# PDB-wide identification of biological assemblies from conserved quaternary structure geometry

Sucharita Dey, David Ritchie, Emmanuel D Levy

# PDB-wide identification of biological assemblies from conserved quaternary structure geometry

Sucharita Dey[1], David W. Ritchie[2], Emmanuel D. Levy[1]*

1. Weizmann Institute of Science, Department of Structural Biology, Rehovot, Israel.

2. Inria Nancy, Villers-les-Nancy, France

* Correspondence: emmanuel.levy@weizmann.ac.il

**Document statistics**

Abstract: 151

Main text: 2978

Methods: 2346

Number of figure: 4

Number of tables: 0

Number of supplementary figures: 7

Number of supplementary tables: 5

**Abstract**

Protein structures are key to understanding bio-molecular mechanisms and diseases, yet their interpretation is hampered by limited knowledge of their biologically relevant quaternary structures (QSs). A critical challenge in obtaining QSs from crystallographic data is to distinguish biological interfaces from crystal packing contacts. We tackled this challenge with two strategies for aligning and comparing QS states, both across homologs (QSalign), and across data repositories (QSbio). QS conservation across homologs was a remarkably strong predictor of biological relevance and allowed annotating of >80,000 biological QS states. QS conservation across methods enabled us to create a meta-predictor, QSbio, from which we inferred confidence estimates for >110,000 assemblies in the Protein Data Bank, which approach the accuracy of manual curation. Based on the dataset obtained, we analyzed interaction interfaces among pairs of structurally conserved QSs. This revealed a striking plasticity of interfaces, which can maintain a similar interaction geometry through widely different chemical properties.

**INTRODUCTION**

A majority of proteins adopt a quaternary structure (QS) by interacting with copies of themselves, thereby forming homomers[1-6]. These protein complexes are present in virtually all biological processes[7-9]. For example, the well-known oncogene p53 is a homomer, as are many transcription factors. In apoptosis, caspase-9 acts as a trigger after homodimer formation[10], and it is the homotrimeric scaffold of the tumor necrosis factor that induces cell death[11]. Metabolic pathways are no exception, with most glycolytic enzymes forming homomers. Besides being ubiquitous, homomers are a common theme in cellular machines, with ATP synthases, proteasomes, chaperones, photosystems, exosomes and nucleosomes having evolved from ancestral homomers[12].

Knowledge of biological QS states is thus key to contribute to the understanding of protein function and evolution. Currently, the richest source of information on protein QS is the Protein Data Bank (PDB), the repository for structural information obtained by X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy[13,14]. Structures solved by X-ray crystallography represent, by far, the majority of the data with now over 118,000 structures deposited. However, X-ray crystallography provides the coordinates of only the asymmetric unit (ASU) within the unit cell of a crystal. At the molecular level, a crystal is formed as an infinite lattice of unit cells (**Fig. 1**), and a protein complex may be made from one or more ASUs, or indeed from parts of several ASUs[15]. A critical challenge underlying analysis and interpretation of protein structures is to discriminate between fortuitous protein-protein contacts that arise from crystal packing and evolved contacts that make up the biologically relevant QS (**Fig. 1**).

Much computational work has been devoted to addressing this challenge. Several properties were found to be important and can be separated into two broad classes. The first class comprises physicochemical and shape properties of protein interfaces and includes size[15-21], planarity[22], atomic packing[19,21,23], predicted binding free energy[15,20], secondary structure[24], as well as the composition[18,19,21-23,25] or entropy of amino acids[26]. The second class is evolutionary conservation, either of individual residues[27-31] or of the geometry of the interaction interface, which can be assessed across different crystal forms[32,33] or across homologs[34,35].

Historically, the protein QS (PQS) server[15] was the first resource to provide information on likely biological assemblies of structures deposited in the PDB and has since been succeeded by another method called PISA[20]. While the predictions of these servers are generally correct, the hard nature of the problem means that ~15% of QS states from the PDB do not reflect a biologically relevant state[20,36]. This limitation prompted the development of PiQSi[36], which contains 1,434 high-confidence and non-redundant manual annotations. However, manual curation is tedious and cannot be applied at a PDB-wide scale, which motivated us to develop novel automated methods approaching the accuracy of manual curation.

Approaches to superpose structurally related protein chains within arbitrary multi-component complexes have been previously reported[37-39]. Here, we focus on the identification and analysis of homologous QSs

sharing the same number of subunits, on the premise that conservation of QS provides a strong indicator for biological relevance (**Fig. 1b**). We first developed a heuristic algorithm named QSalign, which structurally aligns complete QSs and infers those conserved as being biologically relevant. QSalign confirmed the biological relevance of 31,257 QS states from the PDB, with an estimated error rate as low as 2%. To annotate monomeric proteins, we inverted the QSalign approach and used the absence of homomeric homologs as predictive information of a protein being monomeric. We called this method anti-QSalign and were able to annotate 46,877 monomeric proteins with it. In addition to comparing QSs across species, we integrated QS conservation across methods (PISA[20], EPPIC[30], and QSalign/anti-QSalign). The resulting predictor, named QSbio (www.QSbio.org), approached the accuracy of manual curation. We finally illustrate the use of QSalign in an analysis of interface properties of structurally similar but evolutionary distant complexes, which reflects the plasticity of interface properties among remote structural homologs.

## RESULTS

### Inferring biological relevance by multi-chain superposition of protein QS

Evolutionary conservation is a powerful means to assess function and can be applied in the context of QS. Previous work indeed reported that structural conservation of interface geometry is a strong indicator of functional relevance[33]. Among dimers, interface conservation directly reflects QS conservation because a single interface is involved. In higher-order oligomers, however, structural similarity cannot be inferred readily from similarity of pairwise interfaces (**Supplementary Fig. S1**), which prompted us to assess the structural similarity of full complexes, without decomposing them into pairwise interfaces.

The task of superposing full protein complexes presented two main obstacles. A first hurdle arose from the size of QSs, because structural superposition algorithms have been classically applied to single chain proteins of a few hundred residues at most. In contrast, a QS can involve multiple chains totalling several thousand residues (**Fig. 2a**). As a solution, we employed the Kpax protein structure alignment algorithm[40], which we found to be fast and robust, even with very large structures. A second obstacle arose from the fact that QS information involves multiple chains, whose coordinates typically appear in PDB files in an arbitrary order. This prompted the development of a heuristic to map chain-chain correspondences (**Methods, Supplementary Fig. S2**). We used this approach to perform pairwise structural superpositions between all homo-oligomeric assemblies from the PDB, and also included assemblies predicted by PISA. We focused on potential matches by comparing homomers with identical numbers of subunits and domain folds, which resulted in >25 million superpositions. We then developed a strategy to integrate these data and infer the biological QS of protein structures. The overall algorithm is described in detail in the Methods as well as in **Supplementary Figures S3 and S4**. Henceforth, we refer to QSalign as the global method that computes structural superposition of protein complexes and uses that information to infer the biological relevance of assemblies in the PDB.

