Structure, Volume *29*

Supplemental Information

PDB-wide identification of physiological hetero-oligomeric assemblies based on conserved quaternary structure geometry

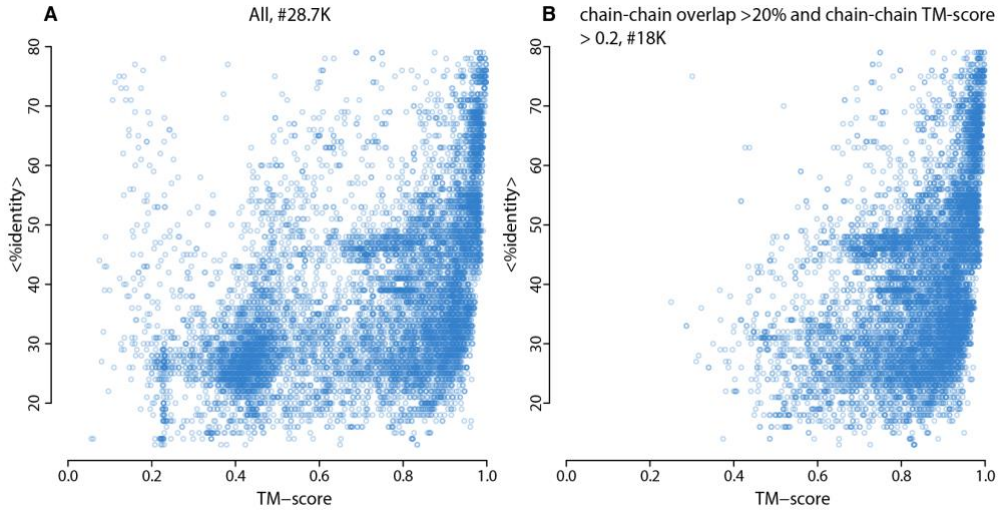
Sucharita Dey and Emmanuel D. Levy

Supplementary Table S2 | All QSalign^{HET} annotation categories and their numbers with an example. The last two columns show the numbers of each annotation type from PiQSi^{HET}, related to Figure 2.

