# Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone

Juan Rodriguez-Rivas[a], Simone Marsili[a,1], David Juan[a,1], and Alfonso Valencia[a]

[a]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, 28029 Madrid, Spain

Edited by Barry Honig, Howard Hughes Medical Institute, Columbia University, New York, NY, and approved November 23, 2016 (received for review August 4, 2016)

**Protein–protein interactions are fundamental for the proper functioning of the cell. As a result, protein interaction surfaces are subject to strong evolutionary constraints. Recent developments have shown that residue coevolution provides accurate predictions of heterodimeric protein interfaces from sequence information. So far these approaches have been limited to the analysis of families of prokaryotic complexes for which large multiple sequence alignments of homologous sequences can be compiled. We explore the hypothesis that coevolution points to structurally conserved contacts at protein–protein interfaces, which can be reliably projected to homologous complexes with distantly related sequences. We introduce a domain-centered protocol to study the interplay between residue coevolution and structural conservation of protein–protein interfaces. We show that sequence-based coevolutionary analysis systematically identifies residue contacts at prokaryotic interfaces that are structurally conserved at the interface of their eukaryotic counterparts. In turn, this allows the prediction of conserved contacts at eukaryotic protein–protein interfaces with high confidence using solely mutational patterns extracted from prokaryotic genomes. Even in the context of high divergence in sequence (the twilight zone), where standard homology modeling of protein complexes is unreliable, our approach provides sequence-based accurate information about specific details of protein interactions at the residue level. Selected examples of the application of prokaryotic coevolutionary analysis to the prediction of eukaryotic interfaces further illustrate the potential of this approach.**

coevolution | protein–protein interaction | protein complex | homology modeling | contact prediction

Cells function as a remarkably synchronized orchestra of finely tuned molecular interactions, and establishing this molecular network has become a major goal of molecular biology. Important methodological and technical advances have led to the identification of a large number of novel protein–protein interactions and to major contributions to our understanding of the functioning of cells and organisms (1, 2). In contrast, and despite relevant advances in EM (3) and crystallography (4), the molecular details of a large number of interactions remain unknown.

When experimental structural data are absent or incomplete, template-based homology modeling of protein complexes represents the most reliable option (5, 6). Similarly to modeling of tertiary structure for single-chain proteins, homology modeling of protein–protein interactions follows a conservation-based approach, in which the quaternary structure of one or more experimentally solved complexes with enough sequence similarity to a target complex (the templates) is projected onto the target. Template-based techniques have provided models for a large number of protein complexes with structurally solved homologous complexes (7–10). Unfortunately, proteins involved in homologous protein dimers tend to systematically preserve their interaction mode only for sequence identities above 30–40% (11). For larger divergences, defining the so-called twilight zone (11, 12), it is not possible to discriminate between complexes having similar or different quaternary structures (11, 13, 14). As a consequence, the quality of the final models strongly depends on the degree of sequence divergence between the target and the available templates.

In contrast to more traditional approaches based on homology detection and sequence conservation, contact prediction supported by residue coevolution (15–25) makes use of sequence variability as an alternative source of information (26). The analysis of residue coevolution has been successfully applied to contact prediction at the interface of protein dimers (27–33), eventually leading to de novo prediction of protein complexes assisted by coevolution (29, 30). In these methods, coevolutionary signals are statistically inferred from the mutational patterns in multiple sequence alignments of interacting proteins. Coevolution-based methods have been shown to be highly reliable predictors of physical contacts in heterodimers, when applied to large protein families with hundreds of nonredundant pairs of interacting proteins (28–30, 34, 35). Unfortunately, these methods cannot be straightforwardly applied to the analysis of eukaryotic complexes where paralogues are abundant, making it very difficult to distinguish their interaction specificities. In consequence, many eukaryotic complexes remain out of reach for both template-based homology modeling (14) and coevolution-guided reconstruction. To address this eukaryotic "blind spot" it is essential to identify long-standing evolutionary constraints that could be used for guiding the reliable projection of structural information from remote homologs.

We test the hypothesis that strong coevolutionary signals identify highly conserved protein–protein contacts, making them particularly adequate for homology-based projections. From a structural modeling point of view, we test whether and when coevolution-based

## Significance

Interacting proteins tend to coevolve through interdependent changes at the interaction interface. This phenomenon leads to patterns of coordinated mutations that can be exploited to systematically predict contacts between interacting proteins in prokaryotes. We explore the hypothesis that coevolving contacts at protein interfaces are preferentially conserved through long evolutionary periods. We demonstrate that coevolving residues in prokaryotes identify interprotein contacts that are particularly well conserved in the corresponding structure of their eukaryotic homologues. Therefore, these contacts have likely been important to maintain protein–protein interactions during evolution. We show that this property can be used to reliably predict interacting residues between eukaryotic proteins with homologues in prokaryotes even if they are very distantly related in sequence.

Author contributions: J.R.-R., S.M., D.J., and A.V. designed research; J.R.-R., S.M., and D.J. performed research; J.R.-R., S.M., and D.J. contributed new reagents/analytic tools; J.R.-R., S.M., and D.J. analyzed data; and J.R.-R., S.M., D.J., and A.V. 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.