### Benchmarking QSalign

We employed a multi-chain version of the TM-score[41] to assess whether two QS states are similar or not. To find the optimal threshold and to examine how sensitive results were with respect to the threshold, we benchmarked QSalign using different TM-score cut-off values, from 0.4 to 0.9. For each value, we estimated the accuracy of QSalign annotations by comparing them to manual annotations from PiQSi. As expected, lower cut-off values gave more annotations but also higher error rates (**Fig. 2b**). Above a TM–score of 0.65, however, the error rate remained stable, but coverage decreased. We thus chose a cut-off equal to 0.65, at which 31,257 biological assemblies from the PDB were validated with an error rate of 2.1% (**Fig. 2b**).

Subsequently, we utilized each validated assembly to search for non-annotated entries with a similar sequence but a different QS. By transitivity, we annotated 11,119 such entries to be potentially incorrect, as illustrated in **Figure 1b**. Benchmarking these annotations against PiQSi showed an error rate estimated at 11.5% (**Fig. 2c**). While this error rate is higher, we note that these annotations are highly valuable because they are specific to erroneous structures.

We next compared the performance of QSalign to two state-of-the-art methods, PISA[20] and EPPIC[30]. Using PiQSi as a benchmark dataset, the prediction of dimers using PISA and EPPIC gave results comparable to published values with 13% and 18% error rates, respectively, while we observed an error rate of 4% with QSalign. Interestingly, among higher order oligomers, QSalign showed the same error rate of 4%, but values for PISA and EPPIC increased to 16% and 32%, respectively. This increase might be due to the fact that high-order oligomers involve several interfaces, and it becomes less likely to predict all of them correctly. We also benchmarked QSalign using a compendium of gold standard datasets, henceforth the “cGS” dataset (Methods), and we observed similar results (**Fig 2d**, **Table S1,** Table S4 **[[here we cite s4 before s2 … problematic]]**).

**QSalign accuracy is supported by residue-level conservation information**

In addition to benchmarking QSalign using a manually curated set of protein structures, we carried out a global quality assessment by comparing the evolutionary conservation of amino acids in biological versus non-biological interfaces. It is known that amino acids present at interaction interfaces tend to be more conserved than surface residues of the same protein[28-31]. Thus, we compared the log-ratio, *r*, of surface over interface evolutionary rate (Methods). If residues are equally conserved in both structural regions, *r* will be close to 0, whereas higher (or lower) conservation of interface residues will yield positive (or negative) *r* values, respectively. In non-biological QS states from PiQSi and QSalign, the distribution of *r* was centered on, or close to zero, indicating that interface residues behaved like surface residues, as expected. Among biological complexes, *r* was above zero on average ($Mean_{PiQSi}$=0.70, *p*=2.5e-15; $Mean_{QSalign}$=0.77, *p*=2.6e-6), indicating that residues at biological interfaces are, on average, twice as conserved as surface residues. The similarity of the distribution of *r* observed among curated structures from PiQSi and among structures annotated with QSalign further reflects the high quality of the annotations derived from the method (**Fig. 2d**). Interestingly, despite the fact that QSalign relies on interface geometry conservation to infer biological

relevance, 5.6% of complexes annotated as biologically relevant show *r* values below 0. Among these, the geometry of the interaction between subunits is conserved, but the residues mediating the interaction are variable.

**Annotating monomers with anti-QSalign**

QSalign was unable to annotate monomeric proteins, which by definition do not contain conserved interfaces. Nonetheless, by "inverting" the QSalign principle, we may consider the absence of interface in homologs to be predictive of monomeric proteins. We call this approach "anti-QSalign".

Our predictor was based on counts of non-redundant homologs having or not having a QS. With this strategy, it is expected that protein families with fast-evolving QS will lead to erroneous annotations, while families with conserved QS will allow robust predictions. As for QSalign, the high degree of redundancy in the PDB enabled annotating a majority of protein structures. Indeed, we found that ~80% of monomeric proteins (46,877) had at least one homolog in the 30% to 90% sequence identity range, and could therefore be classified as being either monomeric or oligomeric by anti-QSalign (Methods).

We benchmarked the predictions of anti-QSalign using both the cGS dataset (**Fig. 3a**) and PiQSi (**Fig. 3b**). In both benchmarks, anti-QSalign proved to be a reliable predictor with areas under the curve (AUC) of receiver-operator characteristic (ROC) plots equal to 0.94 (cGS) and 0.90 (PiQSi). These results compared favorably to PISA and EPPIC predictions, which showed maximal AUC values of 0.85.

**QSbio integrates predictions and provides QS confidence estimates on a PDB-wide scale**

Together, QSalign and anti-QSalign enabled annotating the QS state of over 80,000 structures (**Table S5**). Next, we asked whether QS conservation across methods could be utilized to create a meta-predictor with even higher accuracy and coverage. We employed the structural superposition heuristic of QSalign to compare QS states across PISA and the PDB, and we also mapped interfaces annotated by EPPIC onto the PDB assemblies (Methods, **Fig. 3c** and **Supplementary Fig. S5**). The comparison of PISA and EPPIC predictions with PDB assemblies enabled consensus predictions to be derived from their agreement or disagreement. We expected that a QS supported by both methods should be more likely to be biologically relevant than a QS supported by only one or neither method.

Among the structures annotated by QSalign or anti-QSalign, QSbio relied on all three methods. For other structures, QSbio combined PISA and EPPIC only. We assessed the performance of QSbio and saw that integrating predictions gave substantial improvements over each method taken individually. Integrating PISA and EPPIC (without QSalign/anti-QSalign) improved AUC values by 0.06, 0.08, and 0.08 for monomers, dimers, and oligomers, to reach values of 0.91, 0.91, and 0.83, respectively (**Fig. 3b**). A notable improvement is seen among dimers where the false positive rate is above 0.20 for either PISA or EPPIC, but decreases to 0.05 when both methods agree.

