Origins of coevolution between residues distant in protein 3D structures

Ivan Anishchenko[a,b], Sergey Ovchinnikov[a,b,c], Hetunandan Kamisetty[d], and David Baker[a,b,e,1]

[a]Department of Biochemistry, University of Washington, Seattle, WA 98105; [b]Institute for Protein Design, University of Washington, Seattle, WA 98105; [c]Molecular and Cellular Biology Program, University of Washington, Seattle, WA 98195; [d]Facebook Inc., Seattle, WA 98109; and [e]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105

Edited by Barry Honig, Howard Hughes Medical Institute, Columbia University, New York, NY, and approved July 6, 2017 (received for review February 15, 2017)

Residue pairs that directly coevolve in protein families are generally close in protein 3D structures. Here we study the exceptions to this general trend—directly coevolving residue pairs that are distant in protein structures—to determine the origins of evolutionary pressure on spatially distant residues and to understand the sources of error in contact-based structure prediction. Over a set of 4,000 protein families, we find that 25% of directly coevolving residue pairs are separated by more than 5 Å in protein structures and 3% by more than 15 Å. The majority (91%) of directly coevolving residue pairs in the 5–15 Å range are found to be in contact in at least one homologous structure—these exceptions arise from structural variation in the family in the region containing the residues. Thirty-five percent of the exceptions greater than 15 Å are at homo-oligomeric interfaces, 19% arise from family structural variation, and 27% are in repeat proteins likely reflecting alignment errors. Of the remaining long-range exceptions (<1% of the total number of coupled pairs), many can be attributed to close interactions in an oligomeric state. Overall, the results suggest that directly coevolving residue pairs not in repeat proteins are spatially proximal in at least one biologically relevant protein conformation within the family; we find little evidence for direct coupling between residues at spatially separated allosteric and functional sites or for increased direct coupling between residue pairs on putative allosteric pathways connecting them.

protein coevolution | structural variation | homo-oligomeric contacts

Natural proteins tread a delicate balance between maintaining structural stability and carrying out their biological function. The amino acid sequence covariation in evolutionarily related proteins arises from protein structural, functional, and stability constraints, and methods have been developed for predicting residue–residue contacts in protein 3D structures from multiple sequence alignment (MSA) data (1–5). These methods disentangle direct couplings between residue pairs from indirect couplings that arise from chaining effects: If A is correlated with B, and B with C, then correlations between A and C will be observed even if A and C do not directly interact with one another. The direct couplings are isolated by expressing the energy of a sequence as the sum of one and two body interactions between residues and then finding values of the one and two body interaction parameters that best account for the observed sequence data; residue pairs with significant two body interaction parameters are considered to be directly coupled. In large protein families many pairs of residues—both close and distant in space—are found to covary because of chaining effects; directly coupled residues, however, are usually close in the 3D structure. There has been considerable recent success in using direct couplings to infer residue–residue contacts for protein structure prediction (6–9), protein–protein complex prediction (10–12), identification of specific interaction partners among two sets of paralogous proteins (13), resolution of ambiguities in protein NMR spectra (14), identification of structurally ordered regions within intrinsically disordered proteins (15), modeling of conformational changes (16, 17), and modeling homo-oligomeric complexes (18, 19).

However, not all directly coupled residues are close in the 3D structure. It has been proposed that coevolution between distant sites can arise from residue–residue interactions through allosteric interaction networks (20), negative design (21), codon effects (22), and spurious phylogenetic correlations (23, 24). Although previous work has examined individual cases in which pairs of directly coevolving residues are distant in the 3D structure, we are not aware of any comprehensive analysis over the protein structure database of the contributions to such potential contact mispredictions.

Coevolution-based structure prediction methods will become even more powerful as more genomes are sequenced, and it is important to understand possible sources of error as well as to gain general understanding of the sources of direct evolutionary coupling between residue pairs. In this paper, we systematically study the extent and source of direct coupling between residues distant in protein 3D structures in the Protein Data Bank (PDB). We find that almost all such coupling can be traced to direct physical interactions in at least one physiologically relevant conformation in the protein family.

Results and Discussion

Distances Between Coevolving Residue Pairs. To investigate the origins of direct evolutionary coupling of residue pairs in protein 3D structures, we focus on the 0.5 × (protein length) most strongly coupled pairs identified by GREMLIN separated by more than six residues in the primary sequence. As found previously, the fraction of these pairs that are close in the 3D structure (shortest distance between heavy atoms less than 5 Å) increases with increasing number and diversity of the sequences in the family (Fig. 1*A*; we use M_{eff} defined in *Methods* to summarize both effects). As the

Significance

Coevolution-derived contact predictions are enabling accurate protein structure modeling. However, coevolving residues are not always in contact, and this is a potential source of error in such modeling efforts. To investigate the sources of such errors and, more generally, the origins of coevolution in protein structures, we provide a global overview of the contributions to the "exceptions" to the general rule that coevolving residues are close in protein three-dimensional structures.

Author contributions: S.O., H.K., and D.B. designed research; I.A. and S.O. performed research; I.A., S.O., and D.B. analyzed data; and I.A. and D.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: MSAs and GREMLIN contact maps are available at gremlin.bakerlab.org/exceptions.

[1]To whom correspondence should be addressed. Email: dabaker@u.washington.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1702664114/-/DCSupplemental.