[1] To whom correspondence may be addressed. Email: simo.marsili@gmail.com or dadejuan@cnio.es.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1611861114/-/DCSupplemental.

contact predictions can be projected to homologous complexes. In particular, we focus on the paradigmatic problem of contact prediction in eukaryotic complexes based on coevolutionary signals detected in distant alignments of prokaryotic sequences. To this aim, we develop a domain-centered protocol to detect coevolving residues in multiple sequence alignments of prokaryotic complexes and evaluate their accuracy in predicting interprotein contacts in eukaryotes. Our results show that when the signal of coevolution in prokaryotic alignments is strong, conserved interprotein contacts in eukaryotes can be reliably predicted solely using prokaryotic sequence information. These results provide the basis for the application of coevolution to assist de novo structure prediction of eukaryotic complexes with homologs in prokaryotes even when they are highly divergent in sequence, a scenario where standard template-based homology modeling is unfeasible or unreliable.

## Results

**Benchmark Dataset to Study the Interplay Between Coevolution and Structural Conservation at Protein Interfaces.** An initial analysis of the human interactome (*SI Text* and Fig. S1) reveals that for ~17% of human protein–protein interactions each interaction partner shares homology with many prokaryotic sequences. Most of these interactions lack reliable structural information. In the following, we propose that interprotein coevolution based on large collections of prokaryote sequences can be an invaluable source of information about those, otherwise unsolved, protein–protein interaction interfaces.

To investigate the relevance of coevolution in the structural conservation at protein–protein interfaces among highly divergent homologs, we created a dataset of prokaryotic and eukaryotic domain–domain interfaces that integrates structural and coevolutionary data at the residue level. We started from the complete dataset of heterodimeric Pfam (36) domain–domain interactions with known 3D structure (37). Coevolutionary analysis of a protein interface requires a large set of paired sequences from the families of two interacting proteins in the complex (28–30). Distant evolutionary relationships can be often unveiled only at the level of domains (38); therefore, we devised a domain-centered protocol that enables the detection and the alignment of many conserved families of interacting domains (*Materials and Methods* and Fig. S2*A*). We searched for homologous sequences of the interacting domains in 15,271 prokaryotic genomes and built a joint alignment by pairing domains in genomic proximity (29, 30). We used proximity in the genome to identify the existence of a specific physical interaction between two domains (39). This protocol retrieved 559 cases of domain–domain pairs (each corresponding to a unique Pfam family–family pair) having 3D structural evidence of interaction in at least one prokaryotic or eukaryotic species and containing more than 500 sequences in the corresponding (nonredundant) joint alignment (Fig. 1*B* and Fig. S2*B*). For every domain–domain pair in this set, we computed coevolutionary z-scores for all of the interdomain residue–residue pairs that quantify the direct mutual influence between two residues (28) in different domains. Finally, we obtained the set of coevolving interdomain pairs of residues by retaining those pairs for which a strong coevolutionary signal was detected (*Materials and Methods*). We classified each domain–domain interface as intra- or interprotein (Fig. 1*A*) if the majority of paired sequences are codified within the same or different genes, respectively; 401 out of 559 domain pairs were classified as intraprotein and 158 as interprotein (Fig. 1*B* and Fig. S2*C*).

We first classified every 3D structure for each domain–domain interaction as prokaryotic or eukaryotic (*SI Text*). To deal with conformational variability in the available experimental structures, we used two different definitions for the set of contacts forming each domain–domain interface (*Materials and Methods*). First, we defined a comprehensive interface by merging all of the interdomain contacts (defined as residues closer than 8 Å) extracted from all known homologous structures. This definition incorporates information from different biological (e.g., conformational changes) and methodological scenarios. Second, we selected the complex that best aligns to the Pfam profile and defined the corresponding contacts as the representative interface. A comprehensive and a representative interface were computed separately for each domain–domain pair and for both eukaryotic and prokaryotic structures. When not specified otherwise, we will refer to the results obtained from the analysis of comprehensive interfaces; however, all of the analyses were performed in parallel for the representative complexes with similar results. All of the collected data were integrated in a dataset (Fig. 1) of 559 domain interactions with their interdomain coevolving residues and their corresponding prokaryotic and/or eukaryotic structural interfaces.

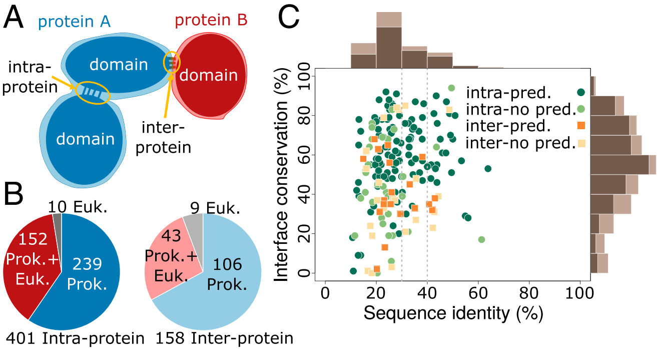

**Fig. 1.** Summary of the coevolutionary/structural dataset generated by our protocol. (*A*) The two kinds of domain–domain interactions discussed in the text: In intraprotein cases the two domains are codified within the same gene; in interprotein cases they are found in different genes. (*B*) Dataset composition according to inter- or intraprotein classification and the availability of 3D structure in prokaryotes and eukaryotes. (*C*) Percentages of interface conservation and sequence identities for 152 intraprotein cases and 43 interprotein cases. Interface conservation was calculated as the proportion of contacts in prokaryotic interfaces that are also in contact in eukaryotes. Sequence identities were calculated as the proportion of identical amino acids between the best aligned prokaryotic PDB sequence and the best aligned eukaryotic PDB sequence (*SI Text*). A domain–domain interaction was labeled as predicted when at least an interdomain pair of residues was classified as coevolving (*Materials and Methods*). Marginal histograms computed on the whole set of cases are in light brown; marginal histograms for predicted cases are in dark brown.