For structures also annotated by QSalign or anti-QSalign, the performance of QSbio was remarkable: among monomer and dimers, the AUC reached 0.95 and 0.97, respectively (**Fig. 3b**). Among higher-order oligomers, the predictive power of QSalign was already maximal (AUC=0.92) and only marginally improved in QSbio.

Overall, QSbio provides QS annotations with a quality approaching that of manual curation for 80% of monomeric and 70% of homo-oligomeric structures, and high-quality annotations for the rest of the structures. We estimated error rates for each class of prediction within the benchmark, and from them derived five levels of annotation confidence: very high, high, medium, low, and very low, corresponding to estimated error rates of 0-2, 2-5, 5-15, 15-50, and 50-100 percent. The numbers of structures assigned to each category were, respectively: 51,050, 18,217, 14,995, 14,335, and 11,499. These annotations will allow the scientific community to select high-confidence QSs to carry out structural analyses. Equally important, the low-confidence classes pinpoint structures in the PDB that may require correction.

**Plasticity of interfaces**

Interaction interfaces place constraints on the structure, chemistry, and evolution of proteins. Structurally, subunits may be required to maintain a precise orientation relative to each other[42]. Chemically, a specific composition at the interface is needed for binding[18,19,21-23], and we know that mutations generally decrease the interaction affinity[43]. As a result, on average, interfaces are hydrophobic, they are enriched in specific sets of amino acids, and their residues are more conserved than those at the surface[27-31].

There is, however, a large degree of variability around these average trends: some interfaces can be as hydrophilic as protein surfaces, while others can be less conserved than protein surfaces (Fig. 2e). It can be hypothesized that such variability reflects different structural requirements found across different protein families. In that case, structurally similar interfaces should deviate from the average in a similar fashion. Alternatively, interfaces may be plastic and be able to drift extensively, while maintaining the overall structure of a protein oligomer. To decide between these two alternatives, we examined the variability of interface properties within and across protein families.

As an initial case study, we analyzed six dimeric alcohol dehydrogenases. These show high structural similarity, with a mean TM-score among pairs equal to 0.81. In these enzymes, NAD (or NADP) is used as a cofactor and is in contact with both subunits, so we expected physicochemical and evolutionary properties of the dimeric interfaces to be conserved. We analyzed three interface properties (interface propensity, interface hydrophobicity, and the ratio of interface to surface conservation) among the six members of the family. We then compared how these properties varied relative to the entire dataset of 4,215 pairs (**Fig. 4a**). We observed that interface properties across the six enzymes spanned nearly the entire range seen in the dataset (**Fig. 4b**). For example, the human enzyme (PDB code 3COS) has a highly hydrophobic interface, while that of

*Rhizobium etli* (PDB code 4DVJ) is predominantly hydrophilic. Similarly, in terms of conservation, the interface of the enzyme from *Escherichia coli* (PDB code 4ILK) is 3.9 times more conserved than the rest of the surface, while its yeast homologue (PDB code 4OAQ) shows interface conservation similar to the rest of the surface. Thus, these dimers do not need their interface to meet strict family-specific physico-chemical requirements. Rather, their interfaces exhibit nearly as much variability as is sampled among dimers in general.

Is such interface plasticity specific to this alcohol dehydrogenase family, or is it general to all protein families? To answer this question, we employed a dataset of 4,215 pairs of homologous interfaces based on the results of QSalign. We then assessed the within-family variability of each property relative to the variability across families. If variability within families is small compared to the variability seen across families, the correlation coefficient will be close to one. However, if the variability within families is as large as the variability across families (as we observed for alcohol dehydrogenases) the correlation coefficient will be close to zero (**Fig. 4c**). When comparing interface properties among homologs sharing 80% to 90% sequence identity, we observe high correlations, with $R^2$ values equal to 0.82, 0.48, and 0.57 for interface conservation, hydrophobicity, and propensity, respectively (**Supplementary Fig. S6**). However, the fact that close homologs share interface properties is expected because they have not had sufficient evolutionary time to diverge. As sequence identity decreases among the homologs being compared, so does the similarity between their interfaces, to the point where the interfaces within two homologous assemblies are no more alike than random interfaces. Indeed, for the three interface properties we considered, the correlation among distant homologs is close to zero, even though they share the same interface geometry and overall structure (**Fig. 4**, **Supplementary Fig. S6**). Thus, this result generalizes previous observations made on heteromeric and transient interactions[44,45] and shows that homo-oligomeric interfaces, which are often large and stable, can nevertheless be highly plastic.

## DISCUSSION

We introduced two complementary approaches, QSalign and anti-QSalign, which allow QS prediction based on conservation across evolution. While we focused on annotating monomers and homo-oligomers, QS conservation could, in principle, also be applied to hetero-oligomers. Such an extension will require additional developments because differences in numbers of subunits may be tolerated among hetero-oligomers, and because subunits of different size make global structure similarity harder to interpret.

The main limitation of QSalign and anti-QSalign is their requirement for homologous structures, but the high species coverage of the PDB allowed annotating ~80% and ~70% of monomeric and homo-oligomeric structures, respectively. We anticipate that these fractions will increase with time, as more structures are solved. Increasing the number of high quality QS annotations will be important for detailed structural analyses of disease-causing mutations[46,47] and the determinants of protein evolution[48,49].

In addition to using QS conservation across evolution, we employed QS comparison to integrate predictions of different methods with QSbio, which improved considerably the prediction accuracy for monomers and dimers. A noteworthy feature of QSbio are the confidence estimates that will facilitate the interpretation of QS information, both for individual structures and for PDB-wide datasets serving in bioinformatics analyses.


## ACKNOWLEDGMENTS

We thank H. Greenblatt for valued help with operating the computer cluster, and O. Dym and S. Rogotner the photo of a protein crystal used in Fig. 1. We thank J. Sussman for feedback on the work and D. Fass for comments on the manuscript. This work was supported by a VATAT fellowship to S. Dey, by the Israel Science Foundation and the I-CORE Program of the Planning and Budgeting Committee (grant nos. 1775/12 and 2179/14), by the Marie Curie CIG Program (project no. 711715), by the HFSP Career Development Award to E. D. Levy (award no. CDA00077/2015), and by a research grant from AM. Boucher. E.D. Levy is incumbent of the Recanati Career Development Chair of Cancer Research.


## AUTHOR CONTRIBUTIONS

SD and EDL designed and performed the experiments. DWR adapted the KPAX algorithm to enable the calculations. SD and EDL wrote the manuscript with input from DWR. All authors corrected and approved the final manuscript.