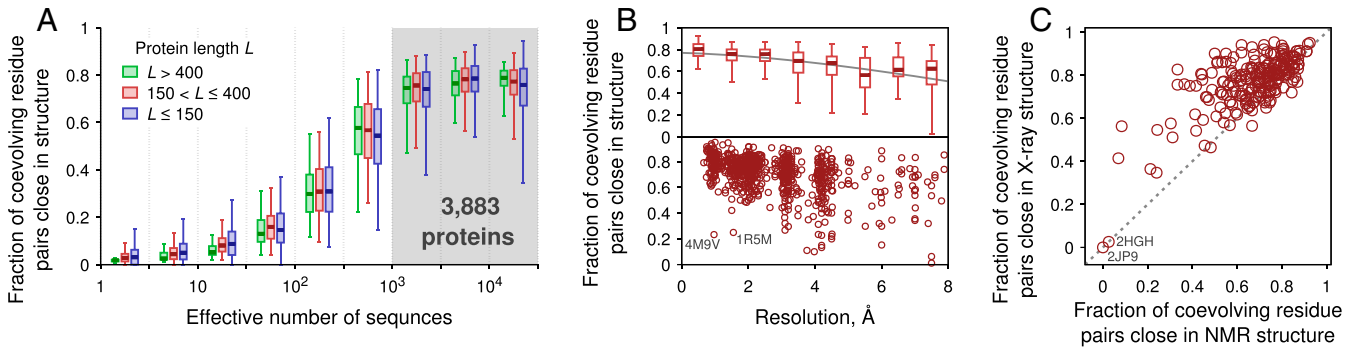

Fig. 1. The frequency with which coevolving directly coupled residues are in contact depends on structure quality and MSA size. (*A*) The contact frequency of directly coupled residue pairs depends on number of sequences in family. Fraction of top 0.5 × (protein length) coevolving directly coupled residue–residue pairs identified by GREMLIN that make contacts in small (protein length L ≤ 150; blue box-and-whiskers), medium (150 < L ≤ 400; red), and large (L > 400; green) protein 3D structures. The contact prediction regime analyzed in this paper (with $M_{eff} > 10^3$) is highlighted in gray. Two residues in the protein 3D structure were considered to be in contact if any pair of heavy atoms are within 5 Å distance. (*B*) The contact frequency of directly coupled residue pairs increases with increasing structure accuracy. The correlation between GREMLIN prediction accuracy and X-ray crystallographic resolution is shown in scatter (*Lower*) and box-and-whiskers (*Upper*) plots (boxes and whiskers comprise 25%, 75% and 2.5%, 97.5% percentiles, respectively; the median is shown by a solid horizontal line). (*C*) Comparison of GREMLIN contact prediction accuracy in X-ray and solution NMR structures for 222 proteins with structures determined using both methods; contact prediction accuracy is consistently higher for the X-ray structures. The outliers marked on *B* and *C* by PDB codes are all repeat proteins.

number of sequences reaches 10^3, the contact prediction accuracy saturates and remains almost constant regardless of the size of the protein (with smaller numbers of sequences, prediction accuracy is smaller for larger proteins) (Fig. 1*A*). Hence, in the following analysis, we focus on a subset of 3,883 proteins with at least 10^3 effective sequences in the MSAs.

This paper is focused on residue pairs that are strongly coupled but are not close in the 3D structure—we refer to such pairs as "exceptions" throughout the text. A first possible source of exceptions is inaccuracy in the 3D structures—residue pairs that are not close in the structure may indeed be close. To investigate this possible contribution, we determined the dependence of the frequency of exceptions on structure quality. For families with $M_{eff} > 10^3$, the frequency of exceptions increases as the X-ray structure resolution decreases (Fig. 1*B*), suggesting that structure inaccuracy does contribute to the exceptions. Exceptions are also higher in solution NMR structures than in high-resolution X-ray structures (Fig. 1*C*). Although NMR structures avoid possible artifacts due to crystal contacts present in X-ray structures, overall this trend suggests these artifacts do not outweigh the overall increased accuracy of high-resolution X-ray structures for which the experimental data are generally considerably more extensive and less ambiguous.

Comparison of the fit of the coevolutionary direct coupling data to structures with moderate resolution ~3 Å suggests that the data can improve model refinement. The multiple different structures solved for the ribosome and the many proteins they contain provide many examples (Fig. S1). For instance, direct coupling-based GREMLIN contact predictions for the 50S ribosomal protein L24 from *Thermus thermophilus* (Fig. S1 *A*–*C*) are clearly more consistent with PDB entry 4V8H than 4V9H, although the latter has higher resolution (2.86 vs. 3.1 Å, respectively). Coevolutionary direct coupling data are also more consistent with the 4V8H structure for four other ribosomal proteins (Fig. S1).

Several structures stand out in Fig. 1 as having a particularly poor agreement with the predicted contacts: the two cases at the bottom-left corner of Fig. 1*C* (PDB entries 2HGH and 2JP9) with poorly predicted contacts in both the NMR and X-ray structures, and the two high-resolution structures 4M9V and 1R5M in Fig. 1*B*. All four of these proteins, as well as many of the others with large numbers of exceptions, are repeat proteins. As has been noted previously (25), the translational symmetry of repeat protein sequences gives rise to pair correlations between residues not in contact, and hence, GREMLIN incorrectly predicts contacts between segments (Fig. S2). Seventeen percent of the strongly coevolving residue pairs identified by GREMLIN with $d_{min} > 15$ Å are in repeat proteins; the average contact prediction accuracy for the 99 repeat proteins in the 3883 proteins is 0.62 compared with 0.75 overall. Coevolution-based contact prediction for repeat proteins clearly should be carried out using techniques that explicitly address the coupling arising from the internal sequence repeats (25).

