THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Expanding Direct Coupling Analysis to Identify Heterodimeric Interfaces from Limited Protein Sequence Data

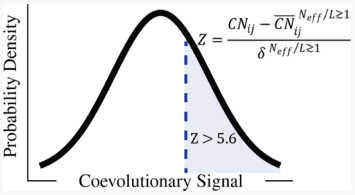
Kareem M. Mehrabiani, Ryan R. Cheng, and José N. Onuchic*

Cite This: *J. Phys. Chem. B* 2021, 125, 11408–11417

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Direct coupling analysis (DCA) is a global statistical approach that uses information encoded in protein sequence data to predict spatial contacts in a three-dimensional structure of a folded protein. DCA has been widely used to predict the monomeric fold at amino acid resolution and to identify biologically relevant interaction sites within a folded protein. Going beyond single proteins, DCA has also been used to identify spatial contacts that stabilize the interaction in protein complex formation. However, extracting this higher order information necessary to predict dimer contacts presents a significant challenge. A DCA evolutionary signal is much stronger at the single protein level (intraprotein contacts) than at the protein−protein interface (interprotein contacts). Therefore, if DCA-derived information is to be used to predict the structure of these complexes, there is a need to identify statistically significant DCA predictions. We propose a simple *Z*-score measure that can filter good predictions despite noisy, limited data. This new methodology not only improves our prediction ability but also provides a quantitative measure for the validity of the prediction.



INTRODUCTION

Reliable structural models of proteins are crucial for understanding the structure-to-function relationship of those proteins in living systems. To this end, experimental approaches such as X-ray crystallography and NMR have successfully produced high-resolution structures of hundreds of proteins, which are readily available in the Protein Data Bank (PDB).[1] Experimental structures, however, are only available for a small fraction of the known proteins of interest.[1] This has led to a significant demand for other information sources that can be exploited to predict structures of proteins and protein−protein complexes. Over the past decade, statistical methodologies[2−19] have been developed that utilize information encoded in protein sequence data to identify spatial contacts present in a three-dimensional protein structure. These methods quantify observed correlations between amino acid identities at different positions (termed protein coevolution) and use this information to infer structural contacts. The key assumption to these sequence-based approaches is that residue coevolution arises from a strong evolutionary constraint to preserve, on average, stable protein interactions over the course of natural selection.

Direct Coupling Analysis (DCA) has had remarkable success in predicting structural contacts in the 3D protein structure by quantifying amino acid coevolution (covariation) within a protein family from their amino acid sequences.[6−10,14,20] It has also been shown that DCA and related approaches can predict spatial contacts that stabilize the interaction between two proteins[5,21−33,57] and can identify the configurational diversity inherent in some protein families[26,34,35,57] Success in identifying interprotein contact prediction is often hampered by weaker interprotein signals compared to intraprotein signals.[36] In addition, only limited sequence data are available for pairs of proteins.

Historically, the degree of coevolution between pairs of residues has been expressed using metrics such as Direct Information[5,8] or Frobenius norm[14,37] where larger values are related to greater amounts of coevolution. Highly coevolving amino acid residue pairs are likely to be in spatial proximity within a 3D structure of a protein or protein complex. Contact pair predictions are ranked by these metrics, and the top-coevolving residue pairs (e.g., top ten) are typically treated as the predictions used in structural determination. As commented above, the prediction of dimer contacts from sequence coevolution is generally more challenging than the prediction of monomeric contacts. On average, monomeric coevolutionary signals are much stronger than dimeric signals, reflecting the importance of preserving monomeric folds over natural selection. This difficulty is exemplified in Figure 1, which shows, as an example, the DCA predictions for the SigK-RskA[38] dimer. Hence, there is a need to quantitatively assess the statistical significance of DCA predictions for dimer contacts, which are often noisy and made using limited data. To address this challenge, we propose a *Z*-score threshold to validate these predictions, which is a statistical metric that measures the relationship between a set of data points and a reference

Received: August 12, 2021
Revised: September 24, 2021
Published: October 7, 2021

https://doi.org/10.1021/acs.jpcb.1c07145
J. Phys. Chem. B 2021, 125, 11408−11417

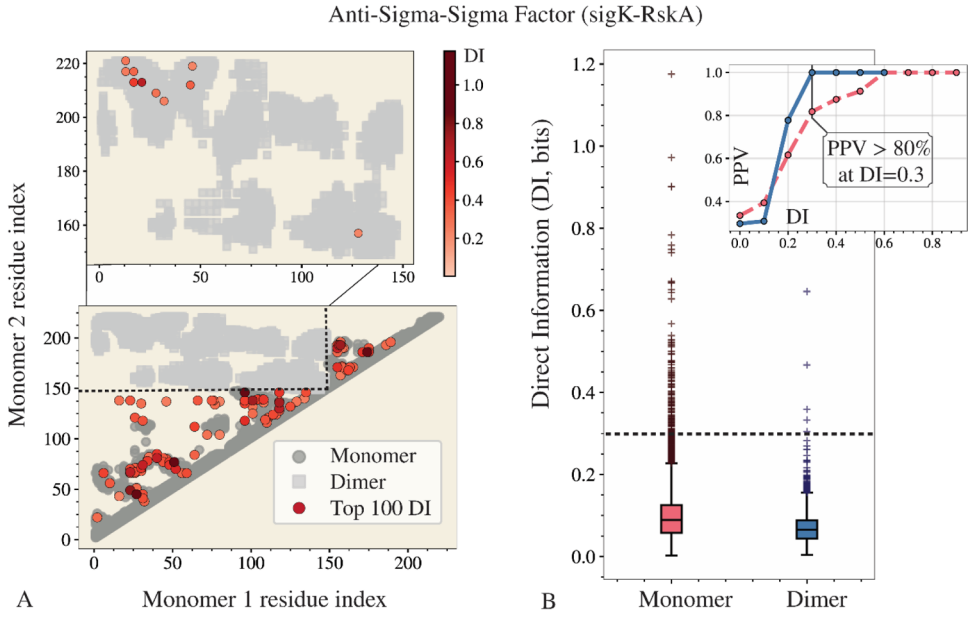


Figure 1. Comparison of DCA coupling strengths for the monomer and dimer signal for MSA with sufficient sequence statistics. We examined the sigma-anti-sigma SigK-RskA dimer (PDB ID: 4NQW; resolution: 2.4 Å). The amount of nonredundant sequence information within a multiple-sequence alignment (MSA) is known to have a large influence on the DCA prediction accuracy. One coarse measure of this sequence information is given by N_{eff}/L, which describes the effective number of sequences over the sequence length L of the MSA (See Methods for more details). MSAs for which $N_{eff}/L \gg 1$ are generally considered to have ample sequence statistics; for SigK-RskA, this measure is $N_{eff}/L = 5.76$. (A) Top 100 Mean-Field DCA residue pair predictions are pictured as red-colored circles; a darker shade indicates increasing direct information. DI (calculated in bits) contains information for both monomeric (intraprotein) and dimeric (interprotein) structural contacts. The dimeric predictions are plotted separately at the top panel for clarity. The PDB crystal contacts for the monomer (dark-gray circles) and dimer interface (light-gray squares) are plotted to compare the DCA predictions against the crystal structure contacts. We have reindexed the PDB residue indices, such that the first monomer begins from 1 and ends at N and the second monomer from $N + 1$ to $M + N$, where N and M are the lengths of monomer one and two, respectively. (B). DI values are shown for the monomeric contacts and for the interface contacts. Recall that DI measures the degree by which amino acid pairs exhibit correlated amino acid identities over natural selection, which often reveals their spatially proximity in the 3D space within a protein structure or set of structures. Above a DI of 0.3 (black-colored dashed line), the positive predictive value, defined as $\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$, is 100% for dimeric predictions (blue-colored line in the inset) and 80% for monomeric predictions (pink-colored dashed line in the inset; TP is the true positive count and FP is the false positive count). Notice that the DI scores for the monomeric contacts are more likely to be above this threshold than dimeric ones. This likely reflects how the maintenance of the protein fold is apparently a stronger evolutionary constraint on sequence selection than the maintenance of dimeric interfaces.

distribution in units of standard deviation. This measure is applied to examine the statistical significance of DCA predictions from 76 dimer systems using DCA information.