## COMPETING FINANCIAL INTERESTS

None

## FIGURES LEGENDS

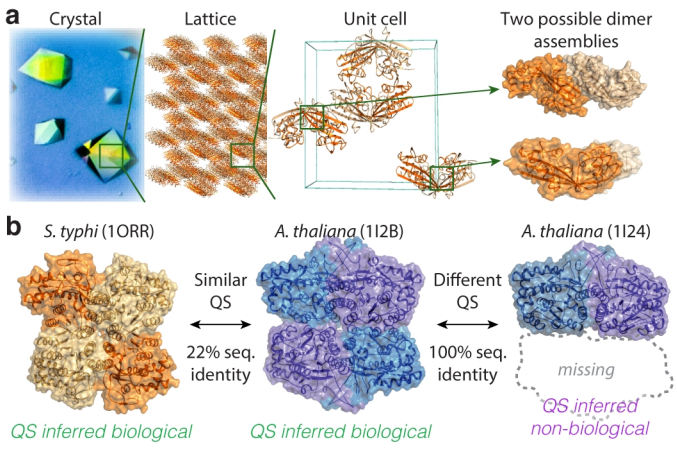

**Figure 1 |** Quaternary structure conservation across species points to biologically relevant crystal contacts. (**a**) Protein X-ray diffraction requires formation of a crystal. At the molecular level, a crystal is formed by a lattice within which the repeated unit is the unit cell. Here, the unit cell contains eight copies of the protein (PDB code 1EX2) in contact with one another. Identifying biologically relevant contacts among these is challenging. For example, the authors of the structure shown[50] assigned the top dimer as biologically significant, while the method PISA predicts the dimeric form underneath to be so. Searching for homologous structures reveals the latter to be conserved across species (*e.g.*, PDB code 4LU1) (**b**) Tyvelose epimerase is a tetrameric enzyme in *Salmonella typhi* (PDB code 1ORR). A similar tetramer is found in *Arabidopsis thaliana* (PDB code 1I2B, RMSD=3.55Å) despite their sequence sharing only 22% identity. Such conservation suggests that both tetramers are biologically relevant. This information enables subsequent correction of entries showing identical sequence but different QS (*e.g.*, PDB code 1I24).

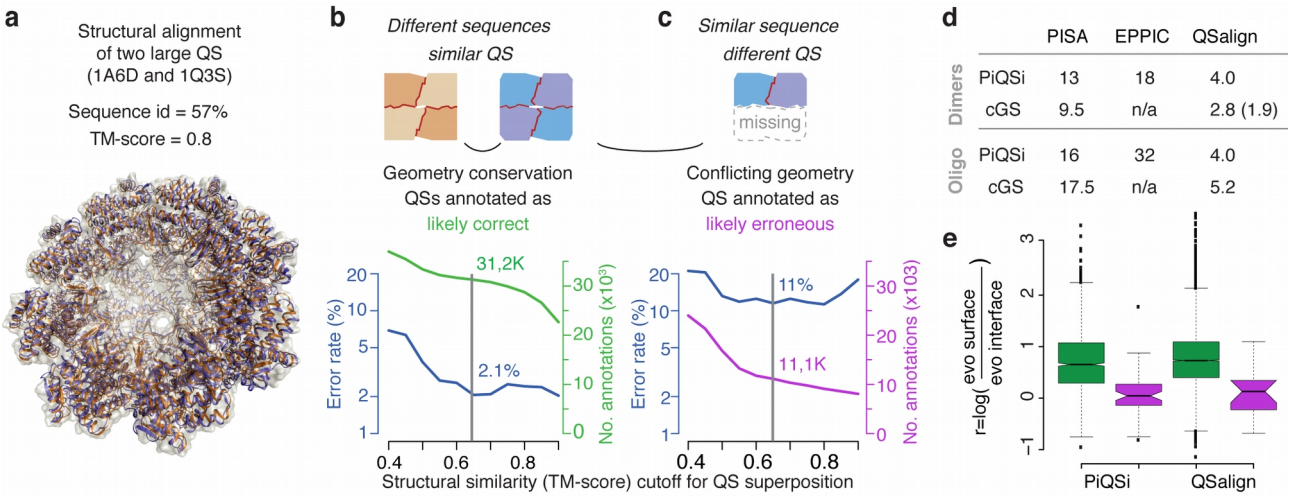


**Figure 2** | Quaternary structure superposition and benchmark of predictions. **(a)** Superposition of two large QSs by QSalign (only half of the structure is shown to facilitate visualization). **(b)** We compared QSalign predictions against a manually curated dataset from PiQSi[36]. QSalign infers two QSs as correct when their structural similarity exceeds a TM-score cut-off. We scanned different cut-off values and for each, recorded the error rate (blue line) and the total number of QSs annotated (green line). **(c)** QSalign then searches for proteins with conflicting QSs and infers those as erroneous. (**d**) Benchmark of QSalign, PISA and EPPIC using the same datasets. The number in parenthesis corresponds to the error rate after discarding a structure for which recent work confirms QSalign prediction. Detailed prediction information for the three methods is given in Supplementary Table S1. **(e)** Sequence conservation of interface residues of QSs predicted to be biological (green) or non-biological (red) in QSalign, and PiQSi. Similar distributions of relative interface conservation are observed ($Mean_{PiQSi}$=0.70, $p$=2.5e-15; $Mean_{QSalign}$=0.77, $p$=2.6e-6; two sided Wilcoxon test).