The analysis in the remainder of the paper focuses on the exceptions in the highest resolution nonrepeat protein X-ray structures where the contribution of structure error is likely quite low. Although only a small fraction of the most strongly coevolving directly coupled residues (the top L/2 residue pairs by GREMLIN score) are distant in the structure, the GREMLIN score distributions for these residues (Fig. S3*A*, green and blue lines) are more similar to those of actually contacting residues in these proteins (Fig. S3*A*, red line) than to those of randomly selected residue pairs (Fig. S3*A*, gray line).

Amino Acid and Distance Distributions of Strongly Coevolving Residue Pairs. It is instructive to examine the closest heavy atom distance distribution for strongly coevolving directly coupled residue pairs, which contains three local maxima at 2.8 Å, 3.7 Å, and 6.8 Å with a long tail at larger distances (Fig. 2*A*). We consider each of the three maxima and the long tail in turn in the following sections. For each, we examine the distribution of amino acid pairs, represented by a symmetric 20 × 20 matrix where an entry (i,j) represents the frequency of residue (i,j) pairs contributing to the population. The log ratios of the observed to expected frequencies of specific amino acid pairs (*Methods*) in each distance bin $d_{min} \leq d < d_{max}$ are depicted in the heat maps in Fig. 2 *B*–*D*.

The first peak in the distance distribution in Fig. 2*A* comprising distances $0 < d_{min} \leq 3$ Å is enriched in residues with charged and polar side chains (Fig. 2*B*). These side-chains can form hydrogen bonds, and the mode of the peak at 2.7–2.8 Å is consistent with donor–acceptor distance of hydrogen bonds in proteins. Pairs with the same charge (involving aspartate and glutamate residues, for example) are less likely to make contacts due to electrostatic repulsion (light-red square near the diagonal at D and E positions on Fig. 2*B*).

The second peak at ~3.7 Å is enriched in hydrophobic residue pairs (Fig. 2*C*). Approximately 3.7 Å is close to twice the van der

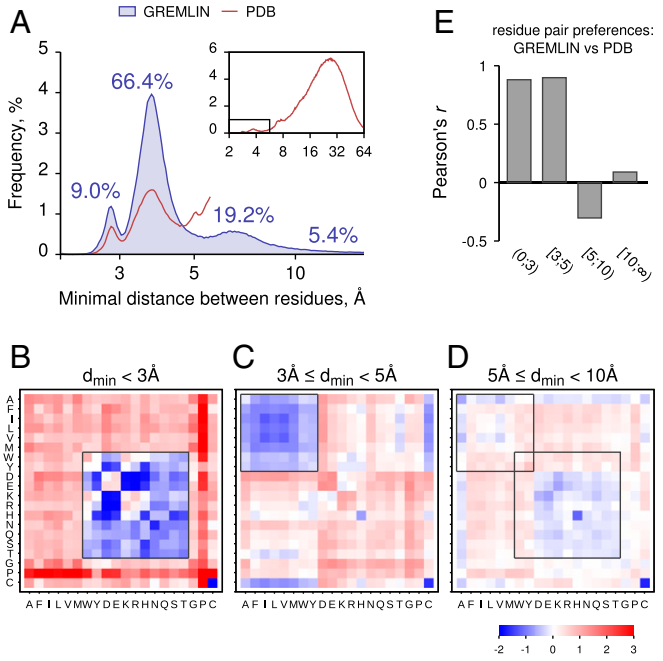

Fig. 2. Amino acid and distance distributions of strongly coevolving directly coupled residue pairs. (*A*) Distribution of distances between directly coupled residue pairs in 3,883 high-resolution X-ray protein structures (see legend to Fig. 1). Numbers indicate the fraction of the population in the ranges (0;3), (3;5), (5;10), and (10;∞). The distribution of distances for all residue pairs in the same set of protein structures is shown in *Inset* and as a red line for short distances in the main panel (scale is arbitrary). (*B–D*) Amino acid pair composition of directly coupled residue pairs at distances $0 < d_{min} \leq 3$ Å (*B*), $3 < d_{min} \leq 5$ Å (*C*), and $5 < d_{min} \leq 10$ Å (*D*) (blue, enriched; red, depleted). (*E*) Pearson correlation coefficients between the amino acid pair distributions in *B–D* and the corresponding distributions derived for all contacts in the structure in the four distance ranges.

Waals radius of the carbon atom; this peak is that expected for interactions between hydrophobic residue pairs. Thus, the first two peaks of the distribution of distances cover direct physical interactions between residues—this is why we chose a 5 Å distance cutoff in this paper for defining exceptions. The positions of the first and second peaks (Fig. 2*A*) and the amino acid compositions of the two peaks (first two bars in Fig. 2*E*) are quite similar in the coevolving residue pairs identified by GREMLIN and in PDB as a whole. It may be possible to obtain more accurate contact distance predictions from less sequence data by using the sequence distributions in Fig. 2 *B–D* to constrain GREMLIN two-body parameter estimates.

The third peak in between 5 and 10 Å has characteristics of both the first and second peaks but somewhat less pronounced (Fig. 2*D*). Unlike the first and second peaks, the residue composition of this peak is substantially different from that in PDB as a whole in this distance range (third bar on Fig. 2*E*). We will return to the origins of this peak and the discrepancy below.