METHODS

Paired Multiple Sequence Alignments were Downloaded from the EVcoupling Database. Extending DCA to predict heterodimers involves concatenating two interacting families together, after determining which individual sequences interact by genomic adjacency. For bacterial proteins, this task is simpler because interacting sequences mostly share the same operon.[21,22] This task is more challenging for eukaryotes where tools such as iterative pair-matching algorithms[39,40] and highest reciprocal identity methods[29,41] are utilized to determine these matches. Interestingly, predicting homodimer contacts does not involve identifying protein partners but has the challenge that interchain contacts may also exist as intrachain contacts.[36] MSAs of 561 concatenated protein family pairs were downloaded from the EVcoupling heterodimeric database,[29] which uses the highest reciprocal identity approach. Of the 561 MSAs available in the data set, we selected the 76 MSAs, which have a corresponding dimeric biological unit in an available PDB structure. Complexes involving more than two proteins were excluded from our analysis. The PDB structures all have resolutions of at least 3 Å.

MSAs Classified by the Amount of Unique Sequence Information. The number of effective sequences of an MSA, N_{eff}, is the number of nonredundant sequences $N_{eff} = \sum_{k=1}^{N} w_k$, where N is the total number of MSA sequences and w_k is the reciprocal of the number of sequences within the MSA that are similar to sequence k, defined by a threshold of 80% or more amino acid identity (see refs8 37, for additional details). The ratio N_{eff}/L, where L is the length of a sequence in the MSA, is often a useful metric for characterizing the quantity of sequence data available for a given system.[21] Therefore, it is used for an initial determination if sufficient information is available to predict protein contacts using coevolutionary inference techniques. MSAs for which $N_{eff}/L \gg 1$ are typically considered to have ample sequence statistics, while MSAs for which $N_{eff}/L \ll 1$ have insufficient sequence statistics. For simplicity, we divided the 76 MSAs into 2 categories: $N_{eff}/L \gtrsim 1$ (n = 28 sequences) and $N_{eff}/L \lesssim 1$ (n = 48 sequences).

Quantifying Coevolutionary Information between Residue Pairs. Highly coevolving residue pairs are presumed to be in spatial proximity within a 3D structure of a protein or protein complex. The coevolutionary residue couplings for each system were calculated using plmDCA,[14,37] which uses pseudolikelihood maximization to infer a statistical model that is consistent with the amino acid correlations observed in the MSA data. The degree of coevolution between pairs of residues