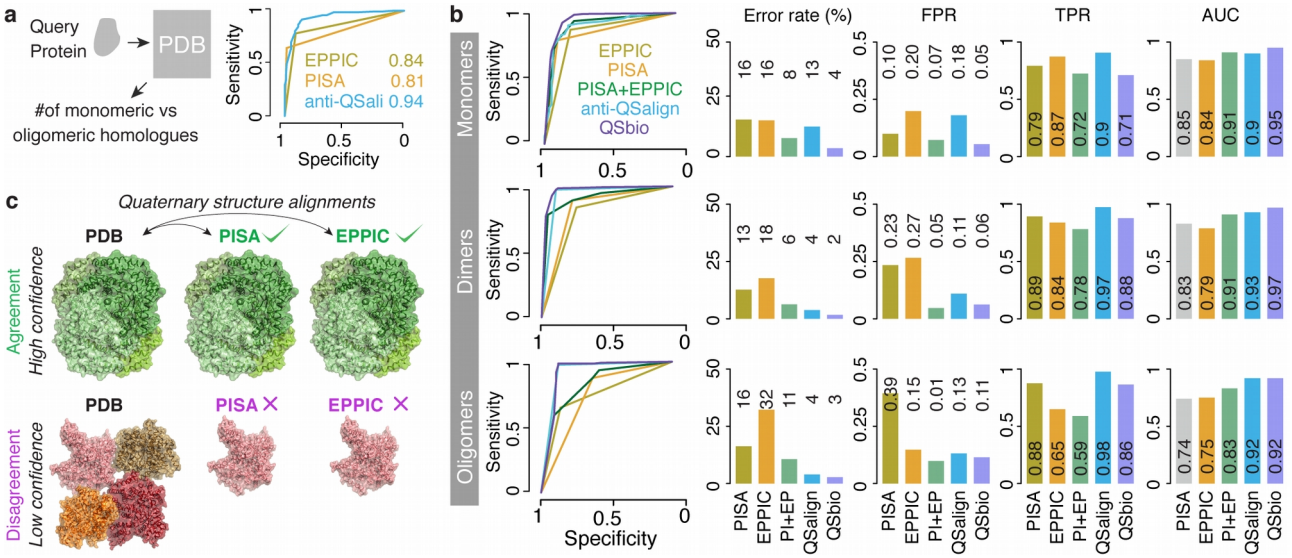


**Figure 3 |** Principle of anti-QSalign and benchmark of QSbio. (**a**) anti-QSalign uses the oligomeric state of homologues to infer whether a protein is monomeric. ROC analysis based the cGS dataset (144 monomers and 187 oligomers). The AUC for anti-QSalign, PISA and EPPIC are shown. Detailed prediction information are given in Supplementary Table S2. (**b**) Benchmark of individual methods and of their integration into QSbio. Among monomers, dimers, and larger oligomers, combining PISA and EPPIC led to 2, 2, and 1.3-fold reduction in error rate. The addition of QSalign reduced the error rate further, by 2, 3, and 3.5-fold, respectively. These improvements are reflected in the AUC values. Detailed prediction information are given in Supplementary Table S3. (**c**). Example of QS states annotated as correct or incorrect in QSbio. The predictions across methods agree for the top QS (green, PDB code 2I56) and disagree for the QS underneath (PDB code 1JQN).

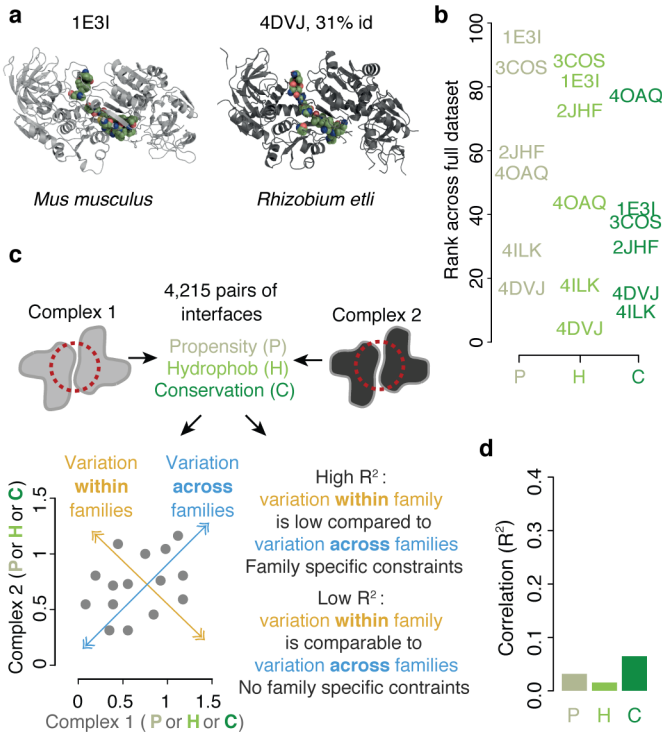


**Figure 4 |** Protein interfaces are plastic. (**a**) We compared the interfaces of structurally similar protein complexes and exemplify interface plasticity in dimeric alcohol dehydrogenases. Two homologous structures sharing 31% sequence identity are shown. The interface, shown with spheres, is conserved structurally but not chemically. (**b**) Rank of interface properties for six alcohol dehydrogenases, where 0 and 100 correspond respectively to the lowest and highest values seen in any complex from the dataset. The six members span nearly the entire range, indicating that their interface properties vary greatly, nearly as much as seen across the entire dataset of 4,215 pairs. (**c**) We evaluated interface plasticity across the entire dataset. The correlation coefficient ($R^2$) of a given property for pairs of structures reflects whether that property is constrained in a family-specific manner. If pairs of homologous interfaces exhibit similar values for the properties examined, a high $R^2$ value is expected. In contrast, if there are no family-specific constraints, homologous interfaces will be as different from each other as two random interfaces can be, and $R^2$ will be low. (**d**). The correlation coefficient among pairs of structurally similar interfaces is low (<0.1) when the proteins compared share less than 30% sequence identity. This result highlights that structurally similar interfaces can be as different from each other chemically and evolutionarily as interfaces from entirely different families, provided they have had sufficient evolutionary time to diverge.

**REFERENCES (Main text only)**


1 Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annual review of biophysics and biomolecular structure* **29**, 105-153 (2000).

2 Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D complex: a structural classification of protein complexes. *PLoS Comput Biol* **2**, e155 (2006).

3 Lukatsky, D. B., Shakhnovich, B. E., Mintseris, J. & Shakhnovich, E. I. Structural similarity enhances interaction propensity of proteins. *J Mol Biol* **365**, 1596-1606 (2007).

4 Andre, I., Strauss, C. E., Kaplan, D. B., Bradley, P. & Baker, D. Emergence of symmetry in homooligomeric biological assemblies. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 16148-16152 (2008).

5 Marsh, J. A. & Teichmann, S. A. Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem* **84**, 551-575 (2015).

6 Ahnert, S. E., Marsh, J. A., Hernandez, H., Robinson, C. V. & Teichmann, S. A. Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).

7 Nooren, I. M. & Thornton, J. M. Diversity of protein-protein interactions. *Embo J* **22**, 3486-3492 (2003).

8 Kuhner, S. *et al.* Proteome organization in a genome-reduced bacterium. *Science* **326**, 1235-1240 (2009).

9 Perica, T. *et al.* The emergence of protein complexes: quaternary structure, dynamics and allostery. Colworth Medal Lecture. *Biochemical Society transactions* **40**, 475-491 (2012).