Our dataset includes 43 interprotein and 152 intraprotein cases (Fig. 1*B*) with structure in both prokaryotes and eukaryotes. For these cases, we quantified the structural interface conservation as the proportion of prokaryotic contacts that are also in contact in their homologous sites in eukaryotes. In this subset, 66% (129 out of 195) of the cases correspond to sequence identities below 30%. Complexes with sequence identities below 30–40% have highly variable values of interface conservation, and conserved interfaces cannot be identified using sequence identity alone (see Fig. 1*C* and Fig. S2*D* for representative interfaces). This variability reflects the difficulties associated to accurate template-based homology modeling in the twilight zone. In our dataset, a naive extrapolation of contacts from prokaryotes to eukaryotes would result in highly unreliable predictions, due to the large divergences. This set of homologous interfaces provides the basis for investigating the structural conservation of coevolving residues between prokaryotic and eukaryotic interfaces even at large sequence distances.

**Coevolving Residues Identify Structurally Conserved Contacts at Protein Interfaces.** We detected strong coevolutionary signals in 20 out of 43 interprotein cases (and in 121 out of 152 intraprotein cases). The proportion of cases with predictions (strong coevolutionary signals) is higher when the structural interface conservation is larger (Fig. 1*C*). This suggests that coevolution is indicative of a greater structural conservation. To gain further insight, we studied the relationship between structural interface conservation and the degree of coevolution detected in each case. To this aim, we calculated a score, called interface coupling, by averaging the z-score of the five strongest interdomain coevolving pairs (33). As shown in Fig. 2*A*, the level of interface coupling determines a lower bound for interface conservation (i.e., the stronger the interface coupling, the higher the minimal interface conservation observed in our dataset). Moreover, large interface coupling values consistently identify domain–domain pairs that interact via a single 3D interaction topology (*SI Text*), suggesting that a single, conserved interface may be an important factor in explaining strong domain–domain coevolution.

A comparison between homologous sites in eukaryotic and prokaryotic structures clearly reveals that pairs of residues that are coevolving and in contact in prokaryotes (interprotein: 52 contacts out of 56 coevolving pairs; intraprotein: 1,070 contacts out of 1,107 coevolving pairs) are systematically found in contact in the 3D structures of the corresponding eukaryotic homologs (Fig. 2*B*). This effect is highly significant compared with the proportion of prokaryotic contacts shared with a eukaryotic homolog expected by chance ($P < 10^{-10}$, one-tailed Fisher exact test for both interprotein and intraprotein cases; *SI Text*) and it is robust to different definitions of coevolution and contacts (Fig. S3 *A* and *B*). The analysis of representative interfaces leads to the same conclusion (Fig. S3 *C* and *D*). Moreover, the structural conservation of coevolving contacts is much higher than expected by chance after considering the conservation in sequence of the coevolving residues (*SI Text* and Fig. S3 *E* and *F*). Remarkably, focusing on the difficult cases in the twilight zone (less than 30% sequence identities, 29 interprotein and 100 intraprotein) we also found a highly significant enrichment in conserved coevolving contacts (Fig. S4, $P < 10^{-6}$, one-tailed Fisher exact test for both interprotein and intraprotein cases, and *SI Text*).

In detail, the proportion of interprotein contacts conserved in prokaryotic and eukaryotic interfaces (37%) increases up to 91% (48 conserved contacts out of 53 coevolving pairs in contact in prokaryotes or eukaryotes) for pairs of coevolving residues (Fig. 2*B*). Interestingly, three out of the four coevolving pairs that apparently are not conserved correspond to residue pairs that are spatially close in the eukaryotic structure (less than 10 Å). For the cases in the twilight zone, 23 out of 25 coevolving contacts are conserved and one of the remaining pairs is at 8.1 Å in eukaryotes. Intraprotein interfaces follow the same trend: The proportion of conserved contacts goes from 50 to 96% for coevolving pairs (Fig. 2*B*; 1,039 conserved contacts out of 1,082 coevolving contacts). Again, we found that coevolving contacts are highly conserved even for interfaces in the twilight zone (583 conserved out of 615). These results are robust to the specific measure of sequence divergence (*SI Text* and Fig. S5). They clearly prove that coevolving contacts have been preferentially conserved during the course of evolution, validating our hypothesis that coevolution identifies structurally conserved contacts. Moreover, when applied to coevolving pairs of residues at prokaryotic interfaces, this property should allow one to predict interface contacts in eukaryotic proteins, in a wide range of evolutionary distances, including the twilight zone.

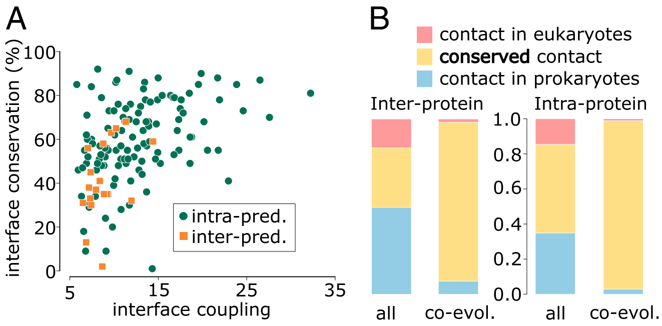

**Fig. 2.** (*A*) Relation between interface structural conservation (defined as in Fig. 1) and interface coupling (the average z-score of the five strongest interdomain coevolving pairs) for 20 interprotein and 121 intraprotein domain–domain interactions with contact predictions (i.e., strong coevolutionary signals) and structurally solved prokaryotic and eukaryotic homologous complexes. (*B*) Proportion of conserved contacts at the homologous sites of prokaryotic/eukaryotic complexes, computed from the total set of contacts and the subset of coevolving contacts, and for inter- and intraprotein cases. Blue: contacts found in a prokaryotic complex and not in the homologous eukaryotic complex. Red: contacts found in a eukaryotic complex and not present in the homologous prokaryotic complex. Yellow: contacts shared by prokaryotic and eukaryotic complexes. Forty-eight out of 52 coevolving contacts in interprotein complexes and 1,039 out of 1,070 coevolving contacts in intraprotein complexes are shared by prokaryotes and eukaryotes.