https://doi.org/10.1021/acs.jpcb.1c07145
J. Phys. Chem. B 2021, 125, 11408−11417

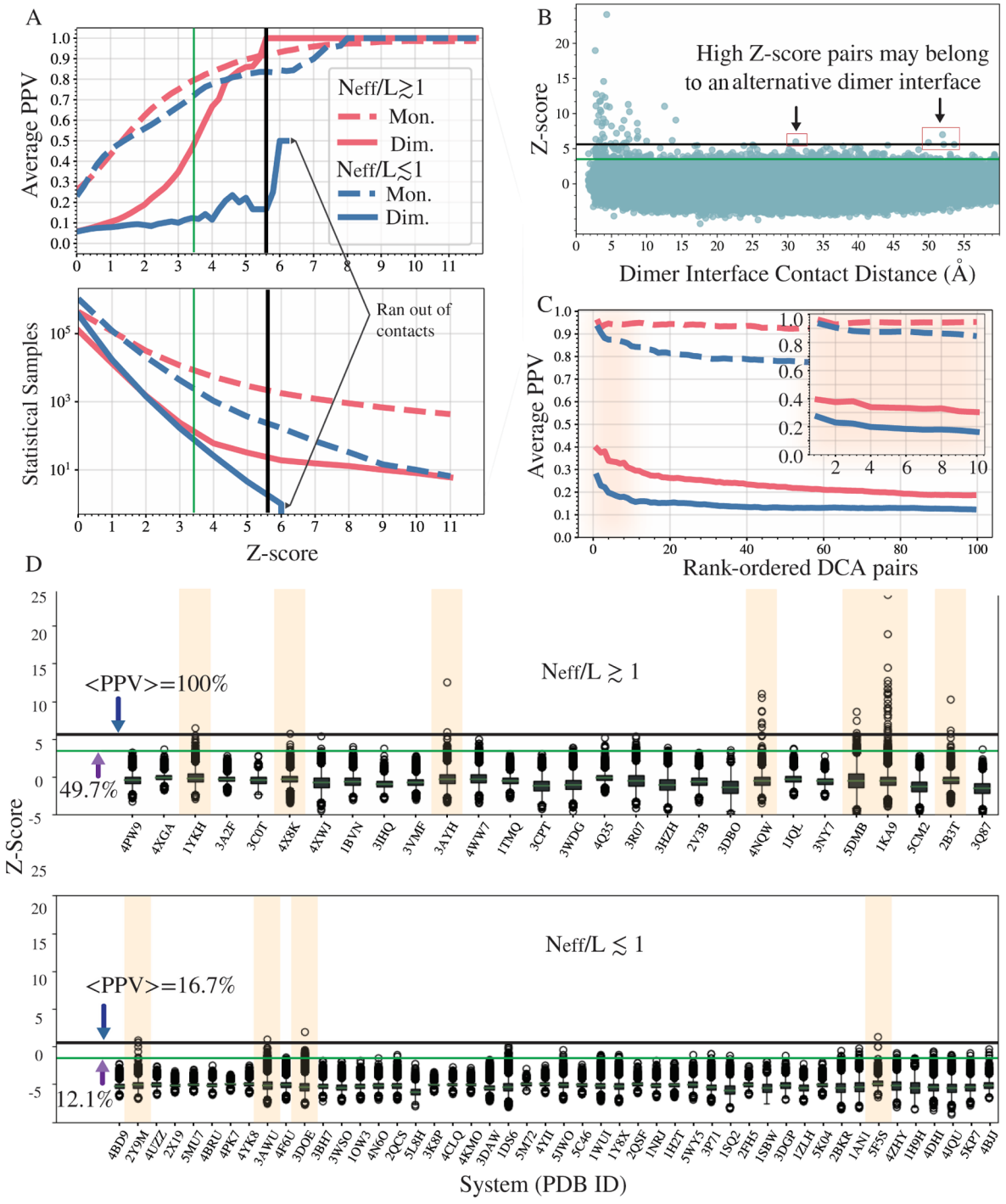


Figure 2. Applying *Z*-score analysis to DCA analysis to filter out statistically insignificant predictions. MSAs of 561 concatenated protein family pairs were downloaded from the EVcoupling database. After filtering for 76 MSAs, which have a corresponding dimeric biological unit in an available PDB structure, we categorized the remaining MSAs into two groups based on their effective number of sequences per length: dimers with sufficient sequence data ($N_{eff}/L \gtrsim 1$) and limited data ($N_{eff}/L \lesssim 1$). (A) Average PPV (top panel) and the number of predicted dimer contacts (bottom panel) as a function of the *Z*-score are shown for the two groups $N_{eff}/L \gtrsim 1$ (pink-colored lines) and $N_{eff}/L \lesssim 1$ (blue-colored lines) and for each group's monomeric (dashed lines) and dimeric predictions (solid lines). For the $N_{eff}/L \gtrsim 1$ systems, at $Z = 3.5$ (green-colored vertical line), the average PPV for true dimer predictions is 49.7%, which increases to 100% at $Z = 5.6$ (black vertical line). (B) *Z*-score values of all DCA dimer predictions and their corresponding residue−residue distances in the PDB crystal structure. There are 358 dimer pair predictions above $Z = 3.5$ and 49 of those are above $Z = 5.6$. Note that at $Z = 5.6$, there are four predictions around 50 Å and one at 30 Å that come from the $N_{eff}/L \lesssim 1$ systems (red-colored rectangles). These long-range contacts currently contribute to the dimer pair false positive count in (A) They may, however, be associated with alternative dimer structures (see discussion in the Results section). (C) Average PPV as a function of the descending rank order of DCA predictions for monomeric predictions (dashed lines) and dimeric predictions (solid lines). The pink lines plotted are the average PPV for systems with $N_{eff}/L \gtrsim 1$, whereas the blue lines correspond to the average PPV for the $N_{eff}/L \lesssim 1$ systems. The average PPV for the top ten predictions is 94.4% (for $N_{eff}/L \gtrsim 1$) and 84.6% (for $N_{eff}/L \lesssim 1$) for monomeric predictions and 30.3% (for $N_{eff}/L \gtrsim 1$) and 16.5% (for $N_{eff}/L \lesssim 1$) for the dimeric predictions. As illustrated in (A), a *Z*-score threshold offers significant improvements in prediction accuracy compared to simply selecting the top rank-order predictions. (D) *Z*-scores of

https://doi.org/10.1021/acs.jpcb.1c07145
J. Phys. Chem. B 2021, 125, 11408−11417

Figure 2. continued

the top coevolving residue pairs for the two groups of MSAs—$N_{eff}/L \gtrsim 1$ top panel and $N_{eff}/L \lesssim 1$ bottom panel. The x-axis is labeled by PDB ID. The systems with $N_{eff}/L \gtrsim 1$ ($n = 28$) were used as the reference distribution. A dimer contact is defined if they are separated by less than 15 Å between two interchain residue heavy atoms. This choice allows for a flexible refining process using molecular dynamics simulations. Two lines are drawn at $Z = 5.6$ (black line) and $Z = 3.5$ (green-colored line), and the arrows show the average PPV at these points. We observed seven systems in the $N_{eff}/L \gtrsim 1$ group to have an average PPV of 100% at $Z = 5.6$ and an additional ten systems to have an average PPV > 49.7% at a threshold $Z = 3.5$.

was expressed using the Frobenius norm of the coupling matrix[14,37] where larger values are related to greater amounts of coevolution. The Frobenius norm score is given by

$$FN_{ij} = \left\|\hat{J}_{ij}\right\|_2 = \sqrt{\sum_{\substack{k,l=1 \\ k,l\neq \text{gap}}}^{q} \hat{J}_{ij}(k,l)^2} \tag{1}$$

was calculated from the DCA coupling matrix $\hat{J}_{ij}$ after which an average product correction was applied to remove entropic and phylogenetic effects,[42] resulting in the so-called corrected-norm (CN-score) $\mathrm{CN}_{ij} = \mathrm{FN}_{ij} - \frac{\mathrm{FN}_{i:}\mathrm{FN}_{:j}}{\mathrm{FN}_{::}}$, where $FN_{:j}$ and $FN_{i:}$ are averaged over columns i and j, respectively. $\mathrm{FN}_{::}$ is the average value of the entire matrix. We could similarly quantify the amount of coevolutionary information using the DI,[5,8] which is a Kullback-Leibler[43] divergence between the inferred global statistical model of coevolving amino-acid pairs, $P_{ij}^{(\mathrm{DCA})}$, and a pair-independent model consisting of the product of the single site occurrence of a particular amino-acid A_i and A_j

$$\mathrm{DI}_{ij} = \sum_{A_i,A_j} P_{ij}^{(\mathrm{DCA})}(A_i,A_j)\ln\frac{P_{ij}^{(\mathrm{DCA})}(A_i,A_j)}{f_i(A_i)f_j(A_j)} \tag{2}$$

Equation 2 measures the amount information encoded in $P_{ij}^{(\mathrm{DCA})}$ relative to a null model where the residue sites i and j are statistically independent (no coevolution); hence, $\mathrm{DI}_{ij} = 0$ when $P_{ij}^{(\mathrm{DCA})}$ can be factorized into independent single-site frequencies.[8]

Determining Statistical Significance of Residue Coupling Strength Using the *Z*-Score Analysis. We defined a dimer contact to be less than 15 Å between two interchain residue heavy atoms, although we also examined contact definitions of 8 and 10 Å shown in the Supporting Information Figure S1. This broad definition of a contact accommodates the subtle conformational rearrangements that occur within a folded protein and can further be refined by combining DCA predictions as constraints in molecular dynamics simulation (see subsection Using DCA predictions as docking constraints). The CN-scores of these predictions were used for the *Z*-score calculation. The *Z*-score is defined as

$$Z_{ij} = \frac{\mathrm{CN}_{ij} - \overline{\mathrm{CN}}_{ij}^{N_{\mathrm{eff}}/L\gtrsim 1}}{\delta^{N_{\mathrm{eff}}/L\gtrsim 1}} \tag{3}$$

where $\overline{\mathrm{CN}}_{ij}^{N_{\mathrm{eff}}/L\gtrsim 1}$ and $\delta^{N_{\mathrm{eff}}/L\gtrsim 1}$ are the average and standard deviation of the CN-score, respectively. The reference distribution combines the DCA predictions for the monomeric contacts for sequences with $N_{eff}/L \gtrsim 1$ ($n = 28$) (e.g., systems with sufficient statistics). The *Z*-score analysis provides the appropriate metric to quantify the statistical significance of any specific prediction relative to the chosen reference distribution. A similar *Z*-score analysis has been applied to analyze phylogenic correlations within a protein family inferred using DCA.[44]

Validation of Residue Coupling *Z*-Scores Using the PPV. The total number of true positives (TP) and false positives (FP) were determined for all 76 sequences described above and used to calculate their PPV, defined as $\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$. PPV estimates the fraction of true contacts (defined by the PDB structure) out of the top n predicted contacts .

Using DCA Predictions as Docking Constraints. Previous studies have demonstrated how predicted DCA dimer contacts can be used to dock monomer structures to obtain a predicted dimer[21,23,29,45] or multimeric complex.[25] We will primarily adopt the docking protocol of dos Santos et al.[23] The monomeric structures are typically available on the PDB or constructed using homology modeling from existing structural data.[43,46−48] Docking is then performed by simulating a coarse-grained structure-based model[49,50] representation of the monomeric subunits and interaction between the subunits; an attractive potential is provided between the interprotein residue pairs predicted by DCA to be in contact. This docking procedure is discussed in greater detail in the Supporting Information for a specific illustrative example. This docked model can be further refined using explicit solvent molecular dynamics.[51] The resulting structural model can then be compared to known crystal structures[52] or, if none exists, validated by analyzing the interaction energetics of the dimer interface of the predicted protein complex. This is accomplished using frustration analysis,[53−55] which can be used to assess how energetically favorable or unfavorable a particular structure is.

RESULTS AND DISCUSSION

DCA uses coevolution information to predict not only spatial contacts in a monomeric protein[6−10,14] but also dimeric contacts that stabilize the interaction between protein complexes.[21,23,29,56] Figure 1 shows representative DCA predictions for the sigma factor SigK-RskA dimer,[57] where the top 100 DCA predicted contacts (containing both intra- and interprotein pairs) are plotted alongside the known crystal structure contacts. As discussed in Figure 1 and in the Introduction, the coevolutionary couplings between intraprotein residues are on average stronger than those for interprotein residue pairs. Ordinarily, the strongest coevolutionary couplings reflect contacts on the monomeric folds, while dimeric contacts can be viewed as a perturbative effect.