10 Renatus, M., Stennicke, H. R., Scott, F. L., Liddington, R. C. & Salvesen, G. S. Dimer formation drives the activation of the cell death protease caspase 9. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 14250-14255 (2001).

11 Tang, P., Hung, M. C. & Klostergaard, J. Human pro-tumor necrosis factor is a homotrimer. *Biochemistry* **35**, 8216-8225 (1996).

12 Pereira-Leal, J. B., Levy, E. D., Kamp, C. & Teichmann, S. A. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol* **8**, R51 (2007).

13 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235-242 (2000).

14 Velankar, S. *et al.* PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic acids research* **44**, D385-395 (2016).

15 Henrick, K. & Thornton, J. M. PQS: a protein quaternary structure file server. *Trends Biochem Sci* **23**, 358-361 (1998).

16 Janin, J. Specific versus non-specific contacts in protein crystals. *Nature structural biology* **4**, 973-974 (1997).

17 Carugo, O. & Argos, P. Protein-protein crystal-packing contacts. *Protein Sci* **6**, 2261-2263 (1997).

18 Ponstingl, H., Henrick, K. & Thornton, J. M. Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* **41**, 47-57 (2000).

19 Zhu, H., Domingues, F. S., Sommer, I. & Lengauer, T. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* **7**, 27 (2006).

20 Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774-797 (2007).

21 Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J. & Poupon, A. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics (Oxford, England)* **24**, 652-658 (2008).

22 Tsuchiya, Y., Nakamura, H. & Kinoshita, K. Discrimination between biological interfaces and crystal-packing contacts. *Adv Appl Bioinform Chem* **1**, 99-113 (2008).

23 Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* **336**, 943-955 (2004).

24 Pal, A., Chakrabarti, P., Bahadur, R., Rodier, F. & Janin, J. Peptide segments in protein-protein interfaces. *J Biosci* **32**, 101-111 (2007).

25 Tina, K. G., Bhadra, R. & Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic acids*

*research* **35**, W473-476 (2007).
26 Liu, Q., Li, Z. & Li, J. Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC Bioinformatics* **15 Suppl 16**, S3 (2014).
27 Elcock, A. H. & McCammon, J. A. Identification of protein oligomerization states by analysis of interface conservation. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 2990-2994 (2001).
28 Guharoy, M. & Chakrabarti, P. Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15447-15452 (2005).
29 Scharer, M. A., Grutter, M. G. & Capitani, G. CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. *Proteins* **78**, 2707-2713 (2010).
30 Baskaran, K., Duarte, J. M., Biyani, N., Bliven, S. & Capitani, G. A PDB-wide, evolution-based assessment of protein-protein interfaces. *BMC Struct Biol* **14**, 22 (2014).
31 Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research* **44**, W344-350 (2016).
32 Xu, Q. *et al.* Statistical analysis of interface similarity in crystals of homologous proteins. *J Mol Biol* **381**, 487-507 (2008).
33 Xu, Q. & Dunbrack, R. L., Jr. The protein common interface database (ProtCID)--a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic acids research* **39**, D761-770 (2011).
34 Shoemaker, B. A. *et al.* IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic acids research* **40**, D834-840 (2012).
35 Faure, G., Andreani, J. & Guerois, R. InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic acids research* **40**, D847-856 (2012).
36 Levy, E. D. PiQSi: Protein Quaternary Structure Investigation. *Structure* **15**, 4 (2007).
37 Sippl, M. J. & Wiederstein, M. Detection of spatial correlations in protein structures and molecular complexes. *Structure* **20**, 718-728 (2012).
38 Koike, R. & Ota, M. SCPC: a method to structurally compare protein complexes. *Bioinformatics (Oxford, England)* **28**, 324-330 (2012).
39 Mukherjee, S. & Zhang, Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic acids research* **37** (2009).
40 Ritchie, D. W., Ghoorah, A. W., Mavridis, L. & Venkatraman, V. Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics (Oxford, England)* **28**, 3274-3281 (2012).
41 Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302-2309 (2005).
42 Perica, T., Chothia, C. & Teichmann, S. A. Evolution of oligomeric state through geometric coupling of protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 8127-8132 (2012).
43 Moal, I. H. & Fernandez-Recio, J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics (Oxford, England)* **28**, 2600-2607 (2012).
44 Andreani, J., Faure, G. & Guerois, R. Versatility and invariance in the evolution of homologous heteromeric interfaces. *PLoS Comput Biol* **8**, e1002677 (2012).
45 Sudha, G., Singh, P., Swapna, L. S. & Srinivasan, N. Weak conservation of structural features in the interfaces of homologous transient protein-protein complexes. *Protein Sci* **24**, 1856-1873 (2015).
46 Shi, Z. & Moult, J. Structural and functional impact of cancer-related missense somatic mutations. *J Mol Biol* **413**, 495-512 (2011).

47 David, A. & Sternberg, M. J. The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease. *J Mol Biol* **427**, 2886-2898 (2015).
48 Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548**, 244-247 (2017).
49 Bloom, J. D., Drummond, D. A., Arnold, F. H. & Wilke, C. O. Structural determinants of the rate of protein evolution in yeast. *Molecular biology and evolution* **23**, 1751-1761 (2006).
50 Minasov, G. *et al.* Functional implications from crystal structures of the conserved Bacillus subtilis protein Maf with and without dUTP. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 6328-6333 (2000).

## METHODS

### Datasets

The dataset of protein structures on which all analyses were performed is based on 3DComplex as of March 2015[2] and consists of 110,498 "biological assemblies" defined by the Protein Data Bank. Note that the total number of structures annotated in QSbio is slightly lower (110,096) because EPPIC lacks some annotations. The dataset is available on the 3DComplex (Version 5) website:

http://shmoo.weizmann.ac.il/elevy/3dcomplexV5/Home.cgi

For each structure, we also included the top prediction from PISA[20] as of April 2015. EPPIC assignments of interfaces were kindly provided by G. Capitani in October 2015.

