorability of the domain-swapped DBRRs. (*A*) Contact maps of unphosphorylated (*lower triangle*) and phosphorylated (*upper triangle*) state structures with DCA pairs. Dashed lines indicate the domain boundaries. Purple boxes show DCA pairs successfully capturing subfamily-specific active dimer interfaces. The full-length coevolution profile included linker-domain residue interactions as well as intra-domain residue interactions. (*B*) Confusion matrix showing the PPVs of the subfamily-specific DCA pairs for predicting the residue contacts for the OmpR and NarL subfamilies within the REC domain. (*C*) Overall mutational favorability F of the domain-swapped DBRRs. Full-length coevolutionary profiles were able to distinguish a subset of the nonfunctional domain-swapped DBRRs computationally.

the active dimer interfaces in REC domain, which are known to vary among the DBRR subfamilies (α4-β5-α5 and α1-α5 interface for OmpR and NarL subfamilies, respectively (4)). As can be assumed from this result, the DCA predictions for the OmpR and NarL subfamilies captured the subfamily specific contacts of the DBRR proteins within the REC domain (Fig. 2 *B*), despite the high similarity of the REC domains across the DBRR subfamilies. Further, strong residue coevolution was found between the linker and domains (linker-REC domain and linker-EFF domain) additionally to intra-domain residue coevolution in both subfamilies (Fig. 2 *A*). This shows that the linker-domain residue interactions were evolutionarily selected and hence are important for the function of the DBRR proteins. This is consistent with the previous observations that the linker and REC domain cooperatively regulate the function of EFF domain and that the amino acid residues in the linker play an important role in DBRR function (52,53).

Full-length coevolution models can computationally predict nonfunctional domain-swapped proteins

To further test our subfamily-specific coevolutionary models, we constructed mutational landscapes and applied them to examine bioengineering experiments where the C-terminal portion of a DBRR protein is removed and replaced with one from a nonnative DBRR protein within the same subfamily (domain-swapped DBRR proteins) (9,41,42). We computed the overall mutational favorability F of these domain-swapped DBRRs relative to our wild-type reference protein(s), from direct coupling matrices given by DCA (see section "materials and methods"). Generally, mutations leading to negative couplings reflect evolutionarily unfavorable amino acid combinations found in sequence data. Based on this, we hypothesized that there exists a threshold value for our favorability index F such that bioengineered proteins below that threshold would not be expected to be functional. We observed that all domain-swapped proteins less than a threshold of $F = -1.8$ were experimentally nonfunctional (Fig. 2 *C*). It is possible that the high similarity between DBRR subfamilies in terms of structure and function results in a similar degree of tolerance for the unfavorable interactions. Further studies, however, are needed to verify this hypothesis. The nonfunctional domain-swapped proteins predicted by this cutoff value accounted for 31.0% (13 out of 42) of nonfunctional domain-swapped OmpR subfamily proteins and 25.0% (four out of 16) of nonfunctional domain-swapped NarL subfamily proteins (Fig. S2; Table S2). Our overall mutational favorability

index F can aid efficient domain-swapping experiments by narrowing down the list of potentially functional domain-swapped DBRRs. This analysis is also expected to be the starting point to develop computational prediction model of DBRR domain-swapping experiments.

Two-domain coevolutionary models predict novel monomer inter-domain residue interactions in addition to known inter-domain residue interactions of inactive monomers

We further examined the inter-domain residue coevolution between the REC and EFF domains that were expected from experimental 3D structures. To purely focus on REC-EFF domain residue coevolution, DCA was performed on MSAs where the linker region was removed (referred to as the two-domain MSAs). Following the same protocol as for the full-length coevolutionary analysis, residue pairs with DCA score $Z_{ij}^{FN\text{-}APC} \geq 1.2$ were identified as DCA pairs in this part of the study. This Z score cutoff resulted in PPV >0.96 against the union sets of residue contacts (Cα-Cα distance <12 Å) from all the PDB structures of the subfamilies (Fig. S3). This chosen cutoff was selected to reliably identify strongly coevolving residue pairs. Although the specific value of the cutoff is an arbitrary choice, this is a reasonable choice and minor variations do not affect the conclusion. This filtering left 450 OmpR subfamily DCA pairs and 374 NarL subfamily DCA pairs. DCA pairs overlapped on the subfamily-distinct dimeric interfaces (purple box in Fig. 3 *A*), showing that our two-domain coevolutionary profiles successfully captured unique characteristics of the two subfamilies.

DCA pairs from the two-domain MSAs included six inter-domain pairs (pairs a–f) in OmpR subfamily and four inter-domain pairs (pairs g–j) in NarL subfamily (Fig. 3 *B*; Table S3). Interestingly, those inter-domain pairs lay on the two separate areas in both subfamilies: A) the area between position 7 and 17 of REC domain (mostly α1 of the protein) and DNA recognition helix (α8 for OmpR subfamily and α9 for NarL subfamily), and B) the area between position 70 and 87 of REC domain (α3-β4-α4) and DNA recognition helix. We will refer to these areas as area A and B in the following text (orange boxes in Fig. 3 *B*). Inter-domain DCA pairs f, i, and j (area B) coincided with residue contacts of closed inactive monomer structures, suggesting that the inactivating mechanism through these inter-domain residue interactions is shared by the two subfamilies.