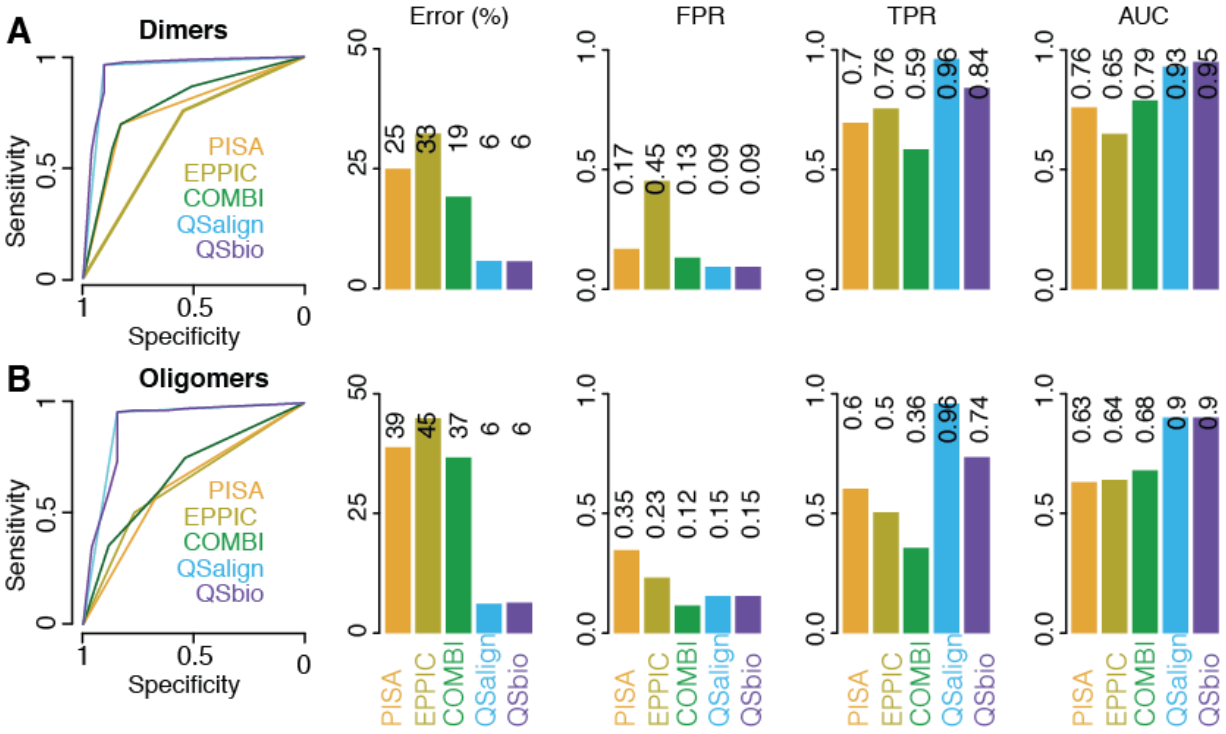
ID	Annotation	#	#NR	Annotation example (sentence)	Detailed explanation **Note**: a complex "including" another means that all subunits of the smaller complex match a respective subunit (sequence identity > 95%) in the larger one.	#NO	#YES
1	*Physiologically relevant*	7810	1747	*Query: 4az0_1; QS geometry similar to that of 1gxs_1 (TM=0.85; seq identity: avg=30%, max=35%)*	A query complex is assigned this annotation when a target complex with a similar QS is found, yielding evidence of QS conservation.	161	4
2	Crystal interface	260	101	*Query: 1bi8_1; QS geometry similar to that of 1g3n_4 (TM=0.93; seq-identity: avg=73%) - Note that 1bi8_1 PISA QS was used to detect structural similarity.*	A query complex is assigned this annotation when the PISA-predicted QS for that same complex is different but likely correct, as inferred from QS conservation. These crystal contacts are the ones we are most confident about because an alternative and conserved interface exists in the lattice of that structure.	2	5
3	Sub-stoichiometry	523	108	*Query: 1mmf; This QS has the same composition as 1iwp_1 (which is likely correct), but 1mmf subunits are in lower stoichiometry.*	A query complex is assigned this annotation when another complex with identical composition but higher stoichiometry has been found, and the higher-stoichiometry was validated by QS conservation.	1	43
4	Sub-composition	84	22	*Query: 1aui; This QS has a subset of subunits present in 1tco_1 (which is likely correct).*	A complex is flagged with this annotation when it is included in another complex. This does not invalidate the QS *per se*, but rather serves to inform on alternative compositions.	2	0
5	Excessive stoichiometry	989	358	*Query: 2r9p; This QS has the same composition as 3p95_1 (which is likely correct) but subunits are in higher stoichiometry with no support from evolutionary conservation.*	A complex is flagged with this annotation when it includes another complex and shares the same subunit composition so the difference lies in the subunit stoichiometries. While evolutionary conservation of QS is detected for the lower-stoichiometry form only, it does not necessarily invalidate the higher-stoichiometry form.	12	6
6	Crystal interface or large conformational change	277	102	*Query: 3dhh; This QS shows different stoichiometry and/or composition to 5tds_1 (which is likely correct) along with significant structural changes (TM-score=0.4991). Chain-chain matching information: B:B:301:92:99.*	A query complex is assigned this annotation when it includes another complex and shows structural differences, hinting at a possible crystal interface or a conformational change. The difference in composition/stoichiometry may be associated with the structural differences detected. This is why we consider this case separately from the next one (#7), where there is no difference in composition/stoichiometry.	3	3
7	Crystal interface or large conformational change	96	24	*Query: 4fxk; This QS shows the same stoichiometry and composition to 5jtw_2 but the structure is different, (TM-score=0.4777). This might reflect that an incorrect QS or may originate in large conformational changes.*	A query complex is assigned this annotation when the size and the composition is the same between the two complexes, but they differ structurally.		
8	Ambiguous	133	43	*Query: 4c3o_1;* PDB and PISA assemblies are different, but both show QS conservation with 4kl8_1 and 5a4f_1 respectively, based on the target 2frv_4.	These are sub-classes of ID#2. PDB and PISA QS are different and both have evidence of structural conservation	3	0
9	Ambiguous	130	53		The number of structures supporting the annotation is less than 5% of the size of the family		

Supplementary Table S4 | Prediction statistics for individual methods, related to Figure 4.

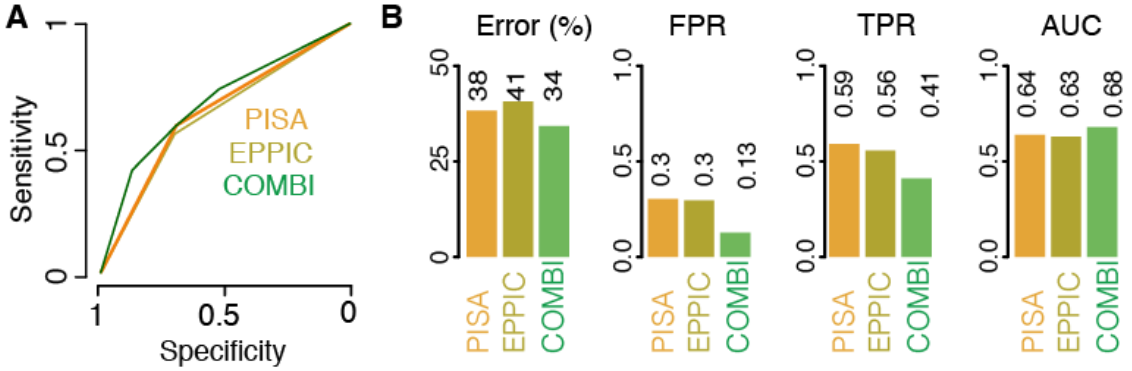
Method	TP	FN (total #positives = 203)	TN	FP (total #negatives = 79)
PISA	130	73	61	18
EPPIC	123	80	49	30
QSalign^{HET}	195	8	70	9
QSbio	158	4	41	9