Historically, DCA predictions have been ranked in the descending order starting from the highest coevolving signal and the top ranked predictions are taken to represent structural contacts. However, selecting dimeric contact predictions in this manner can be problematic. The weakness of dimeric coevolutionary signals relative to monomeric signals makes dimer predictions more sensitive to noise associated with limited data. Therefore, a more rigorous approach is needed to determine if these predictions are statistically significant and meaningful. Toward this goal, we propose a quantitative approach to establish a threshold below which all DCA predictions should be excluded because no sufficient information is available. Figure

https://doi.org/10.1021/acs.jpcb.1c07145
J. Phys. Chem. B 2021, 125, 11408−11417

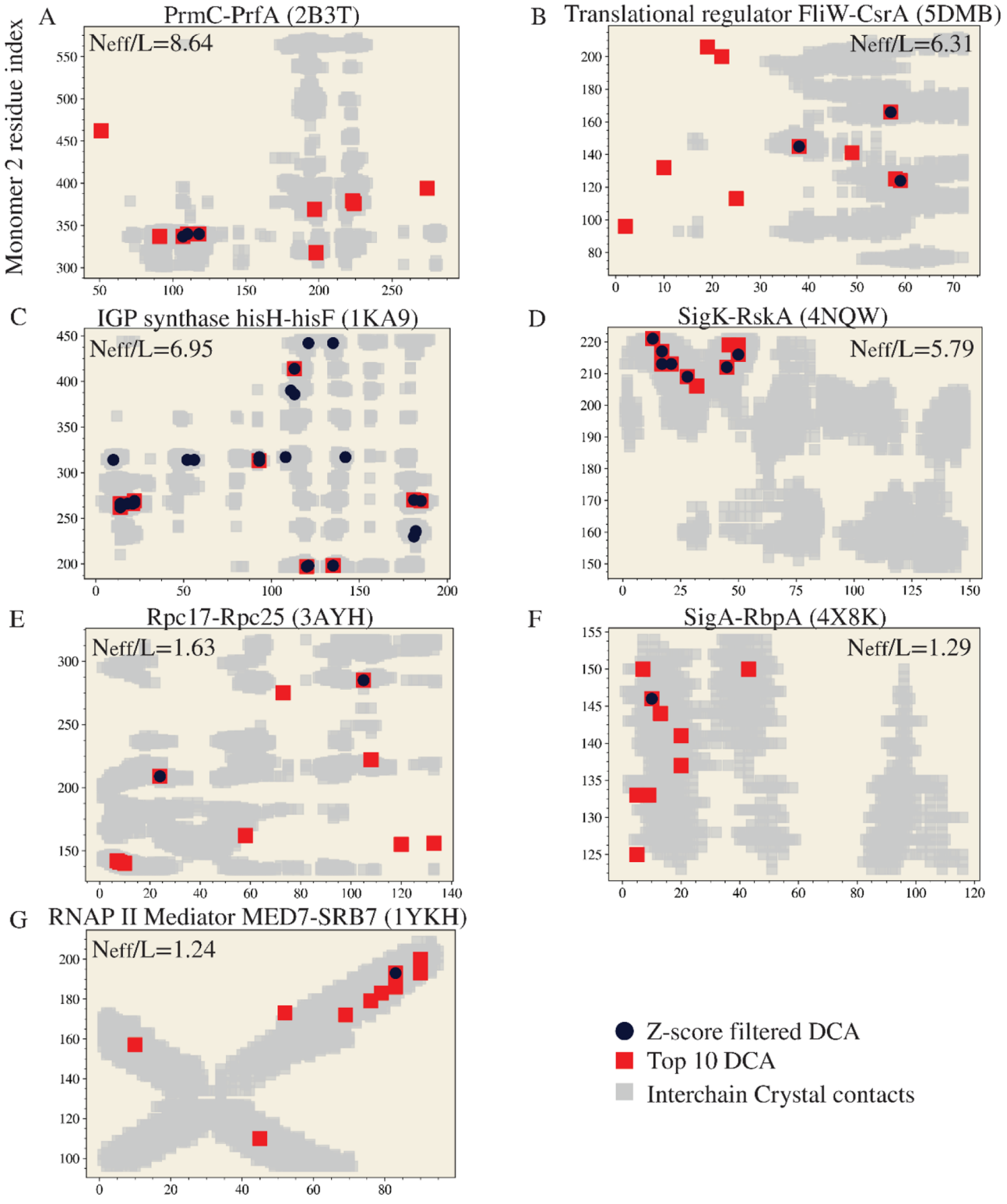


Figure 3. *Z*-score threshold removes incorrect DCA predictions for $N_{eff}/L \gtrsim 1$ systems with only true contacts for *Z*-scores above 5.6 (black circles). For comparison, the top ten DCA predictions are shown as red-colored squares and the crystal dimer contacts from the PDB are shown as gray-colored squares. (A–G). Seven out of the twenty-eight $N_{eff}/L \gtrsim 1$ dimers have true interface contacts above this threshold. False positive predictions are present when using the top ten predictions but were eliminated using this *Z*-score threshold. The remaining dimer systems in this group did not have any dimer predictions available (see Figure 2A bottom panel), but predictions at lower *Z*-score thresholds exist at the cost of a lower PPV, that is, not every DCA pair is a crystal interface contact.

1B illustrates the confidence level of DCA predictions at different values of the chosen threshold.

To quantitatively determine suitable thresholds to filter DCA predictions, a statistical significance measure using a *Z*-score analysis is implemented. As previously described, to test this