**Contact Prediction at Eukaryotic Protein Interfaces.** We assessed the precision of prokaryotic coevolving pairs in predicting contacts in prokaryotic and eukaryotic structures for cases with predictions in structurally solved regions, both in prokaryotic and eukaryotic interfaces (19 interprotein and 120 intraprotein). The vast majority of these cases have a high precision in the two superkingdoms (Fig. S6). Only 1 out of 19 interprotein cases in prokaryotes (6 out of 120 in intraprotein) was predicted with a precision lower than 0.6 (Fig. S6). For eukaryotes, these numbers are only slightly higher with 2 out of 19 interprotein (11 out of 120 in intraprotein; Fig. S6). The few additional cases with low precision found for representative interfaces are evenly distributed in prokaryotes and eukaryotes, suggesting that they are not related with the projection procedure (Fig. S6). Most false positives occur in cases within the twilight zone with low structural interface conservation (Fig. S5 *A* and *C*). This low structural conservation could result in poorly aligned eukaryotic sequences. We evaluated the impact of alignment quality on the projection of contact predictions from prokaryotes to eukaryotes by computing the averaged expected alignment accuracy for residues at the eukaryotic homologous sites of the prokaryotic interface (*SI Text*). Indeed, most of the cases with low-quality predictions in eukaryotes but not in prokaryotes correspond to low-quality sequence alignments, both for comprehensive (Fig. S7 *A* and *B*) and representative interfaces (Fig. S7 *C* and *D*).

As discussed above, the high reliability of coevolution as a predictor of contacts in prokaryotes and the preferential conservation of coevolving contacts allows one to predict contacts in eukaryotes without any prior structural information. To further assess this point, we quantified the quality of eukaryotic contact prediction for all cases in which a eukaryotic structure was available to check the resulting predictions (51 interprotein and 162 intraprotein; Fig. 1*B*). We detected 62 coevolving pairs in 22 interprotein cases (approximately three predictions per case) and 1,140 pairs in 124 intraprotein cases (approximately nine per case). We found that the precision in eukaryotes is very high both in interprotein (precision = 0.81, Fig. 3*A*) and in intraprotein cases (precision = 0.95, Fig. 3*B*) and it is only slightly lower than the precision obtained in prokaryotes (Fig. 3 *C* and *D*; precision interprotein = 0.86 and precision intraprotein = 0.95). We repeated the analysis after removing cases with low alignment quality, using a filter based on the pairs of coevolving residues (*SI Text*). In line with the discussion in the previous paragraph, the results suggest that an a priori filter can detect cases in which projected predictions have a lower precision (Fig. S7 *E* and *F* and Table S1).

**Application to Mammalian Complexes.** The pyruvate dehydrogenase complex, responsible for the catalysis of pyruvate to acetyl-CoA and $CO_2$, is the complex in our dataset with the highest interface coupling in eukaryotes. Its E1 component forms a homodimer of heterodimers of its α and β subunits (40). The coevolving contacts detected by our protocol are distributed over the interface between the two subunits and are well conserved in the eukaryotic interface.

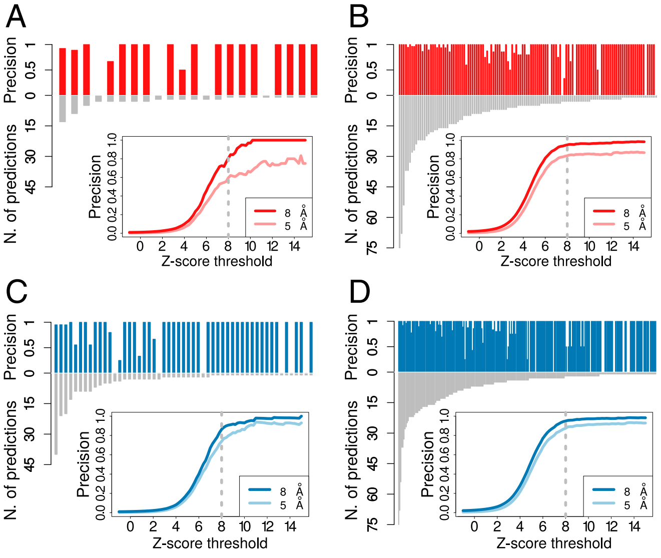

**Fig. 3.** Precision and number of predictions for each of the 22 interprotein (*A*) and 124 intraprotein (*B*) domain–domain interfaces in eukaryotes with at least one detected coevolving pairs. Coevolutionary z-scores were computed for all residue pairs for each domain–domain interface, and those pairs having a z-score larger than 8 were classified as coevolving (*Materials and Methods*). (*Insets*) The fraction of coevolving pairs closer than 8 Å (dark blue line) and 5 Å (light blue line) as a function of the threshold for the z-score. The dashed gray lines highlight the reference z-score threshold (z-score = 8) for detection of coevolution. We detected 62 coevolving pairs for interprotein interfaces (approximately three predictions per case on average) with an average precision of 0.81 for a contact distance of 8 Å (and 0.6 at 5 Å), and 1,140 pairs in the intraprotein case (approximately nine per case, on average) with a precision of 0.95 at 8 Å (and 0.83 at 5 Å). Precision and number of predictions for each of the 53 interprotein (*C*) and 245 intraprotein (*D*) domain–domain interfaces in prokaryotes with at least one detected coevolving pair. We obtained a precision of 0.86 at 8 Å and 0.74 at 5 Å for interprotein domain–domain interfaces (respectively 0.95 and 0.87 in the case of intraprotein interfaces).