Analysis of Coevolving Residue Pairs at Longer Distances. Twenty-five percent of coevolving directly coupled residue pairs are separated by more than 5 Å, and 3% by more than 15 Å in the structure. Although only a small fraction of all direct couplings, the absolute number in the latter category is quite significant (12,755 pairs in our high-resolution PDB subset). In the following sections, we examine the origins of these exceptions.

Homo-oligomer interfaces. Most proteins in nature perform their functions by interacting with other proteins. Around three-fourths of multimeric biological units in PDB represent homo-oligomeric assemblies (26), and coevolution occurs at homo-oligomeric interfaces (27), leading to contact predictions by GREMLIN across interfaces. To determine the contribution of contacts across homo-oligomeric interfaces to the total number of exceptions, we analyzed GREMLIN direct couplings in all biological unit files originating from the same PDB entry. In each biounit, we first searched for all chains that share at least 70% sequence identity with the query chain and then checked whether any of the highly coevolving directly coupled residue pairs distant within the monomer were separated by short distances across homo-oligomeric interfaces. If several biounit files were available for one PDB entry, the one with the highest number of GREMLIN contacts below 5 Å was used as the reference structure. The fraction of the coevolving residue pairs distant in the monomers (exceptions), which interact across homo-oligomeric interfaces, is summarized on Fig. 3*A*. Thirty-four percent of directly coupled residue pairs that are far apart ($d_{min} > 15$ Å) in the 3D structure within the monomer make a direct physical interchain contact in homo-oligomers (an arrow from category III to category I in Fig. 3*A*). For example, the contact map for the P5CR oxidoreductase (Δ^1-pyrroline-5-carboxylate reductase) predicted by GREMLIN contains a large group of contacts that are not present in the X-ray structure (PDB entry 1YQG) of a single chain (group of bright red dots on Fig. S4*A*). The catalytic unit of this protein is a homodimer, and most of the directly coupled pairs not close in the monomer are close in the homodimer. Interactions across homo-oligomeric interfaces are the primary source of strong coevolution between residue pairs separated by more than 15 Å (Fig. S5).

Structural variation. The MSAs from which GREMLIN and other methods identify strongly coevolving residue pairs contain information on all members of the family, and evolutionary coupled residue pairs that are not close in a structure of one family member may be close in another. Thus, structural variation within a protein family could account for some fraction of the exceptions. Indeed, 81% of the exceptions at distances of 5–15 Å (II → I arrow on Fig. 3*B*) are less than 5 Å apart in at least one homolog structure, and 19% of the exceptions at distances >15 Å (Fig. 3*B*; we chose to separate exceptions above and below 15 Å into two classes as homo-oligomer interfaces dominate the former and structural variation the latter). The shift of strong residue pairs toward shorter distances in homologs is much greater than for residue pairs selected at random (Fig. S6*B*): Although some nonvanishing transitions occur between adjacent distance bins, reductions to $d_{min} < 5$ Å (II → I and III → I arrows in Fig. S6*B*) are much less pronounced than for strongly coevolving sites.

Examples of exceptions arising from structural variation are shown in Fig. 3*E*. The majority of strongly coevolving residue pairs that are not in contact in the X-ray structure of the α-mannosidase (PDB entry 4AYO) do interact directly in one of the five homologous structures (Fig. 3*E*). Unlike interchain contacts in homo-oligomers (Fig. S4*A*), this type of exception is spread all around the contact map (Fig. S4*B*) as small structural variations within the family can occur throughout a structure. Errors arising in contact-based structure prediction from this class of exceptions are likely to correlate with structural variation in the family.

Structural variation with a protein family is also likely responsible for the third mode at ~7 Å in the distance distribution on Fig. 2*A*: Strongly coupled residue pairs that do not make a direct physical contact in the input PDB structure but are still relatively close in space are likely to interact directly in one of the homologs (transitions II → I in Fig. 3*B* have 81% probability). Structural changes that bring such residues in contact likely are accompanied by change of identities of the corresponding amino acids to make a contact favorable, explaining why the residue pair composition matrix for this mode (Fig. 2*D*) resembles the superposition of the matrices for the two shorter modes. This resemblance is even stronger when one considers the full sequence