https://doi.org/10.1021/acs.jpcb.1c07145
J. Phys. Chem. B 2021, 125, 11408–11417

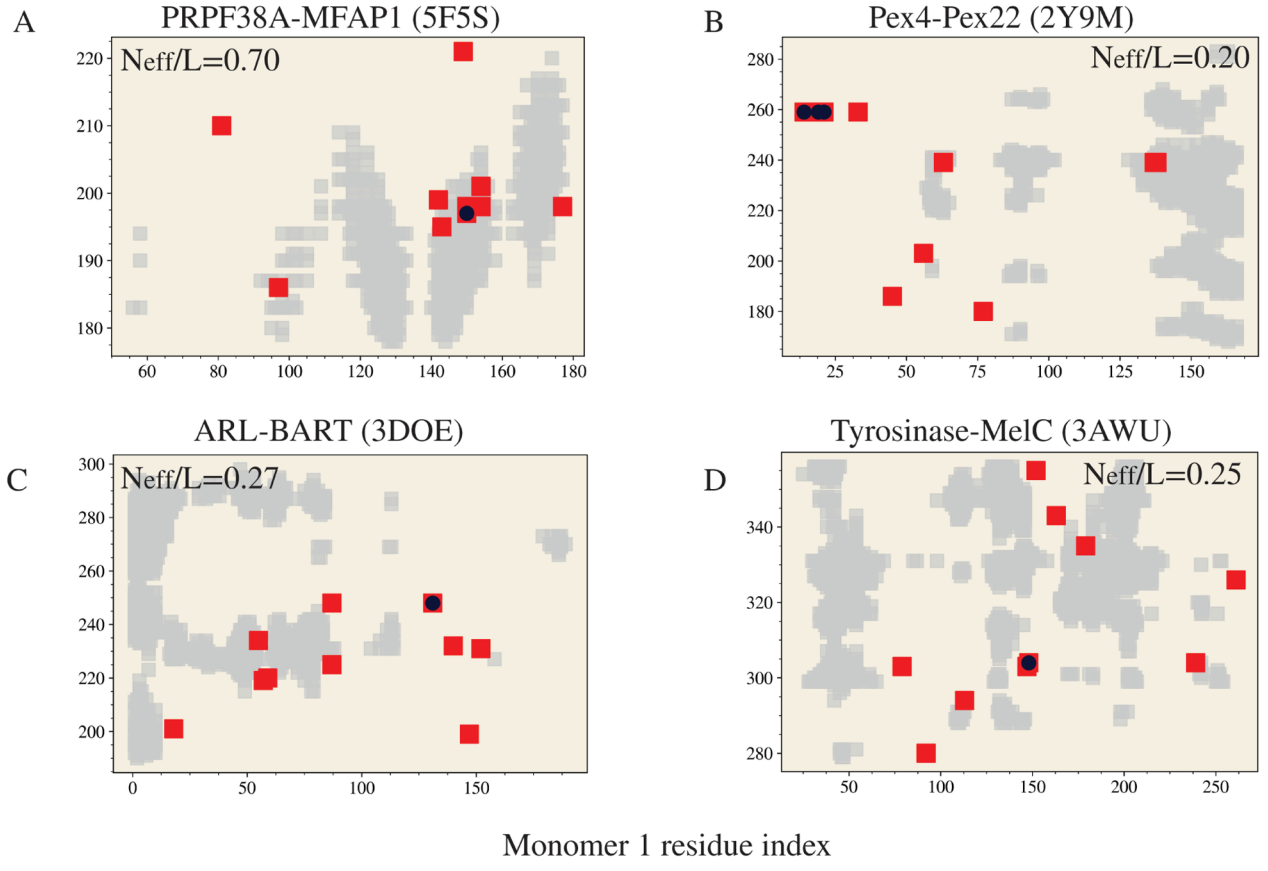


Figure 4. Applying *Z*-score threshold to systems with limited data ($N_{eff}/L \lesssim 1$). Using a *Z*-score threshold of 5.6 reveals that systems with limited data generally have no statistically significant predictions. (A) Only 1 of the 48 dimers has a true interface contact prediction (black circle) that overlaps with the crystal contact (gray squares). Interestingly, (B–D) three representative dimer systems that have high scoring contacts ($Z > 5.6$) are not in spatial proximity within a known dimeric structure. These predictions may capture alternative dimer structure because there is a significant high covariation signal revealed by using a *Z*-score threshold.

procedure, we investigated the 561 paired protein families from the EVcoupling heterodimeric webserver[29] and filtered sequence coevolution information for 76 families (presented inFigure 2), which were associated with independent dimer biological complexes in the PDB structure.

The strengths of the coevolutionary signals were determined using plmDCA. This information was used to validate the interprotein predictions using *Z*-score analysis, and we compared these results to the rank-ordered predictions traditionally used to predict the 3D structure. Figure 2A (top panel) shows the average PPV as a function of the *Z*-score (eq 3) for all dimer systems with sufficient data ($N_{eff}/L \gtrsim 1$; pink-colored lines) and systems with limited data ($N_{eff}/L \lesssim 1$; blue-colored lines). Again, this *Z*-score threshold, which quantifies the statistical significance of DCA-based predictions, is particularly needed for any dimer prediction. For more information on how the *Z*-score analysis is performed or how the two subgroups of dimers are selected (i.e., $N_{eff}/L \lesssim 1$ and $N_{eff}/L \gtrsim 1$); see the Materials and Methods section.

At a *Z*-score threshold of about 5.6 (shown as a black line in Figure 2A–D), we achieved a PPV of effectively 100%, meaning that all of the dimer predictions greater than or equal to this threshold are confirmed to be spatially proximal contacts within a known crystal structure. A comparative plot of PPV as a function of *Z*-score is shown in Figure S1 for different contact distances, showing that the general conclusions of Figure 2 are not sensitive to the specific definition of contact distance. However, setting an appropriate threshold for refining predictions involves a tradeoff between a higher PPV and a decreasing number of predictions remaining (i.e., less systems covered), as shown in the bottom panel of Figure 2A. For consistency, Figure 2B shows an aggregated scatter plot of the *Z*-score for all DCA interprotein predictions and the distances between interprotein residue pairs within a crystal structure. Interestingly, for the $N_{eff}/L \lesssim 1$ data set, we observed a few interprotein DCA predictions with a *Z*-score above 5.6 that are not a spatial contact in the crystal (actual distances >30 Å). Such a contradiction may be resolved by the existence of alternative dimeric conformations or higher order assemblies. This possibility is explored at the end of this section.

When selecting the top rank-order DCA predictions rather than applying a *Z*-score threshold, the top 10 dimeric DCA predictions had an average PPV below 25% for $N_{eff}/L \lesssim 1$ and below 40% for the $N_{eff}/L \gtrsim 1$ data sets (Figure 2C blue and red solid lines, respectively). Generally, selecting the top DCA predictions for monomeric contacts is an accurate predictor of intraprotein contacts; for example, selecting the top 10 predictions for monomeric contacts in all 61 dimers leads to an average PPV above 90% (Figure 2C inset—blue dashed lines). Filtering the top DCA predictions using a *Z*-score threshold for the dimeric contacts provides an improvement over using the top 10 dimer predictions for systems with $N_{eff}/L \gtrsim 1$.

https://doi.org/10.1021/acs.jpcb.1c07145
J. Phys. Chem. B 2021, 125, 11408–11417

Figure 2D illustrates the application of the Z-score thresholding to the DCA dimer predictions for each system belonging to the $N_{eff}/L \gtrsim 1$ and $N_{eff}/L \lesssim 1$ groups, respectively. The threshold of $Z = 5.6$ yields the most accurate dimer prediction, while still making predictions for 8 out of the 28 dimer systems with $N_{eff}/L \gtrsim 1$ (Figure 2D). As discussed previously, these results demonstrate that good predictions are not available to every system, that is, having sufficient interprotein contacts above the desired threshold (e.g., $Z > 5.6$ in Figure 2D). Two notable examples of systems with many statistically significant predicted dimer contacts were SigK-RskA (PDB ID: 4NQW) and hisH-hisF (PDB ID: 1KA9); both systems had abundant sequence data ($N_{eff}/L \gg 1$) that was accurately able to characterize the coevolutionary signals that reflect spatial contacts in both the monomeric folds and dimeric interfaces. Shown in Figure 3, using a Z-score threshold on dimer predictions with sufficient data ($N_{eff}/L \gtrsim 1$) removes false positives (panels A,B,E−G) and in the two cases (panels C,D) for which the top 10 predictions are all true positives, applying the Z-score threshold performs equally well.

Figure 2D shows very few DCA dimer contact predictions above $Z = 5.6$ for systems with limited data ($N_{eff}/L \lesssim 1$). Several representative examples are shown in Figure 4. For systems with limited data, only one system (PRPF38A-MFAP1) has a single DCA prediction above $Z = 5.6$ that is also a true positive dimer contact. Interestingly, three systems with limited data exhibit dimer contacts above $Z = 5.6$ that are not observed in a known experimental structure (Figure 4B−D). The fact that these predictions contain statistically significant covariance information suggests that they may not be false positives; they could potentially be associated with relevant alternative conformations of certain dimer complexes. We investigate this possibility for the Pex4p-Pex22p[58] system (included in the data limited subset; i.e., $N_{eff}/L \lesssim 1$). This dimer contains three predicted interprotein contacts with a Z-score between 5.6 and 5.8 (black circles shown in Figure 5B left panel) that are not observed in the known crystal structure. It should also be noted that using the top 10 DCA dimer predictions yielded two true positive contacts with a statistical significance above the threshold of $Z = 3.5$ (see Figure 5B right panel). Therefore, these largest Z-score predictions provide an interesting example case for which an alternative complex structure may exist (see Figure 5), a possibility that is further explored below.