Among the 13 coevolving pairs, 10 are in contact at the subunits interface of the branched-chain 2-oxo acid dehydrogenase remote homolog in the *Thermus thermophilus* structure (Fig. 4*A*) and are conserved in the human pyruvate dehydrogenase complex (Fig. 4*B*). Two out of three apparent false positives do actually correspond to contacts at the homodimer interface. These results show that coevolution has been key in the conservation of quaternary structure in the pyruvate dehydrogenase E1 component.

The translocon complex, one of the main mechanisms of transporting proteins across the membrane, is a good example of a conserved mode of interaction with very low sequence conservation. The α and γ subunits of the mammalian Sec61 are homologous to the bacterial proteins SecY and SecE, respectively. Despite the low sequence identity between these proteins in *T. thermophilus* and *Canis lupus* (18.8% SecY/Sec61α and 10.5% SecE/Sec61γ), and a strong structural divergence of the domains, two out of three coevolving contacts (Fig. 4*C*) have been conserved (Fig. 4*D*). In fact, seven out the nine residue pairs in the crystal structure for *T. thermophilus* with the highest coevolutionary z-scores are in contact and six of them are structurally conserved in *C. lupus.* The lower resolution (6.8 Å) of the available EM in *C. lupus* introduces some uncertainty on the definition of the interface. Still, our predictions support the overall arrangement of the interaction given in this experimental structure and highlight the potential use of our approach to refine atomic details of cryo-EM experiments.

Among the 20 cases of interprotein interfaces with structural information in both eukaryotes and prokaryotes and with strong coevolutionary signals, we only detected one case where a strong coevolutionary signal does not go together with an, at least partially, conserved interface: the interaction between the α and β subunits of the phenylalanyl tRNA synthetase (PheRS). PheRS catalyzes the attachment of a phenylalanine amino acid to its cognate transfer RNA molecule. Despite important differences between the prokaryotic and the eukaryotic PheRS complexes (41), several homologous domains can be found in both the α subunit (core catalytic domain) and β subunit (B5 and B3/4 domains) between prokaryotes and eukaryotes (Fig. S8 *A* and *B*). The coevolutionary analysis of the interaction between the core catalytic domain and the B3/4 domain detects two coevolving pairs located at the *T. thermophilus* interface (Fig. S8*C*). These pairs are no longer aligned in the human B3/4 domain due to an insertion in the *T. thermophilus* PheRS compared with the human cytosolic complex as proved by a structural alignment (Fig. S8*D*) and therefore cannot be projected to the corresponding interface. Notably, this interacting region in human is deleted just at one of the two turns where the interaction takes place (Fig. S8*D*). Moreover, the α subunit also interacts with the