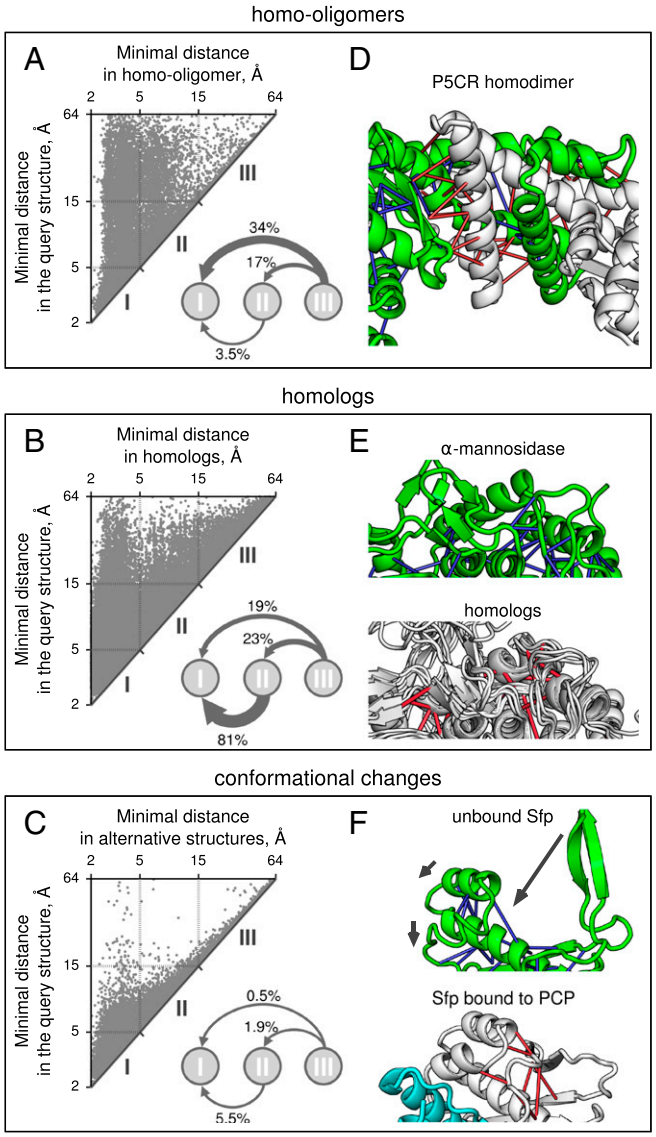

Fig. 3. Origins of exceptions. The top 0.5 × (protein length) directly coupled residue pairs were analyzed for every chain from the test set of 3,883 proteins with M_{eff} > 1,000. In *Top*, for each residue pair, the distance within the monomer (*y* axis) is plotted versus the shortest distance observed for the residue pair in *A* in homo-oligomeric assemblies in the PDB biological unit, (*B*) homologous PDB structures detected by the HHsearch program (HMM constructed from the initial MSA was searched against the database of HMMs for the entire PDB to select matches with E-value < 1E-20), and (*C*) close homologs with sequence identity >95% to capture possible conformational changes. The scatter plot data are summarized in the transition diagrams below; I are pairs in physical contact, II are pairs between 5 and 15 Å, and III are pairs separated by more than 15 Å; arrows indicate the frequency with which contacts at long distance shift to shorter distances, with thicker arrows corresponding to more probable transitions. Corresponding background rates are in Fig. S7 *A*–*C*. Crystal structures exemplifying each source of exceptions are shown in *D*–*F*: (*D*) the homodimeric complex of the P5CR oxidoreductase (PDB entry 1YQG), (*E*) the α-mannosidase (hydrolase) (4AYO) along with five homologous structures (1DL2, 1NXC, 1HCU, 1X9D, 2RI9) overlayed with one another, and (*F*) the Sfp transferase with (white; 1QR0) and without (green; 4MRT, chain A) the PCP (cyan; 4MRT, chain C). Blue sticks in the structures indicate residue pairs that are in contact (d_{min} < 5 Å) in the query PDB file, and red sticks represent additional residue pairs that are adjacent at the homooligomeric interface (*D*), in homologous structures (*E*), and in the bound conformation of the Sfp protein (*F*). Full structures and corresponding contact maps are in Fig. S4.

variation in each MSA (Fig. S7) rather than just the sequences of the query PDB files.

Conformational change is a special case of the structural variation explanation for exceptions. If two or more structures have been solved for the same or very closely related protein sequences, and the different structures represent different functional conformations of the protein, then a directly coupled pair that is an exception when referred to one structure may be close in space in another. To identify such pairs, for every query protein, we identified protein chains with sequence identity >95%. Only a small fraction of strongly coevolving residue pairs are associated with conformational changes that lead to substantial spatial reorganization of the protein structure (long-range transitions III → I in Fig. 3*C* occur with only 0.5% probability). An example is shown in Fig. 3*F*. The Sfp protein is responsible for activation of the peptidyl carrier protein (PCP) domains of surfactin synthetase by transferring the phosphopantetheine group from CoA to the PCP domain. Upon binding to PCP, two α-helices relocate, and a β-hairpin at the C terminus changes orientation (shown by arrows on Fig. 3*F*). Residue pairs that come in contact as the result of this structural rearrangement exhibit strong direct couplings (red dots on the contact map, Fig. S4*C*).

About half of the directly coupled residue pairs that are distant in the query protein can be attributed to either interactions across interfaces or structural variation within families (Fig. 4 and Fig. S5). Unlike the repeat protein case, these exceptions are not spurious as they contain information on homologous protein structures or on homo-oligomeric assemblies. The classification of exceptions as due to structural variation within a family is clearly sensitive to the number of structures that have been solved for family members. In families with hundreds of structures solved, structural variation can explain up to 95% of short-distance and 45% of long-distance exceptions that are unexplained by other sources (Fig. S8). By extrapolation, if all families had this many members with known structures, around three-fourths of long-distance and almost all short-distance contacts could be explained in this way (Fig. 4*A*). The breakdown of the contributions from the three major sources of exceptions at large distances