Statistically significant DCA predictions that are not found in a known crystallographic protein structure pose an interesting question. Are these residue pairs real contacts in an alternative biologically relevant conformation? We explore this possibility by examining the case of the Pex4p−Pex22p complex, one of the dimers where this situation occurs (see Figure 5). In the peroxisome of a eukaryotic cell, E1, E2, and E3 enzymes participate in a series of ubiquitin-associated events that result in the transfer of ubiquitin to a substrate targeted for either degradation or translocation. Ubiquitin-conjugating E2 enzymes are involved in ubiquitin coordination to substrates.[58] The peroxisome-associated E2 enzyme, Pex4p, binds to the peroxisomal membrane protein Pex22p (pink and light-purple subunits shown in Figure 5A). None of the contacts that overlap with the crystal structure interface shows a high Z-score DCA pair. This protein complex binding is necessary for Pex4p to coordinate ubiquitin transfer to the target substrate via another enzyme Pex5p. The active site of Pex4p is located at a cysteine residue (labeled in Figure 5A and represented as a purple residue) and is implicated in ubiquitin binding.[58] No other known binding interface between Pex22p and Pex4p has been reported. The three DCA-predicted interchain pairs are shown

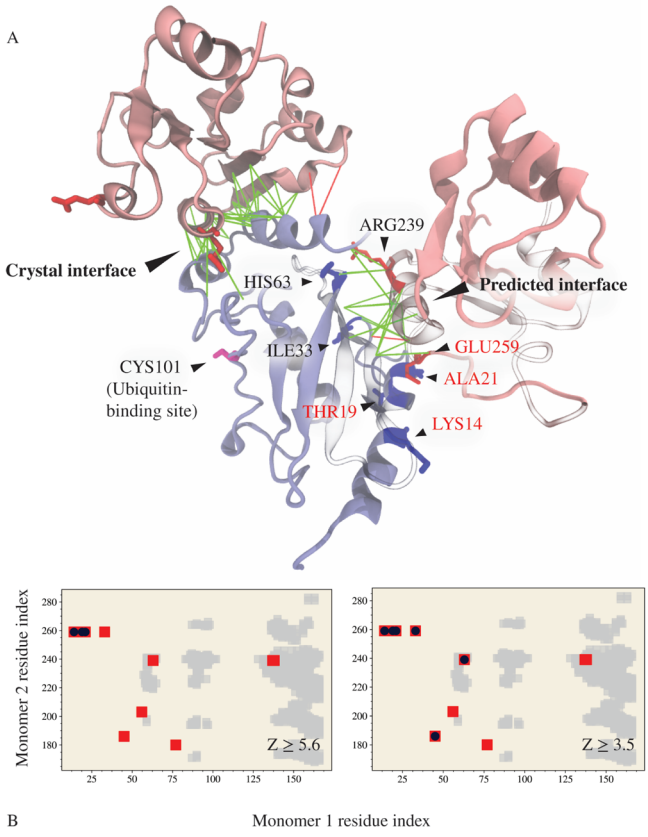


Figure 5. DCA predicts an alternative binding site at $Z > 5.6$ for the Pex4p-Pex22p dimer. Ubiquitin-conjugating enzymes play an important role in protein degradation via ubiquitin-coordination to substrates. (A) Peroxisome-associated E2 enzyme Pex4p, shown in a light-purple-colored new-cartoon representation, is bound to the peroxisomal membrane protein Pex22p (pink-colored new-cartoon). The crystal interface (top left, pink-colored subunit) is shown and compared to our predicted docked interface (top right, pink-colored subunit). We analyzed the energy landscape of both the crystal interface and our newly predicted (alternative) interface using the configurational frustration index, which is a proxy for the interaction energy between pairs of residues where a distribution of various decoy residue pair configurations is compared to the native configuration and expressed as a Z-score. Minimally and highly frustrated regions between pairs involved both interfaces are plotted as green- and red-colored lines, respectively. While the crystal interface is larger, our predicted interface is also minimally frustrated except for one highly frustrated interaction between glutamic acid and aspartic acid. The red text labels correspond to the Z-filtered DCA pairs. The peroxisomal membrane protein is a noncanonical E2 binding partner and thus affects the specific function of E2. An alternate binding site is revealed by these three contacts. We added two more contacts to stabilize the alternative dimer structure ($Z = 5.0$ and $Z = 4.3$, respectively; refer to Supporting Information csv file). (PDB ID: 2Y9M, resolution 2.6 Å.) Note that the PDB residues were reindexed to start from 1 (as described in Figure 1 caption) and thus are shifted away from the original PDB indexing. (B) Contact map showing the top ten dimeric predictions (red-colored squares), crystal interface contacts (gray squares), and Z-score filtered dimeric predictions (black circles). All three predicted interchain contacts (left panel) had Z-scores ≥5.6 and appear to be false positives because they do not exist in a known crystal structure. When a less stringent filter at $Z = 3.5$ (right panel) is applied, we obtain our first true dimer contact at a distance of 8 Å, which is present in the known crystal structure.

https://doi.org/10.1021/acs.jpcb.1c07145
J. Phys. Chem. B 2021, 125, 11408−11417

as filled black circles in Figure 5B left panel (labeled by the red-colored three-letter residue name). Figure 5A shows that their Z-scores are >5.6 but are separated by more than 30 Å in the crystal structure (PDB: 2Y9M). If we consider all interprotein residue pairs with a Z-scores above 5.6 as real interface contacts, these three new contacts reveal a potentially new binding interface between Pex22p and Pex4p. The first true dimer contact (8 Å distance in the crystal structure) is identified in our DCA predictions by relaxing the Z-score threshold to a cutoff of 3.5 (Figure 5B right panel). To explore the idea of a potentially new interface between Pex22p and Pex4p, we performed a molecular docking simulation where the top five contacts—the top three described above and two additional ones with $Z = 5.0$ and $Z = 4.3$, respectively (see Supporting Information csv file), are included to further stabilize this proposed new interface. To verify if this new interface is energetically stable, we utilized frustration analysis,[53,54] which quantitatively determines if the interface interactions are physically favorable. This approach is an energy landscape-based method for quantifying energetic interactions between residue pairs to determine their stability. We find that the predicted interface (see Figure 5A) is minimally frustrated and thus is a plausible alternate dimer structure. The predicted interface may also exist concurrently with the complex present in the crystal dimer, as part of a larger complex structure that may delay ubiquitination in a nonlinear manner. Because Pex22p has a membrane-bound domain (not yet crystallized), a rearrangement of either the membrane or of the domain may need to occur to allow for the formation of this predicted complex.

In summary, we are confident that DCA can yield a strong crystal dimer signal when the Z-scores are sufficiently large. Few true DCA dimer predictions are found for systems with limited data ($N_{eff}/L \lesssim 1$) (Figure 2D), illustrating the lack of statistical significance of the majority of predictions for these systems. Still, although limited, statistically significant dimer predictions can be found for these systems (Figure 4). Moreover, for the case of the Pex4p−Pex22p system ($N_{eff}/L = 0.20$), these predictions suggest an alternative dimer conformation. To probe the plausibility of this conformation, we integrated the high Z-score-filtered DCA pairs into a structure-based model to simulate the alternative docked structure and utilized physical methods to check its stability. It is of interest to note that dimer interfaces may not be unique to a single structure; homologues of bound substrates may be present in the DCA predictions.