We next explored the possibility that the inter-domain DCA pairs (pairs a–e, g, and h, mostly from area A) are spatial contacts, despite not being observed in any experimental structures. The DCA scores of these unverified DCA pairs were

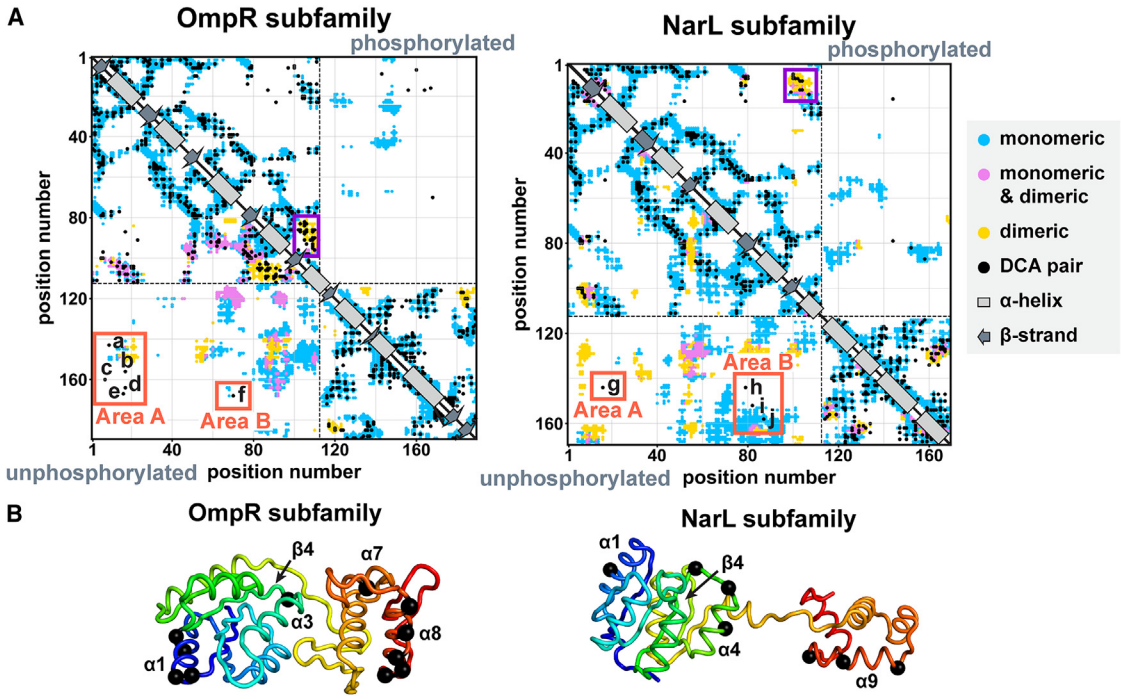

FIGURE 3 Residue contacts, DCA pairs, and 3D structures. (*A*) Contact maps of unphosphorylated (*lower triangle*) and phosphorylated (*upper triangle*) state structures with DCA pairs. Dashed lines indicate the domain boundaries. Purple boxes show DCA pairs successfully capturing subfamily-specific phosphorylated active dimer interfaces. Both subfamilies had inter-domain DCA pairs in two areas, namely areas A and B. Most of the DCA pairs in area B corresponded to known inactive monomeric contacts. (*B*) Inter-domain DCA pairs. The Cα atoms of these pairs are shown in black spheres on representative structures of the subfamilies. REC domain residues were located around α1 and a region over α3-β4-α4. EFF domain residues were mostly located in DNA recognition helix.

comparable to those of contacting inter-domain pairs (pairs f, i, and j from area B) (Table S3); note that all of these pairs satisfy the DCA score cutoff for statistical significance. Generally, the stronger coevolutionary signals indicate stronger direct couplings between the positions, suggesting higher likelihood of having direct residue interactions. Therefore, we explored the idea of these unverified inter-domain DCA pairs forming residue contacts, specifically monomer residue contacts, by MD simulation and structural analysis. We focused on monomeric interactions, because the regulation of DNA-binding activity through monomeric residue interactions has been established (7,54). It is plausible that these inter-domain DCA pairs are involved in the regulation of DBRR activity by inhibiting DNA binding of the DBRR proteins, for instance.

Simulations reveal alternative monomer conformations of DBRR proteins

Three separate simulations were conducted where the unverified inter-domain DCA pairs were treated as spatial contacts between the REC and EFF domains: one on OmpR subfamily protein with the five inter-domain DCA pairs (pairs a–e), one on NarL subfamily protein with pair h, and one on NarL subfamily protein with pair g. Inter-domain DCA pairs f, i, and j were excluded because they matched residue contacts of known PDB structures. Pairs g and h of NarL subfamily were treated separately because they cannot be satisfied simultaneously without disrupting the fold of REC domain, which is out of the focus of this particular work (Fig. S4).

The resulting structures were analyzed using PCA-based clustering (Table S4). In OmpR subfamily, the clustering identified two dominant clusters, the models of which will be referred to as OmpR-A1 and OmpR-A2. The clusters represented by the former was the predominant one and comprised 1416 samples (70.9%) (Fig. S5 *A*). In NarL subfamily, the clustering left a single cluster in each simulation case, resulting in two models in total: NarL-A1 with DCA pair g from area A (2000 samples) and NarL-B1 with DCA pair h from area B (1989 samples) (Fig. S5 *B* and *C*). The mean pairwise backbone root-mean-square deviation was 2.28, 2.96, 6.37, and 5.87 Å for clusters represented by OmpR-A1, OmpR-A2, NarL-A1, and NarL-B1, respectively. The larger root-mean-square deviations of NarL subfamily models can be explained by the fewer inter-domain DCA constraints added to the simulation.