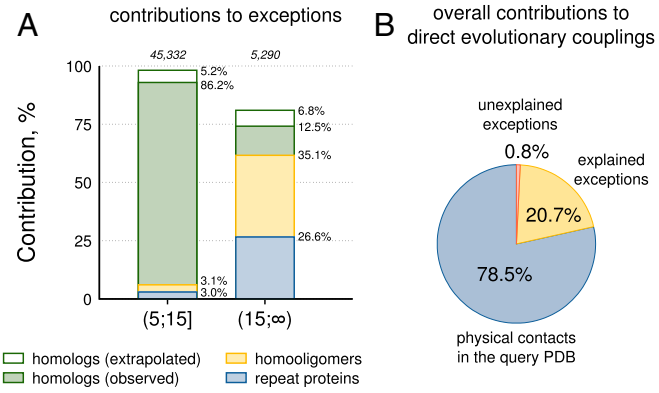

Fig. 4. With the exception of repeat proteins, directly coupled residue pairs in proteins are in direct physical contact. (*A*) Contributions from the three major sources of exceptions at intermediate (5 Å < d_{min} ≤ 15 Å; left bar) and long (d_{min} > 15 Å; right bar) distances are shown: repeat proteins (blue), homooligomeric interfaces (yellow), and homologs (both close and distant; green). Extrapolated contribution from homologs (white bars) is calculated based on the data from Fig. S4. The total number of contacts in each category is shown to the right of the corresponding bars. (*B*) Overall contributions to direct evolutionary couplings: Colors indicate residue pairs that are within 5 Å in the query PDB structure (blue), explained (yellow) and unexplained (red) exceptions. A subset of 235,644 directly coevolving pairs with GREMLIN scores > 0.5 were analyzed.

d_{min} > 15 Å is quite robust to the exact choice of the threshold value (Fig. S5).

Exceptions arising from homo-oligomeric interfaces, structural variation, and conformational change have GREMLIN coupling strength distributions essentially identical to coevolving residue pairs in contact in the monomeric structure (Fig. S3*B*), suggesting that the former and the latter are under equivalent evolutionary selection pressure. In contrast, the GREMLIN score distribution of the remaining unexplained exceptions is closer to the background distribution for residue pairs picked at random (Fig. S3*B*). Thus, a significant portion of the unexplained exceptions have lower GREMLIN scores and are likely mispredictions. Imposing a threshold of 0.5 on the GREMLIN score removes a considerable fraction of the unexplained exceptions; all but 0.8% of top coevolving pairs with GREMLIN scores above this value are either directly in contact or are in one of the three classes of explained exceptions (Fig. 4*B*).

There are still ∼25% (or 1,000 in absolute number) of directly coupled long-distance (d_{min} > 15 Å) exceptions that remain unexplained. Inspection of several such cases suggests that even in this class, many of the residue pairs are in close contact in a biologically relevant conformation. For example, the CysB protein from Gram-negative bacteria (PDB entry 1AL3), while a monomer in the crystal and in the biological unit, is a tetramer (dimer of dimers) in solution (28), and nearly all of the unexplained coevolving residue pairs are in contact across a crystal packing interface. Similarly, the bacterial LuxO protein, in the AAA+ ATPase superfamily of ring-shaped assemblies, is a monomer in the structure and biological unit structure (PDB entry 5EP2) (29) but clearly hexameric, and the unexplained exceptions are across the hexamer interface. In both of these two cases, the biological unit definitions provided in the PDB file are likely incorrect. Also in this category are unexplained exceptions consistent with the homodimeric structure in the asymmetric unit of PDB entry 3IHUA for which the biounit is listed as monomeric; the structure is unpublished. A different source of error is exemplified by the heterodimer of two homologous but not identical proteins ChsE4 and ChsE5 from *Mycobacterium tuberculosis* (PDB entry 4X28) (30). GREMLIN identifies strongly coupled residue pairs across the heterodimer interface that are (incorrectly) also predicted for the individual proteins, as both are included in the MSA for the family. Similarly, the heterodimeric interface contacts between ketosynthase and chain length factor from PDB entry 1TQY (31) are incorrectly attributed to each of the two subunits. In all of these cases, the unexplained exceptions correspond to biologically relevant contacts but were missed in our large-scale analysis above due to incomplete (or improperly annotated) PDB data or because of divergence of homo-oligomers into heteroligomers.

It has been suggested that residue–residue coupling can arise at long distances due to allosteric networks between residues. For example, residues distant from the active site in dihydrofolate reductase (DHFR) influence enzyme catalysis, and statistical coupling analysis (32) has suggested these residues are strongly correlated with a sector of evolutionary coupled sites within the protein (33). However, all of the directly coupled sites in DHFR are close in the 3D structure (Fig. 5*A* and Fig. S9*A*), suggesting that the distant residue pairs identified previously are no more coupled than any pair of residues connected by a chain of contacting residue pairs in the structure. There are also no direct couplings between residues interacting with an allosteric inhibitor of the enzyme cathepsin K (34) and the active site (Fig. 5*B* and Fig. S9*B*). The correlations observed in the previous studies evidently result from chaining together of the direct couplings between physically contacting residues that are the immediate subject of evolutionary selection. We also did not observe any stronger direct couplings between contacting residues on possible paths between allosteric and functional sites than between other contacting residue pairs in the structure (Fig. 5 *A* and *B*; the most strongly coupled pairs are connected by yellow tubes).