CONCLUSIONS

Gaining a mechanistic understanding of proteins and their functional interactions requires knowledge of their three-dimensional structures. Much of this information is obtained from experiments (e.g., X-ray crystallography), but it is limited to a subset of known proteins. Sequence coevolutionary information helps fill this gap by providing additional information that can be used in protein structural predictions.[5−10,14,21,23,25,33,35,45,55]

DCA quantifies this coevolutionary information to predict structural contacts in a folded protein or protein complex. While DCA has found great success in predicting monomeric contacts,[6−10,14] generating predictions for the dimer complex (and higher order complexes) presents a significantly greater challenge because interprotein amino acid correlations are much weaker than intraprotein correlations. This is expected because evolutionary information is much stronger at the individual protein level relative to the protein−protein interface. Hence, selecting the top dimer predictions from a simple rank-order may not be sufficient to guarantee a good prediction for these interface contacts because the top predictions may not be statistically significant. To this end, we have introduced a simple Z-score analysis to assess the statistical significance of any DCA prediction. We have shown that setting a Z-score threshold on these predictions involves a tradeoff between predictive accuracy and the number of predictions that are made. In the extreme case, a stringent threshold of $Z = 5.6$ accurately predicts dimer contacts, although many systems that we examined did not have enough information to meet this threshold. Interestingly, DCA predictions for which $Z > 5.6$ are excellent candidates for alternative structural contacts when these contacts are not observed in known structures. We have explored the Pex4p−Pex22p dimer system as an illustrative example, generating a plausible alternative interface that is supported by our predicted dimer contacts. Predictions where $Z > 5.6$ may not always exist; yet DCA predictions at a lower Z-score threshold can still be used, albeit with a lower predictive accuracy.

Predictions for systems with limited data ($N_{eff}/L \lesssim 1$) further pose an interesting challenge. While we have no control over the amount of sequence data that is available, we can still potentially make progress in structure prediction for these limited data systems. For example, we can incorporate additional constraints on our model or perhaps reduce the amino acid representation to a reduced, coarse-grained representation. The contact predictions of our model can further be combined with physical modeling (e.g., docking), which would corroborate interprotein contact predictions.

Recent advances have demonstrated how machine learning from PDB structures can be used to create remarkable predictors of the protein structure.[59] Coevolutionary information is nevertheless a nonstructural source of information that can provide additional structural information. The strength of coevolutionary methods is further supported by developments in high throughput sequencing, producing new sequence data at a much higher rate than experimental structural data can be obtained. This type of approach offers a complementary source of information that can augment the new, cutting-edge structure prediction methods.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c07145.

Pexp4−Pexp22 DCA Z-scores (XLSX)

Average positive predictive value for the two groups and DCA-Docking Protocol (PDF)

AUTHOR INFORMATION

Corresponding Author

José N. Onuchic − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States; Systems, Synthetic, and Physical Biology, Department of Physics & Astronomy, Department of Chemistry, and Department of Biosciences, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0002-9448-0388; Phone: (713) 348-4197; Email: jonuchic@rice.edu

Authors

Kareem M. Mehrabiani − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States;*

Systems, Synthetic, and Physical Biology, Rice University, Houston, Texas 77005, United States; orcid.org/0000-0001-6733-2577

Ryan R. Cheng − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0001-6378-295X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c07145

Author Contributions

K.M.M conducted the analysis and research. The manuscript was written through contributions of all authors.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

K.M.M. thanks George Britton, Xingcheng Lin, and Brian Sirovetz for helpful discussion. This research was supported by the Center for Theoretical Biological Physics sponsored by the National Science Foundation (grant PHY-2019745). Additional support was provided by the National Science Foundation (NSF) grants CHE-1614101. JNO is a Cancer Prevention and Research Institute of Texas (CPRIT) Scholar in Cancer Research.

■ REFERENCES

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(2) Lockless, S. W.; Ranganathan, R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **1999**, *286*, 295−299.

(3) Bartlett, G. J.; Taylor, W. R. Using Scores Derived from Statistical Coupling Analysis to Distinguish Correct and Incorrect Folds in De-Novo Protein Structure Prediction. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 950−959.

(4) Burger, L.; van Nimwegen, E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.* **2008**, *4*, 165.

(5) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 67−72.

(6) Balakrishnan, S.; Kamisetty, H.; Carbonell, J. G.; Lee, S.-I.; Langmead, C. J. Learning Generative Models for Protein Fold Families. *Proteins* **2011**, *79*, 1061−1078.

(7) Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One* **2011**, *6*, No. e28766.

(8) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, E1293−E1301.

(9) Hopf, T. A.; Colwell, R. S.; Sheridan, R.; Rost, B.; Sander, C.; Marks, D. S. Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **2012**, *149*, 1607−1621.

(10) Sulkowska, J. I.; Morcos, F.; Weigt, M.; Hwa, T.; Onuchic, J. N. Genomics-Aided Structure Prediction. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 10340−10345.

(11) Kamisetty, H.; Ovchinnikov, S.; Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 15674−15679.

(12) de Juan, D.; Pazos, F.; Valencia, A. Emerging Methods in Protein Co-Evolution. *Nat. Rev. Genet.* **2013**, *14*, 249−261.

(13) Skwark, M. J.; Abdel-Rehim, A.; Elofsson, A. PconsC: Combination of Direct Information Methods and Alignments Improves Contact Prediction. *Bioinformatics* **2013**, *29*, 1815−1816.

(14) Ekeberg, M.; Hartonen, T.; Aurell, E. Fast Pseudolikelihood Maximization for Direct-Coupling Analysis of Protein Structure from Many Homologous Amino-Acid Sequences. *J. Comput. Phys.* **2014**, *276*, 341−356.

(15) Seemayer, S.; Gruber, M.; Söding, J. CCMpred—Fast and Precise Prediction of Protein Residue−Residue Contacts from Correlated Mutations. *Bioinformatics* **2014**, *30*, 3128−3130.

(16) Hopf, T. A.; Ingraham, J. B.; Poelwijk, F. J.; Schärfe, C. P. I.; Springer, M.; Sander, C.; Marks, D. S. Mutation Effects Predicted from Sequence Co-Variation. *Nat. Biotechnol.* **2017**, *35*, 128−135.

(17) Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **2017**, *13*, No. e1005324.

(18) Jones, D. T.; Kandathil, S. M. High Precision in Protein Contact Prediction Using Fully Convolutional Neural Networks and Minimal Sequence Features. *Bioinformatics* **2018**, *34*, 3308−3315.

(19) Hopf, T. A.; Green, A. G.; Schubert, B.; Mersmann, S.; Schärfe, C. P. I.; Ingraham, J. B.; Toth-Petroczy, A.; Brock, K.; Riesselman, A. J.; Palmedo, P.; et al. The EVcouplings Python Framework for Coevolutionary Sequence Analysis. *Bioinformatics* **2019**, *35*, 1582−1584.

(20) Jarmolinska, A. I.; Zhou, Q.; Sulkowska, J. I.; Morcos, F. A PyMOL Plugin To Analyze Direct Evolutionary Couplings. *J. Chem. Inf. Model.* **2019**, *59*, 625−629.

(21) Ovchinnikov, S.; Kamisetty, H.; Baker, D. Robust and Accurate Prediction of Residue−Residue Interactions across Protein Interfaces Using Evolutionary Information. *eLife* **2014**, *3*, No. e02030.

(22) Cheng, R. R.; Morcos, F.; Levine, H.; Onuchic, J. N. Toward Rationally Redesigning Bacterial Two-Component Signaling Systems Using Coevolutionary Information. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, E563−E571.

(23) dos Santos, R. N.; Morcos, F.; Jana, B.; Andricopulo, A. D.; Onuchic, J. N. Dimeric Interactions and Complex Formation Using Direct Coevolutionary Couplings. *Sci. Rep.* **2015**, *5*, 13652.

(24) Feinauer, C.; Szurmant, H.; Weigt, M.; Pagnani, A. Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. *PLoS One* **2016**, *11*, No. e0149166.

(25) Krepel, D.; Cheng, R. R.; Di Pierro, M.; Onuchic, J. N. Deciphering the Structure of the Condensin Protein Complex. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 11911−11916.

(26) Szurmant, H.; Weigt, M. Inter-Residue, Inter-Protein and Inter-Family Coevolution: Bridging the Scales. *Curr. Opin. Struct. Biol.* **2018**, *50*, 26−32.