OmpR subfamily models: OmpR-A1 and OmpR-A2

The N-terminal of EFF domain (a β sheet of β6, β7, β8, and β9) was dissociated from the cognate α6 in our models (purple boxes in Fig. 4 *A*). To assess the feasibility of this dissociation, we examined "PDB: 6IJU" (55), "PDB: 6IS4" (56) and "PDB: 6KYX" of an OmpR subfamily protein, "UniProt: Q9KJN4". These were identified to be structurally similar to EFF domain of OmpR-A1 (Dali Z score $\geqq$7.3). In these PDB entries, the N terminus of EFF domain partially unraveled at residues of the β7-β8 loop, splitting apart the contacts between β6 and β7 with α6 (the number of secondary structure elements here follows that of OmpR-A1). This unraveling suggests the increased flexibility of the EFF domain linkers, which can also enable our model structures to be formed. Furthermore, alanine in the β9-α6 linker of our model protein was predicted to be flexible based on its amino acid sequence (Fig. S6 *A*). This residue is responsible for the different arrangement of the β sheet and α6 in our models (Fig. S6 *B*). These, too, support the notion that the partial dissociation of EFF domain in our models can occur. To better assess the feasibility of our models, local conformational frustration was examined. This quantity reveals the energetic favorability of the spatial placements of the residue pairs in the given structure compared to decoy structures where the interacting pairs are moved (57). The high frustration indicates the energetic unfavorability of the residues in the given placement. The general pattern of the local conformational frustration of intra-domain contacts in our OmpR models was consistent with that of the active monomer PDB structure (Fig. 4 *A*). This confirms that a partial dissociation of EFF domain or the newly identified domain arrangement did not affect energetic stability of intra-domain contacts, further supporting the feasibility of our models.

The local conformational frustration can also be used to understand biological functions of the proteins. The highly frustrated regions include binding sites where the binding to the ligands or biomolecular partners resolves the frustration (57,58). The fewer highly frustrated residue contacts were observed in two contact patches composing α1 (REC domain, including HK interface residues) and α8 (EFF domain, including DNA interface residues) in both models than reference structure (brown boxes in Fig. 4 *A*). This observation was confirmed by χ^2 goodness-of-fit test, which compared the fraction of three classes of contacts, i.e., highly frustrated, neutral, and minimally frustrated contacts (statistical significance, 0.05; Table S5). Less frustration in these areas suggests our models are less likely to bind to HK or DNA.

OmpR-A1 and OmpR-A2 exhibited six shared inter-domain contact patches. The only patch that contained the noncontacting inter-domain DCA pairs (predicted contacts) involved residues 8–30 vs. 175–201 (position 5–28 vs. 142–168). Four other contact patches were found between residue 10–44 and 135–158 (position 7–42 vs. unmapped 125) (Fig. 4 *A* and *B*). The contacts in these patches explain the major difference of OmpR-A1 and OmpR-A2. The last inter-domain contact patch, patch X, lay between residue 102–106 and 187–192 (position 100–104 vs. 154–159, mainly β5-α5 loop vs. α8) (Fig. 4 *A* and *B*). It should be