### Structural superposition of QSs

Structural superposition was carried out using Kpax (version 3)[40]; Briefly, Kpax first aligns two structures by placing all possible pairs of 5-mer fragments at the coordinate origin by exploiting the tetrahedral geometry of their central Cα atoms, and by using dynamic programming (DP) to find the optimal combination of pairs of fragments, and hence pairs of central Cα atoms. The aligned Cα atoms are then superposed by least-squares fitting, and the resulting superposition is refined using further rounds of DP on the atomic Gaussian overlap of the superposed Cα pairs. Each Gaussian represents the volume of one Cα atom. Thus, if two Cα atoms overlap perfectly they contribute one volume unit to the overall superposition score, and if two aligned protein chains superpose perfectly then the total Gaussian overlap score directly corresponds to the volume of their overlaid Cα atoms. Kpax can align structures with multiple chains, but only in the given chain order. To find the maximum aligned overlap volume between two multi-chain oligomers, it is necessary to determine the best chain order of one structure with respect to the other. We therefore employed a two-step heuristic procedure to avoid the computational cost of running Kpax repeatedly on different chain order permutations. A first alignment and superposition is carried out. The coordinates of the superposed QSs are then analyzed to map corresponding chains across the two structures. Scores between chain pairs are calculated and correspond to the number of Cα atoms of one chain closer than 2 Å from any Cα from another chain. The highest-ranking chain pairs provide correspondences that are used to rewrite the PDB file in matching chain order. The re-ordered coordinates are then used in a second round of structural superposition, which yields the final TM-score. Pseudo-code describing this procedure is provided in a

Supplemental Note. An executable version of the structural superposition program is available at https://github.com/elevywis/QSalign.

**PDB-wide superposition of QSs**

The annotation process of QSalign first required structural superposition of homo-oligomers. To save computation time, we carried out structural alignments between potential matches only, that is, between pairs of structures sharing structural homology as reflected in their SCOP or PFAM domain architecture, or sharing sequence homology (>30% sequence identity). Ultimately, we measured the structural similarity of 25,965,020 QS pairs and recorded the TM-score for each, which was stored in a MySQL database table. We give the pseudocode for the comparison of one query structure with multiple target structures as a function "QSalign" in a Supplementary Note.

**Annotation procedure**

The execution of structural alignments was followed by an inference process described in the pseudocode "QSinfer" given in Supplementary Note. Briefly, a query QS was annotated as correct if another structurally similar QS (TM score > threshold) was found for a homolog with less than 80% sequence identity. Pairs of QSs sharing more than 80% sequence identity were not considered in order to reduce the risk of the same crystal packing being formed due to protein surface similarity. The annotation process was carried out for each symmetry group separately, starting with those containing larger numbers of subunits. This condition ensured that a lower-order oligomer (*e.g.*, a dimer) would be annotated as correct only if no evidence for high-order oligomerization (*e.g.*, a tetramer) was found. Once all QSs from a symmetry group were processed, those annotated as correct where used to search for possible errors. The overall strategy is described in the pseudocode "QSpropagate", given in Supplementary Note. The correction step consists of identifying proteins with an identical sequence and a structure different from that of a QS annotated as being correct (**Supplementary Figures S3 and S4**). Then, the annotation depends on the number of subunits of the QS to be annotated (QS2) and the QS that is supposedly correct (QS1). We distinguish three cases: (i) If QS2 has more subunits than QS1, we analyze the consistency of QS2 given its number of subunits and its symmetry. We call QS2 consistent when the number of subunits is expected from the point group symmetry (*e.g.*, dimer for C2, tetramer for D2, *etc*). Consistent structures are flagged as "ambiguous" and not annotated as errors, while inconsistent structures are annotated as errors. (ii) If QS2 has the same number of subunits as QS1, we infer that QS2 is incorrect. (iii) Lastly, when QS2 has fewer subunits than QS1, we infer that interfaces are missing from QS2, which is also annotated as incorrect.

To optimize the TM-score threshold at which two QSs were considered equivalent, we carried out several full cycles of annotation (*e.g.*, inference + propagation steps) using different TM-scores cut-off (**Fig. 2**) and benchmarked the results after each cycle based on PiQSi[36].

**Benchmarking the annotations of QSalign for TM-score optimization**

After each cycle of the annotation procedure, we counted the number of structures in the following categories: TP=annotated as correct by QSalign and PiQSi; FP=annotated as correct by QSalign and as

incorrect by PiQSi; FN=annotated as incorrect by QSalign and as correct by PiQSi; and TN=annotated as incorrect by both QSalign and PiQSi. We calculated the error rate of "correct" annotations by the false discovery rate FDR = FP / (TP+FP) and the error rate of "incorrect" annotations by the false omission rate FOR = FN / (FN+TN). The dataset of structures on which these rates were calculated was filtered at a level of 90% sequence identity. In addition, we used only high confidence annotations from PiQSi, which did not contain the tag "probable". The resulting number of structures from PiQSi used for the benchmark was 1434.

**Annotating monomers with anti-QSalign**

We annotated monomers based on the QS state - considered either as monomeric or oligomeric - of known homologs. The underlying assumption of this approach is that QS is generally conserved during evolution[51], so that structures homologous to a true monomer should be more likely to be monomeric than oligomeric. Thus, for a particular protein sequence, the enrichment of monomeric over oligomeric homologs was used to derive a probability score to be monomeric. These probabilities were estimated based on PDB structures not found in any benchmark, as follows. We defined six bins of "numbers of homologs": 0, 1, 2-3, 4-7, 7-14, and >14, and each protein was assigned to two such bins, one for the number of monomeric homologs and the other for the number of oligomeric homologs. We then recorded frequencies of monomers in each bin combination, *e.g.*, 4.2% of monomers have 2 or 3 monomeric homologs and a single oligomeric homolog. Then, we recorded the same frequencies for oligomers, *e.g.*, 1.35% of oligomers have 2 or 3 monomeric homologs and a single oligomeric homolog. The frequencies were subsequently converted into probabilities, *e.g.*, we estimated that a protein has a probability of 4.2/(4.2+1.35)=0.76 to be monomeric if it has 2 or 3 monomeric homologs and a single oligomeric homolog. This process is illustrated in **Supplementary Fig. S7**. We considered proteins sharing a minimum of 30% and a maximum of 90% sequence identity to be homologs. We also imposed a minimum of 60% overlap of the sequence alignment relative to the longest of the two proteins being aligned.

**Comparative benchmarks against PiQSi**

We compared the set of annotations obtained with QSalign and anti-QSalign to predictions derived from PISA and EPPIC. Importantly, PiQSi annotations refer to specific PDB structures. Therefore, we could directly compare the annotations of PiQSi to the annotations of PISA and EPPIC for the corresponding structure. Positives and negatives were counted when the methods supported (positive, P) or not (negative, N) a particular PDB structure. These predictions, together with PiQSi annotations, allowed a contingency table to be created, and performance statistics to be derived. ROC curves were plotted using R[52] and the pROC package[53]. The numbers of structures (broken down by oligomeric state) used in this comparative benchmark are given in **Table S4**.