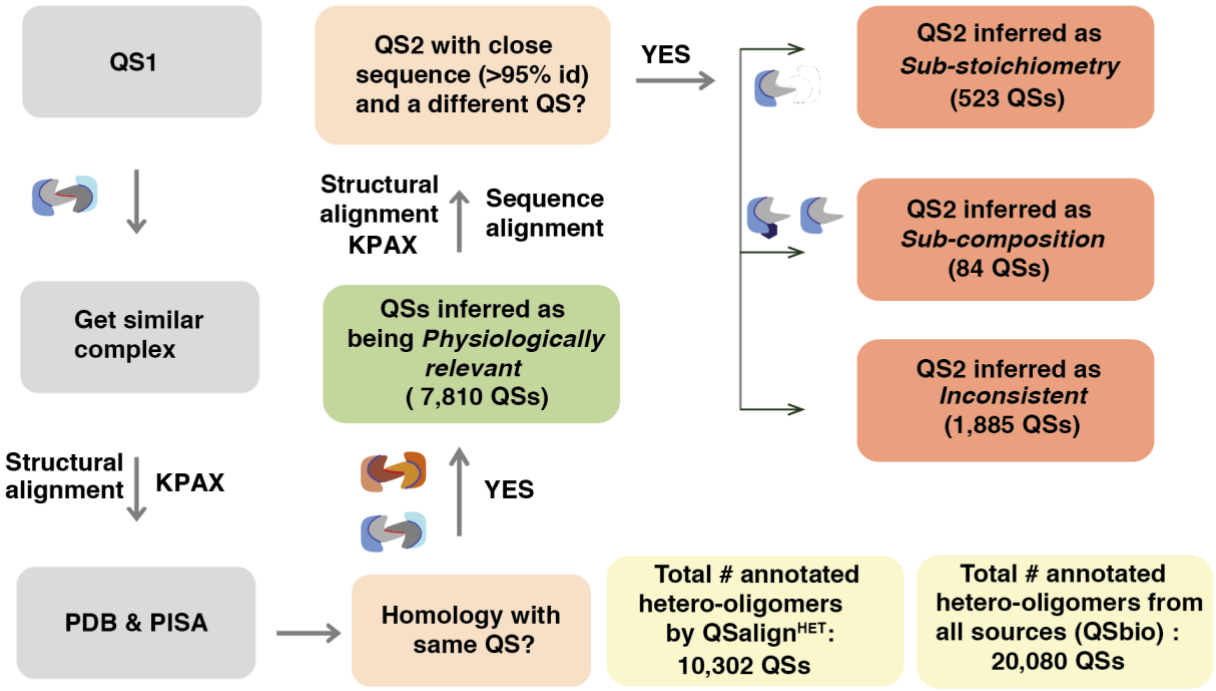
Supplementary Figure S1. TM-scores of pairs of complexes compared by QSalign^{HET} **A**. When comparing the TM-score as a function of sequence identity, an explosion of data is expected at values below 0.5, which corresponds to very distant and unrelated structures. However, here we only compare the structure of complexes for which the composition is matched in the first place based on sequence similarity or domain architecture when no sequence similarity is detected. Hence, low TM-score values are comparatively rare and arise from a lack of quaternary structure conservation rather than from a lack of subunit structure conservation. Altogether there are ~28,700 pairs of QS pairs at a redundancy level of 90% .**B**. We show the same information with two added constraints enforced in QSalign^{HET} to infer that two QSs are conserved. First, matched chains across two complexes must overlap at least 20% (i.e., the shortest chain must cover at least 20% of the longest chain). Second, the TM-score of individual chains is calculated based on the global alignment, and we require a minimum chain-chain score of 0.2 (indicating that chains are at least occupying a similar position in the complex). Most of the pairs with low TM-score are eliminated with these constraints. As a result, most pairs show TM-scores > 0.5, which is why the optimization shown in Fig. 3A appears relatively independent of the TM-score value. Related to Figure 3.