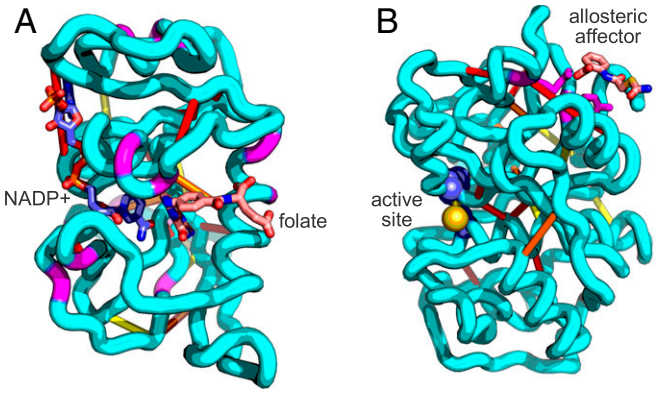

Fig. 5. Coevolutionary direct coupling in allosterically regulated proteins is between spatially adjacent residues. (*A*) Crystal structure (PDB entry 1RX2) of the DHFR with a cofactor NADP+ (nicotinamide adenine dinucleotide phosphate, oxidized form) and a substrate molecule (folate); 14 putative allosteric sites from ref. 33 are highlighted in magenta. (*B*) Crystal structure of the cathepsin K protein (PDB entry 1ATK) bound to an allosteric inhibitor through residues Tyr169 and Arg198 (in magenta). Catalytic dyad Cys25 and His162 are shown in spheres. The strongest GREMLIN contacts are shown as yellow (top 1–5), orange (top 6–10), and red (top 11–20) sticks in the structures. No residue pairs identified by GREMLIN are distant in the structure. Corresponding contact maps are in Fig. S9.

Conclusion

We can account for a substantial fraction of the directly coupled coevolving residue pairs that are distant within monomeric structures. The vast majority of short-distance exceptions (5 Å < d_{min} ≤ 15 Å) are likely due to structural variation within a family. At longer distances (d_{min} > 15 Å), the major source of exceptions are interactions across homo-oligomeric interfaces. A substantial fraction of the remaining exceptions are in repeat proteins. We see little evidence for a contribution from long-range allosteric coupling.

These observations have implications for coevolution guided structure prediction. First, repeat protein structure prediction using coevolutionary information should be undertaken only with very carefully constructed MSAs and using methods that explicitly account for the translational symmetry (25). Second, for very large families, contacts should be predicted using the subset of sequences most closely related to the query sequence to reduce the effects of structural variation within the family (alternatively, when evaluating the fit of contacts to models, the extent of direct evolutionary coupling between residue pairs in the 5–15 Å range should be assessed in the neighborhood of the query sequence). Third, coevolving directly coupled residue pairs separated by more than 15 Å in predicted monomer structures should be used to guide homo-oligomer docking calculations.

More generally, our results support the idea that evolution operates on physically interacting residue pairs very much more strongly than residues involved in long-range allosteric networks.

Methods

Protein Structure Datasets. The analyses in this paper are based on the following three datasets prepared from the PDB.

X-ray set. A nonredundant set of 9,846 protein chains was collected by the PISCES server (35) (accessed on August 2, 2016). Only X-ray structures with resolution ≤ 2.0 Å, R-work ≤ 0.25 (R-free does not exceed 0.32), and at least 40 residues per chain were selected. Redundancies were removed at 25% sequence identity cutoff. Out of 9,846 collected proteins, 3,883 have large enough MSAs with M_{eff} > 10^3 (see *Identification of Coevolving Directly Coupled Residue Pairs*).

X-ray set of different resolutions. PISCES server was run several times to select X-ray structures with different crystallographic resolutions ranging from 0 to 8 Å at 1 Å resolution bins. As previously, only chains with at least 40 residues were considered; no R-factor cutoff was imposed. Majority of X-ray structures in the PDB have resolution in the 1–4 Å range. The number of structures selected in ranges [(1;2), (2;3) and (3:4)] was limited to 500 to avoid unnecessary large samples of structures within these resolution bins.

NMR set. PDB (36) was searched for solution NMR structures with ≥40 residues per chain, resulting in an initial set of 5,758 proteins. We selected 922 proteins with highly populated MSAs ($M_{eff} > 10^3$). For each of the selected proteins, we then checked whether an alternative X-ray structure with resolution ≤3.5 Å exists in PDB. The final set contains 222 proteins with both solution NMR and at least one crystallographic structure resolved.

Identification of Coevolving Directly Coupled Residue Pairs. For every protein sequence in each PDB set, an MSA was first constructed by the HHblits program (37) (with parameters –n 8, –e 1E-20, –maxfilt ∞, –neffmax 20, –nodiff, –realign_max ∞) run against the UniProt database (38) and then filtered by HHfilter to exclude highly similar sequences at 90% identity cutoff as well as sequences with coverage <75%. MSA positions that contain >25% of gaps were also eliminated. The GREMLIN pseudolikelihood method (2, 39) was then used to identify directly coupled residue–residue pairs from the resulting MSA. In the GREMLIN model, the probability of a sequence of length L is proportional to $exp(\sum_{i=1}^{L} v_i + \sum_{i\neq j}^{L} w_{ij})$, where v_i is the one body energy of residue i and w_{ij} is the two body energy of residues i,j; the v_i and w_{ij} are obtained by maximizing the L_2-regularized pseudolikelihood of all of the observed sequences. The 21 × 21 matrices of the inferred couplings w_{ij} (20 amino acids + 1 gap) are then converted into single values s^*_{ij} by computing their vector 2-norm for nongap entries: $s^*_{ij} = (\sum_{a=1}^{20}\sum_{b=1}^{20}(w^{ab}_{ij})^2)^{1/2}$. To get the final scores s_{ij}, the average product correction (40) is applied: $s_{ij} = s^*_{ij} - s^*_{i\cdot}s^*_{\cdot j}/s^*_{\cdot\cdot}$, where $s^*_{\cdot j}$, $s^*_{i\cdot}$, and $s^*_{\cdot\cdot}$ are row, column, and full s^*_{ij} matrix averages, respectively. In most figures, we show top 0.5 × (protein length) residue pairs with the largest s_{ij}. We summarize sequence depth and diversity using the effective number of sequences M_{eff}—defined as the sum $M_{eff} = \sum_{i=1}^{N}\frac{1}{m_i}$ over all N sequences in the MSA, where m_i is the number of all sequences in the MSA (including itself), which share at least 80% sequence identity with the current sequence i.