The specific conditions to count positives and negatives depend on the oligomeric state of the PDB structure considered. First, for monomeric proteins, the conditions for counting positives and negatives were:

- anti-QSalign: the "monomer probability score" associated with the PDB code was above (positive) or below (negative) 0.4.
- PISA: no assembly was predicted (positive), or an assembly was predicted (negative)
- EPPIC: no biological interface was predicted (positive), or at least one was (negative).

Second, for oligomers with two or more subunits, the conditions for counting positives and negatives were:

- QSalign: the structure is annotated as correct (positive) or incorrect (negative).
- PISA: the structure matches the PDB structure (positive, TM-score > 0.9) or not (negative, TM-score < 0.9)
- EPPIC: when biological interfaces are mapped onto the PDB structure, they either maintain all subunits in contact (positive) or they do not (negative).

**Comparative benchmarks against the compendium gold-standard dataset (cGS).**

We assembled a second gold standard dataset based on previously published datasets. We used datasets published by Bahadur et al.[54], Ponstigl et al.[18] and Duarte et al.[55] (for monomers only), which gave a total of 338 structures including 144 monomers, 137 dimers, and 57 oligomers (**Table S4**). We call the so-obtained dataset the consolidated gold standard (cGS).

An important difference to note between cGS and PiQSi is that we only used "PDB code identifiers" for the cGS and not actual structures. Therefore, we used the "number of subunits" to assess predictions. For monomers, the conditions for counting positives and negatives were identical to those used in the PiQSi benchmark. For oligomers, however, they differed as follow:

- QSalign: the number of subunits predicted by QSalign is equal to the number of subunits in the benchmark dataset (positive) or is different (negative)
- PISA: the number of subunits in the PISA assembly is equal to the number of subunits in the benchmark dataset (positive) or is different (negative)
- EPPIC: we could not infer the number of subunits solely from interface information, and so we were unable to include EPPIC in this benchmark.

**Integrating QS information with QSbio**

QSbio compares QSs from the PDB with predictions of EPPIC, PISA, and QSalign/anti-QSalign when available. The comparison with PISA was carried out by structural superposition, as described above, and pairs of QSs with a TM-score above 0.9 were considered identical. To compare PDB and EPPIC QSs, we mapped interface groups from EPPIC onto PDB structures. For the mapping, we decomposed each PDB assembly into pairs of contacting chains, and each pair was superposed onto the interfaces from EPPIC to identify matching pairs (**Supplementary Figure S5)**. We then derived a weighted score for a PDB structure based on its support by the different methods. For monomers, the score was calculated as: $S_{MONO}$ = 0.4*antiQSalign + 0.4*PISA + 0.2*EPPIC. For oligomers, the score was calculated as: $S_{OLIGO}$ = 0.7*QSalign + 0.2*PISA + 0.1*EPPIC. When the annotation of (anti)-QSalign was not available, the score was re-normalized to 1.0, but the same weights for PISA and EPPIC were used. We estimated the expected error rates (err) in QSbio by: err = FPR * fN / (FPR * fN + TPR * fP), where TPR and FPR are the true and false positive rates measured in each score class, and fP and fN are the estimated (unknown) fractions of correct and incorrect structures in the PDB. We aimed for conservative error rates, so we used relatively high values for fN and fP, namely fN=0.2 and fP=0.8. Importantly, the error rates estimated for each class also depend on the benchmark dataset and on the assumed pattern of errors in the PDB. In that respect, their absolute value

should be interpreted with care. We do expect, however, that the accuracy rank of the different classes will be robust. To reflect this limitation, as well as to simplify the use of these predictions by end-users, we created five confidence categories: "very high", "high", "medium", "low" and "very low", corresponding respectively to estimated error rates of 0-2, 2-5, 5-15, 15-50, and 50-100 percent.

**Interface properties**

We selected pairs of homologous dimers matching the following criteria: the TM-score was above 0.65, each structure's resolution was below 2.5Å, each structure was a representative of a non-redundant set at 90% sequence identity and each structure was annotated as biological by QSalign. All properties were calculated on interface core residues, as defined in Levy[56]. To calculate sequence conservation, each protein was used to search for homologs in Uniref90[57] and homologs with >40% sequence identity and 80-100% coverage were retained for multiple alignments using MUSCLE[58]. Alignments with less than 5 sequences were discarded. We subsequently used rate4site with default parameters[59] to obtain amino acid specific evolutionary rates. Rates were rescaled to positive values by subtracting the minimum score in each alignment. Interface-core and surface evolutionary rates were obtained by averaging the scores of residues composing each region respectively. The relative conservation of interfaces was obtained by the log-ratio $r$ log (evolutionary rate at surface / evolutionary rate at interface), so $r > 0$ indicates the interface is more conserved (*i.e.*, lower rate of evolution) compared to the surface. Differences in the distribution of $r$ were tested using a two-sided Wilcoxon test. Interface hydrophobicity was calculated by the ratio of hydrophobic residues present at the interface. Interface residue propensity was calculated by averaging individual amino acid propensities taken from Dey *et al.*[60] and normalizing them by the size of the interface.

**Data Availability**

All the annotations and confidence categories derived from this work are available for download at www.QSbio.org.

**Code Availability**

The code implementing the heuristic procedure to be used with the Kpax algorithm for QS alignment is available at https://github.com/elevywis/QSalign.

**REFERENCES (Methods)**


51 Levy, E. D., Boeri Erba, E., Robinson, C. V. & Teichmann, S. A. Assembly reflects evolution of protein complexes. *Nature* **453**, 1262-1265 (2008).

52 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. (2016).

53 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

54 Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. Dissecting subunit interfaces in

homodimeric proteins. *Proteins* **53**, 708-719 (2003).
55 Duarte, J. M., Srebniak, A., Scharer, M. A. & Capitani, G. Protein interface classification by evolutionary analysis. *BMC Bioinformatics* **13**, 334 (2012).
56 Levy, E. D. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol* **403**, 660-670 (2010).
57 Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)* **31**, 926-932 (2015).
58 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797 (2004).
59 Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S71-77 (2002).
60 Dey, S., Pal, A., Chakrabarti, P. & Janin, J. The subunit interfaces of weakly associated homodimeric proteins. *J Mol Biol* **398**, 146-160 (2010).