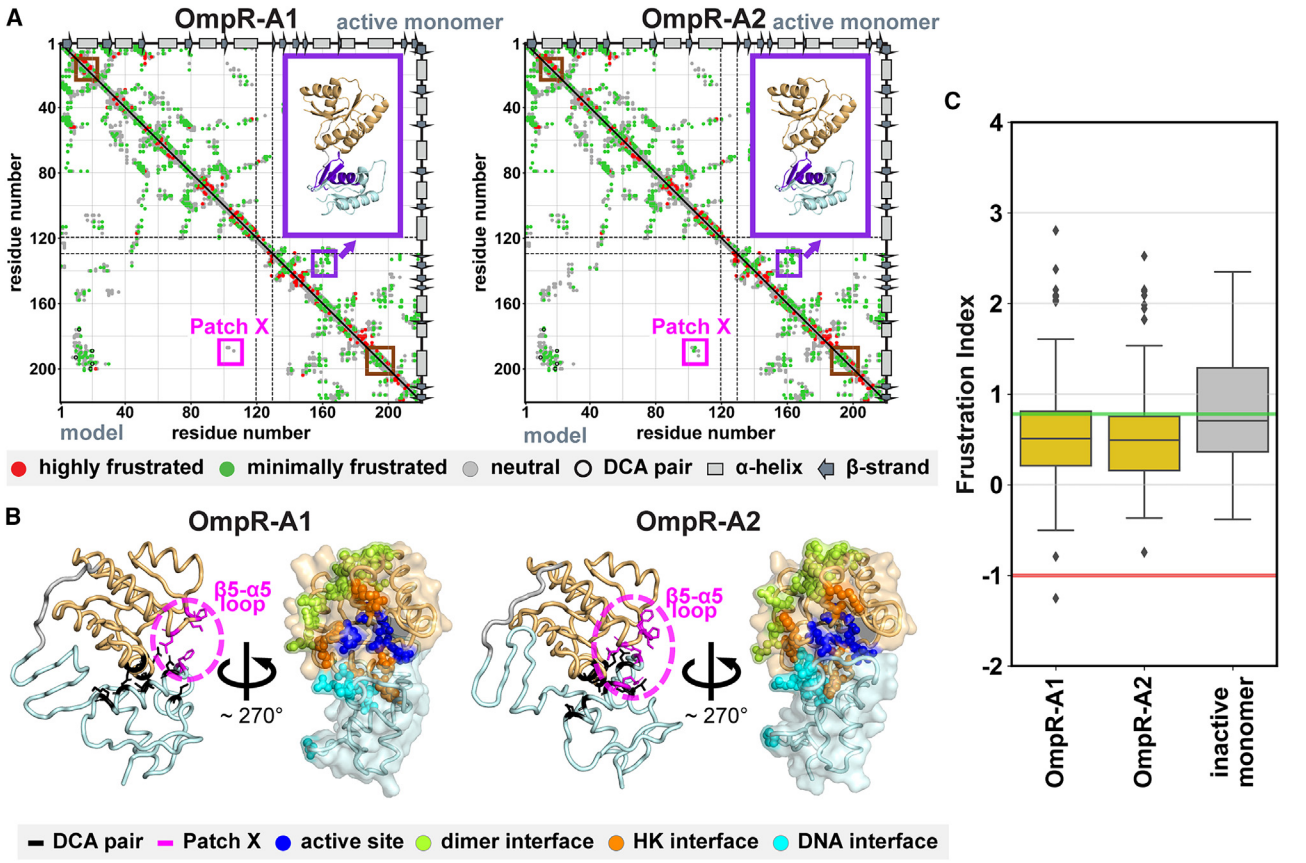

FIGURE 4 OmpR subfamily models. (*A*) Contact maps of the model (*lower triangle*) and the active monomer template structure (*upper triangle*) colored by the degree of local conformational frustration. Residue number follows the reindexed template 3D structure. Purple boxes show the EFF domain contacts lost in our models with the template structure. Two brown patches in α1 and α8 were less frustrated in our models, suggesting less capability of HK and DNA binding of our models. An inter-domain contact patch, patch X, formed as a consequence of forming contacts at the inter-domain DCA pairs. (*B*) Model structures showing DCA pairs and residue contacts in patch X (*left*) and functional sites (*right*). HK interface and DNA interface residues in our models were less exposed than the active monomer reference structure. (*C*) Boxplots of frustration index of inter-domain residue contacts. Green and red lines show the threshold frustration index of minimally and highly frustrated contacts, respectively.

reminded that five contact patches, including patch X, formed as a consequence of five inter-domain DCA pairs in contact. Among the inter-domain residue contacts of these models, a patch X contact of OmpR-A2 coincided with a known residue contact of inactive monomers. Additionally, five and six inter-domain contacts from OmpR-A1 and OmpR-A2 (respectively) were observed to get in proximity (Cα-Cα $\leqq$ 15 Å) in known inactive monomer PDB structures. Those inter-domain contacts included ones from contact patch with noncontacting DCA pairs and patch X. This supports the feasibility of our new inter-domain interface and implies our models to be inactive monomer structures. The local conformational frustration of the inter-domain interfaces of OmpR-A1 and OmpR-A2 were comparable to those of known unphosphorylated inactive monomers (Fig. 4 *C*; Mann-Whitney U test comparing the distribution of the frustration index between the model and the template at the significant level 0.05; *p*-value 0.12 for OmpR-A1 and *p*-value 0.084 for OmpR-A2). The frustration analysis of intra- and inter-domain contacts further supports the feasibility of our models. However, monomeric conformations employing our novel inter-domain interface have not yet been experimentally observed.

Interestingly, the residues of patch X included functionally important residues (Table S6). REC domain residues involved a conserved lysine (K_A) and the three cognate residues to the C terminus ($K_A + 1$, $K_A + 2$, and $K_A + 4$). K_A is one of the active site residues (59) and its two neighboring residues govern mainly the autophosphorylation kinetics of RRs (60,61). $K_A + 4$ was reported to cooperatively function with residues in α1 in single-domain RR through allosteric interactions (61). In our models, α1 and these functional residues interact with DNA recognition helix simultaneously. The residue interactions in patch X together with inter-domain contacts predicted by DCA pairs may be a molecular mechanism for optimizing the phosphorylation timescale of OmpR subfamily DBRRs in accordance with the biological responses determined by EFF domains in DBRRs.

The SASAs of the functional sites of the models were compared against those of the active monomer structure where all the functional sites are assumed to be exposed. More than 35% of SASAs were lost in the DNA interface residues and

TABLE 1 SASA of functional sites in OmpR subfamily models and the reference structure

Functional sites	Active monomer (Å^2)	OmpR-A1 (Å^2)	OmpR-A1 (%)	OmpR-A2 (Å^2)	OmpR-A2 (%)
Active site	94.42	163.15	172.8	99.07	104.9
Dimer interface	595.45	632.06	106.1	725.98	121.9
HK interface	553.67	355.53	64.2	313.41	56.6
DNA interface	548.14	327.86	59.8	351.57	64.1

The % symbol denotes the ratio of the SASA of our computational model to the SASA of the active monomer as a percentage.

HK interface residues in our models (Table 1; Fig. 4 *B*). This suggests DNA and HK are less likely to access their binding interface in OmpR-A1 and OmpR-A2, consistent with our previous discussion on less frustration on HK and DNA interfaces and known inactive monomer contacts. Integrating the discussions above, the noncontacting inter-domain DCA pairs a–e in area A of OmpR subfamily are associated with the potential inactive monomer conformation that inhibits DNA-binding activity of the OmpR subfamily proteins. Additionally, our analyses suggest that these conformations should tune the duration of DBRR phosphorylation to biological responses determined by EFF domains.

NarL subfamily models: NarL-A1 and NarL-B1

Intra-domain residue contacts and their frustration pattern of our models agreed well with those of the template active monomer structure (Fig. 4 *A*). This confirms that the novel domain arrangements did not result in a drastic change in the energetics of the intra-domain residue contacts. In NarL-A1, α1 was less frustrated than the reference, which was statistically significant (p-value 0.015 at χ^2 goodness-of-fit test, significance level 0.05) (Brown box in Fig. 5 *A*). The less frustrated α1 indicates that NarL-A1 is less likely to bind to HK.