Supplementary Figure S2. Benchmarking of the individual methods and their combination into QSbio separately for dimers and oligomers (assemblies with three subunits and more). Related to Figure 4.



Supplementary Figure S3. Results of PISA and EPPIC benchmark on the full manually curated dataset. **A.** ROC curves show the area under the curve (AUC) for each method for dimers and higher-order oligomers altogether. **B.** Values of statistics derived from the benchmark are shown in the barplots. FPR, false-positive rate; TPR, true positive rate; AUC, area under the curve. Related to Figure 4.



Supplementary Figure S4. Schematic representation of the workflow involved in annotating the hetero-oligomers by QSalign$^{HET.}$ Related to STAR Methods.

Supplementary Methods 1. Description of QSinfer^{HET} and QSpropagate^{HET} routines with pseudo-code. Related to STAR Methods.

Function QSinfer^{HET}:

Retrieve list **L1** of "symmetry type (**SYM**) - number of subunits (**SUB**)" pairs, sorted in decreasing order by number of subunits

For pairs (**SYMi, SUBi**) in **L1**:

Retrieve list **L2** of structure pairs **PDB1, PDB2** that meet the following criteria, sorted by increasing minimum sequence identity.
- Symmetry == **SYMi**
- Number of subunits == **SUBi**
- Maximum Sequence identity < 80%
- Minimum sequence identity > 10%
- Global QS alignment with TM-Score > 0.6
- Minimum TM-score of a chain pair > 0.45
- Overlap of sequence alignment > 0.6
- Number of chains of **PDB1** not having mapped chains in **PDB2** == **ngaps** = 0

 # Note: PDB2_i can be from PISA but is sorted after the match with the PDB structure if it exists

For pairs (**PDB1_i, PDB2_i**) in **L2**:

if **PDB1_i** is not already annotated:

Mark **PDB1_i** as likely correct "*Interface geometry is similar to that of PDB2_i*"
Mark **PDB1_i** as annotated

if **PDB2_i** is not already annotated:

if **PDB2_i** is from PDB:
Mark PDB2_i as likely correct "*Interface geometry is similar to that of PDB1_i*"

elsif **PDB2_i** is generated by PISA:
Mark **PDB2_i** as likely incorrect "*Interface geometry is similar to that of PDB2_i but was detected based on PISA and does not appear in the PDB assembly*"
Mark **PDB2_i** as annotated

Call: QSpropagate^{HET}(**SYMi, SUBi**)

Function QSpropagate^{HET}(SYMi, SUBi):

Retrieve List **L3** of structure pairs **PDB1, PDB2** that meet the following criteria:
- **PDB1** is annotated as likely correct
- **PDB1** symmetry == **SYMi**
- **PDB2** is not yet annotated
- Minimum sequence identity between **PDB1** and **PDB2** > 95%
- Number of chains from the query complex that are missing in the target complex, i.e., number of 'gaps' (defined as **ngaps)**

For pairs (**PDB1_j, PDB2_j**) in **L3**:

Define **#PDB1_j** and **#PDB2_j** as numbers of subunits in **PDB1_j** and **PDB2_j** respectively

Case 1: **#PDB2_j** < **#PDB1_j** and **ngaps_j** = **(#PDB2_j - #PDB1_j)** and **T** > 0.9 and matched composition:
Mark **PDB2_j** as *sub-stoichiometry "This QS has the same composition as **PDB1_j** but subunits are in lower stoichiometry"*

Case 2: **#PDB2_j** < **#PDB1_j** and **ngaps_j** = **(#PDB2_j - #PDB1_j)** and **T** > 0.9 and unmatched composition:

Mark **PDB2_j** as *sub-composition "This QS has a subset of the subunits present in **PDB1_j**"*

Case 3: **#PDB2_j** > **#PDB1_j** and **ngaps_j** = 0 and **T** > 0.9:
Mark **PDB2_j** as *Excessive-stoichiometry "This QS is included in **PDB1_j**"*

Case 4: **#PDB2_j** != **#PDB1_j** and **T** < 0.9:
Mark **PDB2_j** as *Crystal interface or larger conformational change "This QS shows different stoichiometry and/or composition as **PDB1_j** along with significant structural changes"*

Case 5: **#PDB2_j** == **#PDB1_j** and **ngaps_j** = 0 and **T** < 0.65:
Mark **PDB2_j** as *Crystal interface or larger conformational change "This QS shows the same stoichiometry and composition as **PDB1_j** but the structure is different. This might reflect an incorrect QS or may originate in large conformational changes"*

Case 6: If two or more different **QS**s are found to have structural homologs, or if the total number of structural homologs supporting a **QS** is < 5%
Mark **QS** as *Ambiguous*