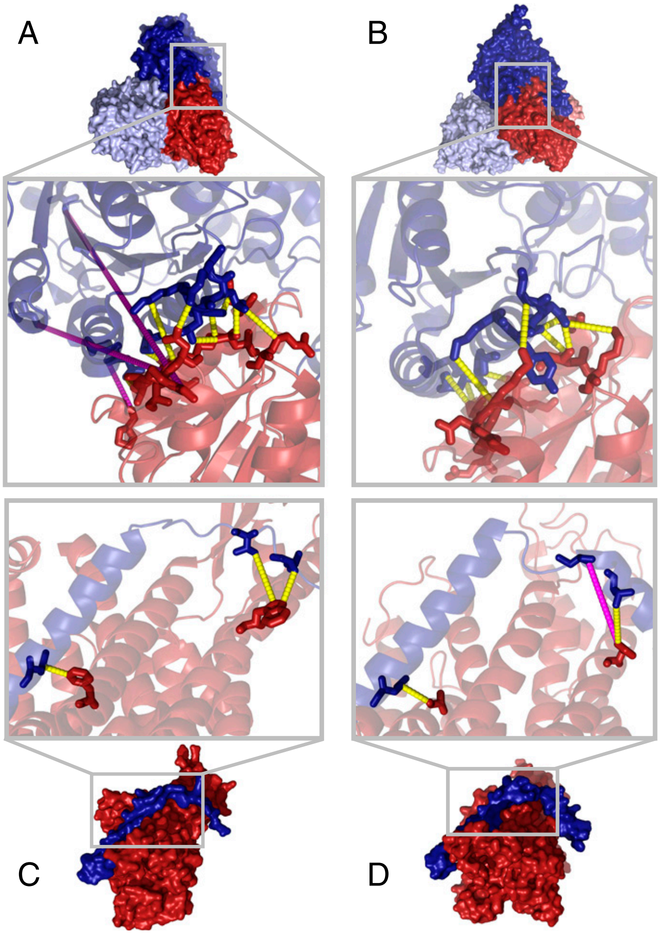

**Fig. 4.** (*A* and *B*) E1 component of the branched-chain 2-oxo acid dehydrogenase in *T. thermophilus* [*A*, PDB ID code 1UMB (61)] and the human mitochondrial pyruvate dehydrogenase E1 component of the pyruvate dehydrogenase complex [*B*, PDB ID code 3EXI (40)]. In the zoomed inset coevolving pairs of residues are shown as sticks and connected by dashed lines (yellow if they are in contact, magenta otherwise). Ten contacts out of 13 coevolving pairs are detected at the *T. thermophilus* interface (*A*). Eleven out of these 13 pairs can be mapped to the human complex where they are all in contact (*B*). (*C* and *D*) Protein transportation channels across bacterial plasma membrane SecYE in *T. thermophilus* X-ray structure of SecY (in dark red) in complex with SecE (in dark blue) [*C*, PDB ID code 2ZJS (62)]. Eukaryotic endoplasmic reticulum Sec61 in *C. lupus* EM structure of Sec61 complex with its α subunit in dark red and γ subunit in dark blue [*D*, PDB ID code 4CG7 (63)]. The three coevolving pairs (drawn as sticks) are in contact in *T. thermophilus* and two of them are conserved in the *C. lupus* structure.

B5 domain of the β subunit and the three coevolving contacts at the prokaryotic interface are completely preserved in the human PheRS (Fig. S8 *G* and *H*). This example illustrates that even after a drastic event, such as removal of a region at the interface in one of the interacting proteins, the remaining coevolving residues can keep pointing to the real interfaces.

## Discussion

In this work we introduce and validate an important property of coevolving contacts at protein interfaces: their propensity to be preferentially conserved at large evolutionary distances. This behavior is confirmed by the analysis of coevolving residues between domains in 15,271 prokaryotic genomes and their homologous sites in 3D structures of eukaryotic complexes. This previously unrecognized aspect of the evolution of protein interfaces highlights the important role of coevolving residues in maintaining quaternary structure and protein–protein interactions. As a first and important consequence of this property, we show that contacts at eukaryotic interfaces can be predicted with high accuracy using solely prokaryotic sequence data, both for protein–protein and for domain–domain interfaces. We tested these conclusions by analyzing a large dataset of prokaryotic/eukaryotic interfaces with a domain-centered protocol. We were able to predict contacts in interprotein eukaryotic complexes with a mean precision >0.8 (Fig. 3 and Table S1). This result is particularly relevant taking into account that this level of accuracy was attained for predictions of contacts in highly divergent complexes (sequence identities lower than 30%), where standard homology modeling is hardly useful. We have also shown that the few errors in these prokaryote–eukaryote projections are generally associated to cases with low structural conservation that can be detected a priori by checking the alignment quality. Moreover, we extended this analysis to domain–domain contact predictions, showing that intraprotein interfaces exhibit even stronger coevolutionary signals, leading to an increased precision in contact prediction. The analysis protocol we propose relies on sequence data only. As a consequence, our strategy can provide useful information on a protein interface both in remote homology-based complex reconstruction and when no structural template is available, and it is inherently complementary to current methods based on the analysis of structural similarity (42) or sequence similarity (6, 7, 10, 43) to a set of available templates.

The main obstacle to structural modeling of eukaryotic protein complexes by means of coevolution-based approaches is the need for a large number of homologous interactions to permit statistical analysis. Eukaryotic complexes present a paradoxical scenario: Large families of eukaryotic proteins are the result of duplication-based expansions, but these duplications make uncertain which paralogues of one family interact with which ones of the other. In the future, improvements aimed to disentangle the network of paralogous interactions will be fundamental to deal with eukaryotic interactions (44–47). Our approach, based on preferential conservation, tackles this problem for proteins with prokaryotic homologs by looking at very divergent, well-populated, and easy-to-couple pairs of interacting prokaryotic proteins. This strategy cannot be applied in some specific contexts; for example, our approach cannot cope with recently evolved interactions, or with disordered—and difficult-to-align—interfacial regions. However, we found enough prokaryotic homologs to perform these analyses for 31,707 experimentally known human interactions without reliable structural templates (an estimated 15% of the human interactome; *SI Text*), suggesting that large-scale prediction of contacts at eukaryotic interfaces is actually possible. The resulting projected contact predictions represent a source of structural information that can be easily incorporated in integrative structural computational methods (48–52) or used to improve the scope of the successful methods that already incorporate coevolutionary information from closer homologs (24, 29, 30, 53–55). At a more general level, these results indicate that coevolving contacts have played a fundamental role in the evolution of interacting surfaces as structurally conserved anchor points.

## Materials and Methods

**Dataset and Joint Alignments.** We extracted a list of 4,556 heterodimeric pairs of interacting Pfam domains with solved 3D structures [3did database (37)]. For each pair of Pfam domains, our protocol searched for proteins containing members of at least one of these two Pfam domain families in 15,271 prokaryotic genomes (56) using HMMER software (version 3.0) (57, 58). Two domains were paired if they were in the same protein, adjacent proteins, or when both had no other paralogous in the corresponding genome. From this set of pairs, a joint alignment was built by aligning each domain to its corresponding Pfam profile. We next applied a stringent set of quality controls (*SI Text*) including alignment coverage (>80%) and redundancy (<80%). Insertions were removed by considering only residues that were assigned to match states of the HMM model. We retained 559 domain–domain interactions with a large number of nonredundant sequences (>500) for further analyses. Each pair of Pfam domains was classified as intra- or interprotein if the majority of paired sequences are codified within the same or different genes, respectively (Fig. S2*C*).