Both models showed two inter-domain contact patches. The patch with the unverified inter-domain DCA pair was found at residue 26–30 vs. 174–180 (position 16 unmapped vs. 142–148: α1 vs. mostly α9) for NarL-A1 and residue 90–92 vs. 174–183 (position 79–81 vs. 142–151: β4-α4 loop vs. mostly α9) in NarL-B1 (Fig. 5 *A*). The other inter-domain patch was found between residue 110–115 and 174–180 (position 99–104 vs. 142–148), corresponding to patch X in OmpR subfamily models (patch X in Fig. 5 *A*). It should be emphasized that contacts in patch X formed as a consequence of DCA pairs that were constrained to form contacts. This patch formed the interface between β5-α5 and DNA recognition helix α9 (patch X in Fig. 5

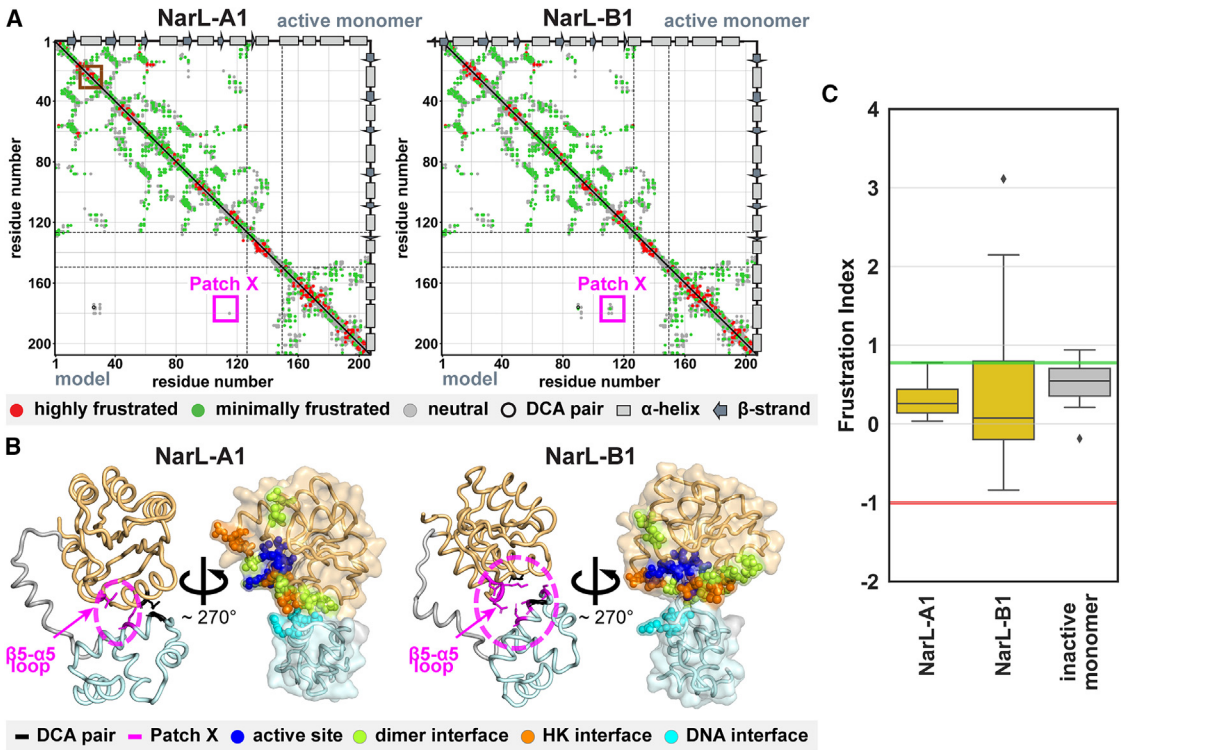

FIGURE 5 NarL subfamily models. (*A*) Contact maps of the model (*bottom triangle*) and the active monomer template structure (*upper triangle*) colored by the level of local conformational frustration. Residue number follows the reindexed template structure. The brown patch shows less frustrated α1 in our model. A conserved inter-domain contact patch in OmpR subfamily, patch X, was observed in NarL models as well. (*B*) Model structures showing DCA pairs and residues involved in patch X (*left*) and functional sites (*right*). DNA interface was less exposed in our models. (*C*) Boxplots of frustration index of inter-domain residue contacts. Data points above the green line are minimally frustrated and those below the red line are highly frustrated. The new inter-domain interfaces in NarL-A1 and NarL-B1 were comparable to that of reference inactive monomer structure.

TABLE 2 SASA of functional sites in NarL subfamily models

Functional sites	Active monomer (Å^2)	NarL-A1 (Å^2)	NarL-A1 (%)	NarL-B1 (Å^2)	NarL-B1 (%)
Active site	96.75	203.98	210.8	97.56	100.8
Dimer interface	1201.86	949.49	82.3	949.02	79.0
HK interface	689.69	605.84	86.7	480.37	68.8
DNA interface	264.70	163.60	61.8	196.13	74.1

The % symbol denotes the ratio of the SASA of our computational model to the SASA of the active monomer as a percentage.

B). No statistically significant differences were shown at significance level of 0.05 by Mann-Whitney U test between the inter-domain contacts of our models and the reference closed inactive monomer structure (*p*-value for NarL-A1, 0.273; *p*-value for NarL-B1, 0.079; Fig. 5 *C*). This suggests that the inter-domain interfaces of our models were energetically comparable to the known inter-domain interface, supporting the feasibility of our NarL subfamily models.