Amino Acid Frequency Distribution Normalization. To obtain normalized frequencies of amino acid pairs of coevolving residues in different distance bins, we first counted the number of GREMLIN directly coupled pairs in the distance bin n^{coev}_{ab} ($d_{min} \le d < d_{max}$) between residues of amino acid types a and b in all 3,883 proteins under study, and then divided by the total number of predicted contacts to get the frequency of the (a,b) residue pair $f^{coev}_{ab} = n^{coev}_{ab} / \sum_{a\ge b} n^{coev}_{ab}$. The expected frequencies are obtained by a similar relation $f^{exp}_{ab} = n^{exp}_{ab} / \sum_{a\ge b} n^{exp}_{ab}$. Here, $n^{exp}_{ab} = \sum_{i=1}^{3883} n^{exp,i}_{ab}$ is an upper estimate of the number of (a,b) contacts under the assumption that every residue of amino acid type a can interact with every other residue of type b within protein i, excluding residue pairs separated by less than six positions in the primary sequence. We then take the negative logarithm of the ratio to obtain $e^{coev}_{ab} = -log(f^{coev}_{ab}/f^{exp}_{ab})$.

ACKNOWLEDGMENTS. This work was supported by National Institute of Health Grants R01GM092802 and R01GM073151, the Howard Hughes Medical Institute, and the Jain Foundation.

1. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108:E1293–E1301.
2. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA* 110:15674–15679.
3. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28:184–190.
4. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys* 87:012707.
5. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6:e28766.
6. Ovchinnikov S, et al. (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* 4:e09248.
7. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci USA* 109:10340–10345.
8. Hayat S, Sander C, Marks DS, Elofsson A (2015) All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. *Proc Natl Acad Sci USA* 112:5413–5418.
9. Hopf TA, et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149:1607–1621.
10. Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* 4:165.
11. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* 3:e02030.
12. Hopf TA, et al. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3: 10.7554/eLife.03430.
13. Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS (2016) Inferring interaction partners from protein sequences. *Proc Natl Acad Sci USA* 113:12180–12185.
14. Tang Y, et al. (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat Methods* 12:751–754.
15. Toth-Petroczy A, et al. (2016) Structured states of disordered proteins from genomic sequences. *Cell* 167:158–170.e12.
16. Dago AE, et al. (2012) Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc Natl Acad Sci USA* 109:E1733–E1742.
17. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc Natl Acad Sci USA* 106:22124–22129.
18. dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN (2015) Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci Rep* 5:13652.
19. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* 106:67–72.
20. Süel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10:59–69.
21. Noivirt-Brik O, Horovitz A, Unger R (2009) Trade-off between positive and negative design of protein stability: From lattice models to real proteins. *PLOS Comput Biol* 5:e1000592.
22. Jacob E, Unger R, Horovitz A (2015) Codon-level information improves predictions of inter-residue contacts in proteins by correlated mutation analysis. *eLife* 4:e08932.
23. Wollenberg KR, Atchley WR (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci USA* 97:3288–3291.
24. Tillier ERM, Lui TWH (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 19:750–755.
25. Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU (2015) Capturing coevolutionary signals inrepeat proteins. *BMC Bioinformatics* 16:207.
26. Rose PW, et al. (2015) The RCSB Protein Data Bank: Views of structural biology for basic and applied research and education. *Nucleic Acids Res* 43:D345–D356.
27. Sutto L, Marsili S, Valencia A, Gervasio FL (2015) From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci USA* 112:13567–13572.
28. Tyrrell R, et al. (1997) The structure of the cofactor-binding fragment of the LysR family member, CysB: A familiar fold with a surprising subunit arrangement. *Structure* 5:1017–1032.
29. Boyaci H, et al. (2016) Structure, regulation, and inhibition of the quorum-sensing signal integrator LuxO. *PLoS Biol* 14:e1002464.
30. Yang M, et al. (2015) Unraveling cholesterol catabolism in Mycobacterium tuberculosis: ChsE4-ChsE5 α2β2 Acyl-CoA dehydrogenase initiates β-oxidation of 3-Oxocholest-4-en-26-oyl CoA. *ACS Infect Dis* 1:110–125.
31. Keatinge-Clay AT, Maltby DA, Medzihradszky KF, Khosla C, Stroud RM (2004) An antibiotic factory caught in action. *Nat Struct Mol Biol* 11:888–893.
32. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295–299.
33. Reynolds KA, McLaughlin RN, Ranganathan R (2011) Hot spots for allosteric regulation on protein surfaces. *Cell* 147:1564–1575.
34. Novinec M, et al. (2014) A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods. *Nat Commun* 5:3287.
35. Wang G, Dunbrack RL, Jr (2005) PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Res* 33(Web Server issue):W94–W98.
36. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242.
37. Remmert M, Biegert A, Hauser A, Söding J (2011) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9:173–175.
38. Magrane M; UniProt Consortium (2011) UniProt Knowledgebase: A hub of integrated protein data. *Database (Oxford)* 2011:bar009.
39. Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ (2011) Learning generative models for protein fold families. *Proteins* 79:1061–1078.
40. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24:333–340.