**Calculation of Coevolutionary Z-Scores.** We retrieved the coevolutionary z-scores by performing a (multinomial) logistic regression of each position in the joint alignment on the remaining positions, a standard network inference strategy (59) that has already been adopted for the analysis of monomeric protein sequences (23) in combination with $l_2$ regularization. For each residue–residue pair we computed the corrected Frobenius norm score (23), a measure of statistical coupling between residues, from the (symmetrized) estimates of the coupling parameters. Finally, these values were standardized to reduce heterogeneity between cases and used as coevolutionary z-scores. An interdomain pair of residues was considered as coevolving when its coevolutionary z-score was higher than a threshold value of 8 (see *SI Text* for details).

**Interface Definition and Contact Prediction Evaluation.** For a given pair of Pfam families, we retrieved, from the Protein Data Bank (PDB) (60), the biological unit for all of the structures of complexes in which two members of the families are in physical contact. The PDB identifiers were retrieved from the 3did annotations. For structures with multiple biological units, we selected the one labeled as first. We extracted the PDB sequences and aligned them to their corresponding Pfam domains. We classified each PDB structure as eukaryotic or prokaryotic (*SI Text*). We defined a comprehensive and a representative interface in one or both superkingdoms depending on the availability of at least one 3D structure in prokaryotes and/or eukaryotes. To that aim, for each pair of Pfam domains (*i*) we recovered the interdomain contacts in all PDB containing those Pfam domains. We used a distance of 8 Å between any heavy atom as the distance threshold for contacts (29, 30). Other contact definitions were used and are shown when appropriate. (*ii*) We mapped all PDB positions to their corresponding positions in the Pfam HMM profiles. (*iii*) We selected the most reliably aligned PDB (according to the alignment bitscores; *SI Text*) as the representative complex in prokaryotes and eukaryotes. (*iv*) Using the alignments of PDB sequences against both Pfam domains, we retrieved the set of PDBs with a 98% or higher percentage of sequence identity with respect to the representative complex. The representative interface is composed by the collection of contacts of the PDBs in this latter set, whereas the comprehensive interface is composed by all of the contacts found in the PDBs containing the Pfam domains. Both interfaces were separately computed for eukaryotes and prokaryotes. Only pairs of interdomain residues that were both aligned and having geometric coordinates in at least one PDB file were used to compute the precision of contact predictions.

**ACKNOWLEDGMENTS.** We thank F. Abascal and M. L. Tress for helpful discussions. This work was supported by Spanish Ministry of Economy and Competitiveness Projects BFU2015-71241-R and BIO2012-40205, cofunded by the European Regional Development Fund.

1. Parrish JR, Gulyas KD, Finley RLJ, Jr (2006) Yeast two-hybrid contributions to interactome mapping. *Curr Opin Biotechnol* 17(4):387–393.

2. Lage K (2014) Protein-protein interactions and genetic diseases: The interactome. *Biochim Biophys Acta* 1842(10):1971–1980.