Similar to OmpR subfamily models, in both NarL models, the REC domain residues of patch X included at least one of K_A, $K_A + 1$, and $K_A + 4$ (Table S7). In NarL-A1, simultaneous inter-domain residue interactions of DNA recognition helix to α1 and patch X were also observed. Applying the same discussion on OmpR subfamily models, these inter-domain residue interactions may be a molecular mechanism to optimize phosphorylation duration of NarL subfamily proteins, perhaps through a conserved mechanism that is shared between OmpR and NarL subfamilies. In NarL-B2, one of the residue interaction partners of DNA recognition helix was replaced from α1 to β4-α4 loop. However, the contact in patch X still involved REC domain residue, which was reported to regulate reaction kinetics of RR, i.e., a residue next to the conserved threonine/serine (60). These interactions imply that inter-domain contacts in NarL-B1 should be important to regulate reaction kinetics of NarL subfamily proteins.

More than 30% loss of SASA in NarL-B1 at HK interface suggests that NarL-B1 is less likely to bind to HK (Table 2). Likewise, loss of at least 25% of SASA in DNA interface suggests that the both NarL subfamily models are less likely to bind to DNA. This tendency was conserved in OmpR subfamily models as well. The discussions above lead us to consider that the noncontacting inter-domain DCA pairs in NarL subfamily are involved in potential closed inactive monomer conformations of NarL subfamily DBRRs.

CONCLUSIONS

We applied DCA on DBRR protein sequence data to thoroughly investigate linker-domain and inter-domain residue interactions for OmpR and NarL subfamilies, the two main subdivisions of DBRR proteins. Our analysis of the individual subfamilies further demonstrated that coevolutionary analysis can detect subtle evolutionary divergences within protein families at the multi-domain level, provided that there are sufficient sequence data to subdivide the family (as is the case for DBRRs). Not only did we detect subfamily-specific contacts within the REC domains but we also found strong linker-domain and REC-EFF domain residue coevolution. The linker-REC domain residue coevolution was consistent with the residue contacts that may be involved in the cooperative functional regulation of the EFF domain.

To examine the applicability of our coevolutionary models on the prediction of biological functionality, we constructed a mutational landscape from our inferred DCA couplings. We applied these mutational landscapes to experiments that replaced the native C-terminal portion of DBRR proteins with one from a nonnative DBRR protein (via domain swapping). Domain-swapped proteins pose a significant challenge for our coevolutionary models, which have previously been used to examine the effects of up to four mutations (24). Nevertheless, we demonstrate that the model offers a high negative predictive value, which can potentially be used to screen out DBRR proteins that are nonfunctional. Our proof-of-concept prediction model will provide a basis for the further development of the coevolution-based prediction models of domain-swapped DBRR functionality.

Finally, we extensively explored inter-domain residue coevolution between the REC and EFF domains. Our analyses detected nontrivial inter-domain residue coevolution in both subfamilies. Most of the inter-domain DCA pairs between α3-β4-α4 vs. DNA recognition helix coincided with residue contacts from known inactive monomer structures. We additionally identified unverified DCA pairs that were highly coevolving DBRR residue pairs that were not observed in any experimental structures between α1 vs. DNA recognition helix. We explored the possibility that these unverified DCA pairs made spatial contacts by simulating them as contacts in the structure-based model. Our findings suggest that these residue pairs potentially capture inactive monomer conformations where DNA binding is inhibited. Interestingly, all of our DCA-guided monomer models showed additional inter-domain contacts, which may facilitate optimization of the lifetime of active DBRRs to the biological output of the DBRRs. Our results suggest that the same functional mechanism of inactivation is encoded in OmpR and NarL classes of DBRRs. Although this particular mechanism of inactivation has not yet been verified experimentally, this work suggests its existence through a combination of coevolutionary analysis, molecular simulation, and frustration analysis.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2024.01.028.

AUTHOR CONTRIBUTIONS

M.S., X.L., J.N.O., and R.R.C. designed the project. M.S. conducted the analysis. J.N.O., K.Y., and R.R.C. provided the materials. M.S., X.L., J.N.O., K.Y., and R.R.C. prepared the manuscript.

ACKNOWLEDGMENTS

Part of this work was done when M.S., X.L., and R.R.C. were at the Center for Theoretical Biological Physics (CTBP). Work at CTBP was supported by the NSF grants PHY-2019745 and PHY-2210291. M.S. thanks TOMODACHI-STEM Women's Leadership and Research Program and Tobitate! (Leap for Tomorrow) Study Abroad Initiative for traveling. J.N.O. is a CPRIT Scholar in Cancer Research sponsored by the Cancer Prevention and Research Institute of Texas. This work was supported in part, by JST, the establishment of university fellowships toward the creation of science technology innovation, grant number JPMJFS2113 and by Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS) (JP23ama121055) from Japan Agency for Medical Research and Development (AMED). The calculation in this research was partly conducted on Chaen, the supercomputer of the Center for Interdisciplinary AI and Data Science at Ochanomizu University.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPORTING CITATIONS

References (62–73) appear in the Supporting Material.