(27) Croce, G.; Gueudré, T.; Ruiz Cuevas, M. V.; Keidel, V.; Figliuzzi, M.; Szurmant, H.; Weigt, M. A Multi-Scale Coevolutionary Approach to Predict Interactions between Protein Domains. *PLoS Comput. Biol.* **2019**, *15*, No. e1006891.

(28) Schug, A.; Weigt, M.; Onuchic, J. N.; Hwa, T.; Szurmant, H. High-Resolution Protein Complexes from Integrating Genomic Information with Molecular Simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 22124−22129.

(29) Green, A. G.; Elhabashy, H.; Brock, K. P.; Maddamsetti, R.; Kohlbacher, O.; Marks, D. S. Large-Scale Discovery of Protein Interactions at Residue Resolution Using Co-Evolution Calculated from Genomic Sequences. *Nat. Commun.* **2021**, *12*, 1396.

(30) Cocco, S.; Feinauer, C.; Figliuzzi, M.; Monasson, R.; Weigt, M. Inverse Statistical Physics of Protein Sequences: A Key Issues Review. *Rep. Prog. Phys.* **2018**, *81*, 032601.

(31) Cheng, R. R.; Nordesjö, O.; Hayes, R. L.; Levine, H.; Flores, S. C.; Onuchic, J. N.; Morcos, F. Connecting the Sequence-Space of Bacterial Signaling Proteins to Phenotypes Using Coevolutionary Landscapes. *Mol. Biol. Evol.* **2016**, *33*, 3054−3064.

(32) Procaccini, A.; Lunt, B.; Szurmant, H.; Hwa, T.; Weigt, M. Dissecting the Specificity of Protein-Protein Interaction in Bacterial

Two-Component Signaling: Orphans and Crosstalks. *PLoS One* **2011**, *6*, No. e19729.
(33) Cheng, R. R.; Haglund, E.; Tiee, N. S.; Morcos, F.; Levine, H.; Adams, J. A.; Jennings, P. A.; Onuchic, J. N. Designing Bacterial Signaling Interactions with Coevolutionary Landscapes. *PLoS One* **2018**, *13*, No. e0201734.
(34) Morcos, F.; Jana, B.; Hwa, T.; Onuchic, J. N. Coevolutionary Signals across Protein Lineages Help Capture Multiple Protein Conformations. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 20533−20538.
(35) Cheng, R. R.; Raghunathan, M.; Noel, J. K.; Onuchic, J. N. Constructing Sequence-Dependent Protein Models Using Coevolutionary Information. *Protein Sci.* **2016**, *25*, 111−122.
(36) Uguzzoni, G.; John Lovis, S.; Oteri, F.; Schug, A.; Szurmant, H.; Weigt, M. Large-Scale Identification of Coevolution Signals across Homo-Oligomeric Protein Interfaces by Direct Coupling Analysis. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E2662−E2671.
(37) Ekeberg, M.; Lövkvist, C.; Lan, Y.; Weigt, M.; Aurell, E. Improved Contact Prediction in Proteins: Using Pseudolikelihoods to Infer Potts Models. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2013**, *87*, 012707.
(38) Shukla, J.; Gupta, R.; Thakur, K. G.; Gokhale, R.; Gopal, B. Structural basis for the redox sensitivity of theMycobacterium tuberculosisSigK-RskA σ-anti-σ complex. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2014**, *70*, 1026−1036.
(39) Bitbol, A.-F.; Dwyer, R. S.; Colwell, L. J.; Wingreen, N. S. Inferring Interaction Partners from Protein Sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 12180−12185.
(40) Gueudré, T.; Baldassi, C.; Zamparo, M.; Weigt, M.; Pagnani, A. Simultaneous Identification of Specifically Interacting Paralogs and Interprotein Contacts by Direct Coupling Analysis. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 12186−12191.
(41) Cong, Q.; Anishchenko, I.; Ovchinnikov, S.; Baker, D. Protein Interaction Networks Revealed by Proteome Coevolution. *Science* **2019**, *365*, 185−189.
(42) Dunn, S. D.; Wahl, L. M.; Gloor, G. B. Mutual Information without the Influence of Phylogeny or Entropy Dramatically Improves Residue Contact Prediction. *Bioinformatics* **2008**, *24*, 333−340.
(43) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B* **2012**, *116*, 8494−8503.
(44) Rodriguez Horta, E.; Weigt, M. On the Effect of Phylogenetic Correlations in Coevolution-Based Contact Prediction in Proteins. *PLoS Comput. Biol.* **2021**, *17*, No. e1008957.
(45) Bai, F.; Morcos, F.; Cheng, R. R.; Jiang, H.; Onuchic, J. N. Elucidating the druggable interface of protein−protein interactions using fragment docking and coevolutionary analysis. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, E8051−E8058.
(46) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46*, W296−W303.
(47) Sirovetz, B. J.; Schafer, N. P.; Wolynes, P. G. Protein structure prediction: making AWSEM AWSEM-ER by adding evolutionary restraints. *Proteins: Struct., Funct., Bioinf.* **2017**, *85*, 2127−2142.
(48) Chen, M.; Lin, X.; Lu, W.; Schafer, N. P.; Onuchic, J. N.; Wolynes, P. G. Template-Guided Protein Structure Prediction and Refinement Using Optimized Folding Landscape Force Fields. *J. Chem. Theory Comput.* **2018**, *14*, 6102−6116.
(49) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput. Biol.* **2016**, *12*, No. e1004794.
(50) Levi, M.; Bandarkar, P.; Yang, H.; Wang, A.; Mohanty, U.; Noel, J. K.; Whitford, P. C. Using SMOG 2 to Simulate Complex Biomolecular Assemblies. In *Biomolecular Simulations: Methods and Protocols*; Bonomi, M., Camilloni, C., Eds.; Methods in Molecular Biology; Springer: New York, NY, 2019, pp 129−151. DOI: 10.1007/978-1-4939-9608-7_6
(51) Lemkul, J. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]. *Living J. Comput. Mol. Sci.* **2018**, *1*, 5068.
(52) Zhang, Y.; Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 702−710.
(53) Ferreiro, D. U.; Hegler, J. A.; Komives, E. A.; Wolynes, P. G. Localizing Frustration in Native Proteins and Protein Assemblies. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19819−19824.
(54) Parra, R. G.; Schafer, N. P.; Radusky, L. G.; Tsai, M.-Y.; Guzovsky, A. B.; Wolynes, P. G.; Ferreiro, D. U. Protein Frustratometer 2: A Tool to Localize Energetic Frustration in Protein Molecules, Now with Electrostatics. *Nucleic Acids Res.* **2016**, *44*, W356−W360.
(55) Thadani, N. N.; Zhou, Q.; Reyes Gamas, K.; Butler, S.; Bueno, C.; Schafer, N. P.; Morcos, F.; Wolynes, P. G.; Suh, J. Frustration and Direct-Coupling Analyses to Predict Formation and Function of Adeno-Associated Virus. *Biophys. J.* **2021**, *120*, 489−503.
(56) Malinverni, D.; Marsili, S.; Barducci, A.; De Los Rios, P. Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. *PLoS Comput. Biol.* **2015**, *11*, No. e1004262.
(57) Altegoer, F.; Rensing, S. A.; Bange, G. Structural Basis for the CsrA-Dependent Modulation of Translation Initiation by an Ancient Regulatory Protein. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 10168−10173.
(58) Williams, C.; van den Berg, M.; Panjikar, S.; Stanley, W. A.; Distel, B.; Wilmanns, M. Insights into ubiquitin-conjugating enzyme/co-activator interactions from the structure of the Pex4p:Pex22p complex. *EMBO J.* **2012**, *31*, 391−402.
(59) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.