3. Nogales E (2016) The development of cryo-EM into a mainstream structural biology technique. *Nat Methods* 13(1):24–27.
4. Barty A, Küpper J, Chapman HN (2013) Molecular imaging using X-ray free-electron lasers. *Annu Rev Phys Chem* 64:415–435.
5. Aloy P, Pichaud M, Russell RB (2005) Protein complexes: Structure prediction challenges for the 21st century. *Curr Opin Struct Biol* 15(1):15–22.
6. Szilagyi A, Zhang Y (2014) Template-based structure modeling of protein-protein interactions. *Curr Opin Struct Biol* 24:10–23.
7. Kundrotas PJ, Zhu Z, Janin J, Vakser IA (2012) Templates are available to model nearly all complexes of structurally characterized proteins. *Proc Natl Acad Sci USA* 109(24): 9438–9441.
8. Andreani J, Faure G, Guerois R (2012) Versatility and invariance in the evolution of homologous heteromeric interfaces. *PLOS Comput Biol* 8(8):e1002677.
9. Tyagi M, Hashimoto K, Shoemaker BA, Wuchty S, Panchenko AR (2012) Large-scale mapping of human protein interactome using structural complexes. *EMBO Rep* 13(3): 266–271.
10. Mosca R, Céol A, Aloy P (2013) Interactome3D: Adding structural details to protein networks. *Nat Methods* 10(1):47–53.
11. Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. *J Mol Biol* 332(5):989–998.
12. Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng* 12(2):85–94.
13. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA (2008) Assembly reflects evolution of protein complexes. *Nature* 453(7199):1262–1265.
14. Negroni J, Mosca R, Aloy P (2014) Assessing the applicability of template-based protein docking in the twilight zone. *Structure* 22(9):1356–1362.
15. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18(4):309–317.
16. Lapedes AS, Giraud BG, And LL, Stormo GD (1999) Correlated mutations in models of protein sequences: Phylogenetic and structural effects. *Stat Mol Biol* 33(May):236–256.
17. Lapedes A, Giraud B, Jarzynski C (2002) Using sequence alignments to predict protein structure and stability with high accuracy. arXiv:1207.2484.
18. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLOS Comput Biol* 6(1):e1000633.
19. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108(49):E1293–E1301.
20. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. *Proteins* 79(4):1061–1078.
21. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.
22. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotechnol* 30(11):1072–1080.
23. Ekeberg M, Hartonen T, Aurell E (2014) Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J Comput Phys* 276:341–356.
24. Michel M, et al. (2014) PconsFold: Improved contact predictions improve protein models. *Bioinformatics* 30(17):i482–i488.
25. 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(44): 13567–13572.
26. de Juan D, Pazos F, Valencia A (2013) Emerging methods in protein co-evolution. *Nat Rev Genet* 14(4):249–261.
27. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. *J Mol Biol* 271(4):511–523.
28. 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(1):67–72.
29. Hopf TA, et al. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3:e03430.
30. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* 3: e02030.
31. Cheng RR, Morcos F, Levine H, Onuchic JN (2014) Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc Natl Acad Sci USA* 111(5):E563–E571.
32. Malinverni D, Marsili S, Barducci A, De Los Rios P (2015) Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones. *PLOS Comput Biol* 11(6):e1004262.
33. Feinauer C, Szurmant H, Weigt M, Pagnani A (2016) Inter-protein sequence co-evolution predicts known physical interactions in bacterial ribosomes and the trp operon. *PLoS One* 11(2):e0149166.
34. Hopf TA, et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149(7):1607–1621.
35. Tamir S, et al. (2014) Integrated strategy reveals the protein interface between cancer targets Bcl-2 and NAF-1. *Proc Natl Acad Sci USA* 111(14):5177–5182.
36. Finn RD, et al. (2014) Pfam: The protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230.
37. Mosca R, Céol A, Stein A, Olivella R, Aloy P (2014) 3did: A catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* 42(Database issue):D374–D379.
38. Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A (2008) Arrangements in the modular evolution of proteins. *Trends Biochem Sci* 33(9):444–451.
39. Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem Sci* 23(9):324–328.
40. Kato M, et al. (2008) Structural basis for inactivation of the human pyruvate dehydrogenase complex by phosphorylation: Role of disordered phosphorylation loops. *Structure* 16(12):1849–1859.
41. Finarov I, Moor N, Kessler N, Klipcan L, Safro MG (2010) Structure of human cytosolic phenylalanyl-tRNA synthetase: Evidence for kingdom-specific design of the active sites and tRNA binding patterns. *Structure* 18(3):343–353.
42. Tuncbag N, Gursoy A, Nussinov R, Keskin O (2011) Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc* 6(9):1341–1354.
43. Andreani J, Guerois R (2014) Evolution of protein interactions: From interactomes to interfaces. *Arch Biochem Biophys* 554:65–75.
44. Ramani AK, Marcotte EM (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol* 327(1):273–284.
45. Izarzugaza JMG, Juan D, Pons C, Pazos F, Valencia A (2008) Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics* 9:35.
46. Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A (2016) Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc Natl Acad Sci USA* 113(43):12186–12191.
47. Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS (2016) Inferring interaction partners from protein sequences. *Proc Natl Acad Sci USA* 113(43):12180–12185.
48. Gray JJ, et al. (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* 331(1):281–299.
49. Mukherjee S, Zhang Y (2011) Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure* 19(7):955–966.
50. Russel D, et al. (2012) Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* 10(1):e1001244.
51. Tang Y, et al. (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat Methods* 12(8):751–754.
52. Rodrigues JPGLM, et al. (2013) Defining the limits of homology modeling in information-driven protein docking. *Proteins* 81(12):2119–2128.
53. 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(52):22124–22129.
54. Hosur R, et al. (2012) A computational framework for boosting confidence in high-throughput protein-protein interaction datasets. *Genome Biol* 13(8):R76.
55. Andreani J, Faure G, Guerois R (2013) InterEvScore: A novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics* 29(14):1742–1749.
56. Kersey PJ, et al. (2014) Ensembl Genomes 2013: Scaling up access to genome-wide data. *Nucleic Acids Res* 42:D546–D552.
57. Eddy SR (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLOS Comput Biol* 4(5):e1000069.
58. HMMER 3.0 (March 2010) HMMER: Biosequence analysis using profile hidden Markov models. Available at hmmer.org/.
59. Ravikumar P, Wainwright MJ, Lafferty JD (2010) High-dimensional Ising model selection using $\ell$1-regularized logistic regression. *Ann Stat* 38(3):1287–1319.
60. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10(12):980.
61. Nakai T, et al. (2004) Ligand-induced conformational changes and a reaction intermediate in branched-chain 2-oxo acid dehydrogenase (E1) from Thermus thermophilus HB8, as revealed by X-ray crystallography. *J Mol Biol* 337(4):1011–1033.
62. Tsukazaki T, et al. (2008) Conformational transition of Sec machinery inferred from bacterial SecYE structures. *Nature* 455(7215):988–991.
63. Gogala M, et al. (2014) Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion. *Nature* 506(7486):107–110.
64. Chatr-Aryamontri A, et al. (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 43(Database issue, D1):D470–D478.
65. Rodriguez JM, Carro A, Valencia A, Tress ML (2015) APPRIS WebServer and WebServices. *Nucleic Acids Res* 43(W1):W455-9.
66. Velankar S, et al. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* 41(Database issue, D1):D483–D489.
67. Sayers EW, et al. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 37(Database issue, Suppl 1):D5–D15.
68. Bairoch A, Apweiler R (1997) The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res* 25(1):31–36.
69. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
70. Li W, Jaroszewski L, Godzik A (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 18(1):77–82.
71. Averick BM, Richard GC, Moré JJ (1993) MINPACK-2 Project. November 1993 (Argonne National Laboratory, Lemont, IL and University of Minnesota, Minneapolis.
72. Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T (2014) PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res* 42(Database issue, D1):D897–D902.
73. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32(1):268–274.
74. Goldgur Y, et al. (1997) The crystal structure of phenylalanyl-tRNA synthetase from Thermus thermophilus complexed with cognate tRNAPhe. *Structure* 5(1):59–68.