REFERENCES

1. Hoch, J. A. 2000. Two-component and phosphorelay signal transduction. *Curr. Opin. Microbiol.* 3:165–170.
2. Galperin, M. Y. 2010. Diversity of structure and function of response regulator output domains. *Curr. Opin. Microbiol.* 13:150–159.
3. Gao, R., T. R. Mack, and A. M. Stock. 2007. Bacterial response regulators: versatile regulatory strategies from common domains. *Trends Biochem. Sci.* 32:225–234.
4. Gao, R., S. Bouillet, and A. M. Stock. 2019. Structural basis of response regulator function. *Annu. Rev. Microbiol.* 73:175–197.
5. Volkman, B. F., D. Lipson, …, D. Kern. 2001. Two-state allosteric behavior in a single-domain signaling protein. *Science.* 291:2429–2433.
6. Corrêa, F., and K. H. Gardner. 2016. Basis of mutual domain inhibition in a bacterial response regulator. *Cell Chem. Biol.* 23:945–954.
7. Barbieri, C. M., T. R. Mack, …, A. M. Stock. 2010. Regulation of response regulator autophosphorylation through interdomain contacts. *J. Biol. Chem.* 285:32325–32335.
8. Barbieri, C. M., T. Wu, and A. M. Stock. 2013. Comprehensive analysis of OmpR phosphorylation, dimerization, and DNA binding supports a canonical model for activation. *J. Mol. Biol.* 425:1612–1626.
9. Schmidl, S. R., F. Ekness, …, J. J. Tabor. 2019. Rewiring bacterial two-component systems by modular DNA-binding domain swapping. *Nat. Chem. Biol.* 15:690–698.
10. Weigt, M., R. A. White, …, T. Hwa. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. USA.* 106:67–72.
11. Morcos, F., A. Pagnani, …, M. Weigt. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA.* 108:E1293–E1301.
12. Krepel, D., R. R. Cheng, …, J. N. Onuchic. 2018. Deciphering the structure of the condensin protein complex. *Proc. Natl. Acad. Sci. USA.* 115:11911–11916.
13. Dos Santos, R. N., F. Morcos, J. N. Onuchic…, 2015. Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci. Rep.* 5:13652.
14. Fantini, M., D. Malinverni, …, A. Pastore. 2017. New techniques for ancient proteins: direct coupling analysis applied on proteins involved in iron sulfur cluster biogenesis. *Front. Mol. Biosci.* 4:40.
15. Galaz-Davison, P., D. U. Ferreiro, and C. A. Ramírez-Sarmiento. 2022. Coevolution-derived native and non-native contacts determine the emergence of a novel fold in a universally conserved family of transcription factors. *Protein Sci.* 31:e4337.
16. Marks, D. S., T. A. Hopf, and C. Sander. 2012. Protein structure prediction from sequence variation. *Nat. Biotechnol.* 30:1072–1080.
17. Trinquier, J., G. Uguzzoni, …, M. Weigt. 2021. Efficient generative modeling of protein sequences using simple autoregressive models. *Nat. Commun.* 12:5800.
18. Malinverni, D., and A. Barducci. 2020. Coevolutionary analysis of protein subfamilies by sequence reweighting. *Entropy.* 21:1127.
19. Uguzzoni, G., S. John Lovis, …, M. Weigt. 2017. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc. Natl. Acad. Sci. USA.* 114:E2662–E2671.
20. Brüderlin, M., R. Böhm, …, B. N. Dubey. 2023. Structural features discriminating hybrid histidine kinase Rec domains from response regulator homologs. *Nat. Commun.* 14:1002.
21. Schug, A., M. Weigt, …, H. Szurmant. 2009. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl. Acad. Sci. USA.* 106:22124–22129.
22. Gueudré, T., C. Baldassi, …, A. Pagnani. 2016. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc. Natl. Acad. Sci. USA.* 113:12186–12191.
23. Cheng, R. R., F. Morcos, …, J. N. Onuchic. 2014. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. USA.* 111:E563–E571.
24. Cheng, R. R., O. Nordesjö, …, F. Morcos. 2016. Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. *Mol. Biol. Evol.* 33:3054–3064.
25. Cheng, R. R., E. Haglund, …, J. N. Onuchic. 2018. Designing bacterial signaling interactions with coevolutionary landscapes. *PLoS One.* 13, e0201734.
26. Figliuzzi, M., H. Jacquier, …, M. Weigt. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.* 33:268–280.
27. Hopf, T. A., J. B. Ingraham, …, D. S. Marks. 2017. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35:128–135.
28. Mistry, J., S. Chuguransky, …, A. Bateman. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49:D412–D419.
29. Maciunas, L. J., N. Porter, …, P. J. Loll. 2021. Structures of full-length VanR from *Streptomyces coelicolor* in both the inactive and activated states. *Acta Crystallogr. D Struct. Biol.* 77:1027–1039.
30. Park, A. K., J. H. Lee, …, H. Park. 2016. Structural characterization of the full-length response regulator spr1814 in complex with a phosphate analogue reveals a novel conformational plasticity of the linker region. *Biochem. Biophys. Res. Commun.* 473:625–629.
31. Paysan-Lafosse, T., M. Blum, …, A. Bateman. 2023. InterPro in 2022. *Nucleic Acids Res.* 51:D418–D427.
32. UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. 2021. *Nucleic Acids Res.* 49:D480–D489.
33. Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.
34. Eddy, S. R. 2011. Accelerated profile HMM searches. *PLoS Comput. Biol.* 7, e1002195.

35. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The protein data bank. *Nucleic Acids Res.* 28:235–242.

36. Ekeberg, M., C. Lövkvist, …, E. Aurell. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E.* 87, 012707.

37. Ekeberg, M., T. Hartonen, and E. Aurell. 2014. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.* 276:341–356.

38. Joosten, R. P., T. A. H. te Beek, …, G. Vriend. 2011. A series of PDB related databases for everyday needs. *Nucleic Acids Res.* 39:D411–D419.

39. Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22:2577–2637.

40. Webb, B., and A. Sali. 2016. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics.* 54:5.6.1–5.6.37.

41. Howell, A., S. Dubrac, …, K. Devine. 2003. Genes controlled by the essential YycG/YycF two-component system of *Bacillus subtilis* revealed through a novel hybrid regulator approach. *Mol. Microbiol.* 49:1639–1655.

42. Tapparel, C., A. Monod, and W. L. Kelley. 2006. The DNA-binding domain of the *Escherichia coli* CpxR two-component response regulator is constitutively active and cannot be fully attenuated by fused adjacent heterologous regulatory domains. *Microbiology.* 152:431–441.

43. Pronk, S., S. Páll, …, E. Lindahl. 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics.* 29:845–854.

44. Noel, J. K., M. Levi, …, P. C. Whitford. 2016. SMOG 2: a versatile software package for generating structure-based models. *PLoS Comput. Biol.* 12, e1004794.

45. Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–953.

46. Lammert, H., A. Schug, and J. N. Onuchic. 2009. Robustness and generalization of structure-based models for protein folding and function. *Proteins.* 77:881–891.

47. Rausch, A. O., M. I. Freiberger, …, R. G. Parra. 2021. FrustratometeR: an R-package to compute local frustration in protein structures, point mutants and MD simulations. *Bioinformatics.* 37:3038–3040.

48. Holm, L. 2022. Dali server: structural unification of protein families. *Nucleic Acids Res.* 50:W210–W215.

49. Vander Meersche, Y., G. Cretin, …, T. Galochkina. 2021. MEDUSA: prediction of protein flexibility from sequence. *J. Mol. Biol.* 433, 166882.

50. Park, A. K., J. H. Moon, …, Y. M. Chi. 2013. Crystal structure of the response regulator spr1814 from *Streptococcus pneumoniae* reveals unique interdomain contacts among NarL family proteins. *Biochem. Biophys. Res. Commun.* 434:65–69.

51. Fraczkiewicz, R., and W. Braun. 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* 19:319–333.

52. Mattison, K., R. Oropeza, and L. J. Kenney. 2002. The linker region plays an important role in the interdomain communication of the response regulator OmpR. *J. Biol. Chem.* 277:32714–32721.

53. Walthers, D., V. K. Tran, and L. J. Kenney. 2003. Interdomain linkers of homologous response regulators determine their mechanism of action. *J. Bacteriol.* 185:317–324.

54. Friedland, N., T. R. Mack, …, A. M. Stock. 2007. Domain orientation in the inactive response regulator *Mycobacterium tuberculosis* MtrA provides a barrier to activation. *Biochemistry.* 46:6733–6743.

55. Yan, H., Q. Wang, …, X. Li. 2019. The DNA-binding mechanism of the TCS response regulator ArlR from *Staphylococcus aureus*. *J. Struct. Biol.* 208, 107388.

56. Ouyang, Z., F. Zheng, …, Y. Wen. 2019. Deciphering the activation and recognition mechanisms of *Staphylococcus aureus* response regulator ArlR. *Nucleic Acids Res.* 47:11418–11429.

57. Ferreiro, D. U., J. A. Hegler, …, P. G. Wolynes. 2007. Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. USA.* 104:19819–19824.

58. Freiberger, M. I., A. B. Guzovsky, …, D. U. Ferreiro. 2019. Local frustration around enzyme active sites. *Proc. Natl. Acad. Sci. USA.* 116:4037–4043.

59. Lukat, G. S., B. H. Lee, …, J. B. Stock. 1991. Roles of the highly conserved aspartate and lysine residues in the response regulator of bacterial chemotaxis. *J. Biol. Chem.* 266:8348–8354.

60. Straughn, P. B., L. R. Vass, …, R. B. Bourret. 2020. Modulation of response regulator CheY reaction kinetics by two variable residues that affect conformation. *J. Bacteriol.* 202:e00089-20–e00020.

61. Foster, C. A., R. E. Silversmith, …, R. B. Bourret. 2021. Role of position K+4 in the phosphorylation and dephosphorylation reaction kinetics of the CheY response regulator. *Biochemistry.* 60:2130–2151.

62. Bryngelson, J. D., and P. G. Wolynes. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA.* 84:7524–7528.

63. Onuchic, J. N., and P. G. Wolynes. 2004. Theory of protein folding. *Curr. Opin. Struct. Biol.* 14:70–75.

64. Leopold, P. E., M. Montal, and J. N. Onuchic. 1992. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. USA.* 89:8721–8725.

65. Sinner, C., B. Lutz, …, A. Schug. 2014. Simulating biomolecular folding and function by native-structure-based/Go-type models. *Isr. J. Chem.* 54:1165–1175.

66. Bruno da Silva, F., V. G. Contessoto, …, V. B. P. Leite. 2018. Nonnative cooperative interactions modulate protein folding rates. *J. Phys. Chem. B.* 122:10817–10824.

67. Noel, J. K., J. I. Sułkowska, and J. N. Onuchic. 2010. Slipknotting upon native-like loop formation in a trefoil knot protein. *Proc. Natl. Acad. Sci. USA.* 107:15403–15408.

68. Kaya, H., and H. S. Chan. 2003. Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? *J. Mol. Biol.* 326:911–931.

69. Chu, X., Z. Suo, and J. Wang. 2022. Investigating the conformational dynamics of a Y-Family DNA polymerase during its folding and binding to DNA and a nucleotide. *JACS Au.* 2:341–356.

70. Noel, J. K., J. Chahine, …, P. C. Whitford. 2014. Capturing transition paths and transition states for conformational rearrangements in the ribosome. *Biophys. J.* 107:2881–2890.

71. Zhao, L., H. P. Lu, and J. Wang. 2018. Exploration of multistate conformational dynamics upon ligand binding of a monomeric enzyme involved in pyrophosphoryl transfer. *J. Phys. Chem. B.* 122:1885–1897.

72. Morcos, F., B. Jana, …, J. N. Onuchic. 2013. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. USA.* 110:20533–20538.

73. Krishnamohan, A., G. L. Hamilton, …, F. Morcos. 2023. Coevolution and smFRET Enhances Conformation Sampling and FRET Experimental Design in Tandem PDZ1-2 Proteins. *J. Phys. Chem. B.* 127:884–898.