Supplementary Data for GOSSIP - Global Structural Alignment

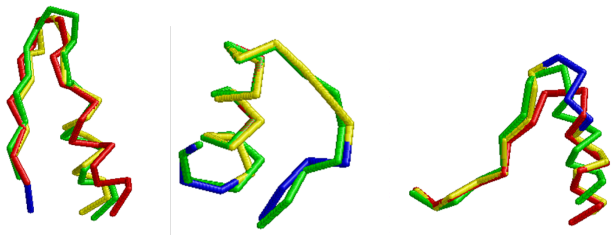
1 Analysis of Fragment Clustering

Figure 1 shows several examples of unique similarities found only by one method, and reflect some of the characteristics of each method. In figure 1(a), MultiProt does not succeed because its algorithm attempts to extend consecutively aligned segments as much as possible. This is clearly seen in the extension of the alignment from the helix to the loop, yet results in a poor alignment to the strand. GOSSIP finds matching quadrilaterals in both the helix and the strand, and the optimal transformation is calculated by minimizing distances on both sides of the loop, resulting in a more balanced alignment that achieves a higher match under $2.0\mathring{A}$. YAKUSA does not find a significant angle match and rules out the alignment as insignificant. In figure 1(b), YAKUSA is only able to detect significantly aligned seeds for the left half of the molecule. The quality of the GOSSIP alignment is slightly worse than that of MultiProt, also due to the right part of the molecule where the query matches less tightly to the target resulting in decreased quadrilateral representation. Increasing the number of refinement iterations enables us to find a closer transformation in this case and many others, but increases running time. We find that in most cases 3 iterations is sufficient, and leave this number as is. Figure 1(c) shows a case in which YAKUSA detects an alignment that was missed by the other two methods under the 80% threshold. However, the superpositions calculated by MultiProt and GOSSIP are identical, suggesting that the difference in alignment sizes between them and YAKUSA is due to the different definition of distance.

Figure 2 presents examples where one method failed to detect a hit detected by the other two methods. In figure 2(a), GOSSIP does not align the helix part of the molecules well due to a lack in closely matching quadrilaterals in that region. Figure 2(b) shows a case where YAKUSA fails to extend its alignment seeds to the edges of the molecules and thus could not detect a large enough alignment. In figure 2(c) MultiProt aligns the left

(a) (b) (c)

Figure 1: Examples of similarities detected by only one of the methods. Query structure is shown in blue, MultiProt superposition is shown in green, GOSSIP superposition is shown in red. As YAKUSA does not output a superposition, we color the parts of the query structure which were aligned by YAKUSA in yellow. (a) An example where GOSSIP detected a unique similarity. YAKUSA did not find an alignment in this case at all. (b) An example of a unique similarity detected by MultiProt (c) An example of a unique similarity detected by YAKUSA. MultiProt and GOSSIP superpositions are identical and are not separable in the figure, suggesting that a better superposition under $2\mathring{A}$ cannot be computed



(a) (b) (c)

Figure 2: Examples of cases for which only one of the methods failed to detect a certain similarity. Query structure is shown in blue, MultiProt superposition is shown in green, GOSSIP superposition is shown in red. As YAKUSA does not output a superposition, we color the parts of the query structure which were aligned by YAKUSA in yellow. (a) A similarity not detected by GOSSIP (b) A similarity not detected by YAKUSA. MultiProt and GOSSIP superpositions are identical and thus barely separable. (c) A similarity not detected by MultiProt.

part better than GOSSIP at the expense of aligning the right part poorly. YAKUSA detects the similar parts well.

2 Supplementary data for CATH and Astral benchmark testing

Supplementary tables 1 and 2 present the numbers used to construct the ROC curves in figure 2 of the manuscript. For clarity, only details for the major similarity thresholds are presented.

Table 1: Method comparison on the Kolodny et al. CATH benchmark

Similarity Threshold	0.6		0.7		0.8		0.9	
	$\sharp TP$	$\sharp$ diff. CA	$\sharp TP$	$\sharp$ diff. CA	$\sharp TP$	$\sharp$ diff. CA	$\sharp TP$	$\sharp$ diff. CA
MultiProt	4,131	491	2,087	86	773	14	113	5
CE	1,994	235	1,250	50	704	8	189	1
GOSSIP	2,325	307	1,529	81	672	15	116	7
YAKUSA	431	141	195	27	62	2	5	2
BLAST	8	0	4	0	4	0	4	0

Comparison of all methods on the CATH sequence-diverse benchmark. Supporting table for the ROC curve presented in figure 2(a) of the manuscript. (*) $\sharp TP$ designates the number of similarities found within the same CATH class. (*) $\sharp$ diff. CA is the number of similarities found that belong to different CA classes of CATH.

Table 2: Method comparison on the Astral 1.55 40% non-redundant benchmark

Similarity Threshold	0.6		0.7		0.8		0.9	
	$\sharp TP$	$\sharp$ diff. fold	$\sharp TP$	$\sharp$ diff. fold	$\sharp TP$	$\sharp$ diff. fold	$\sharp TP$	$\sharp$ diff. fold
MultiProt	5,408	961	3,891	79	2,038	7	461	1
CE	3,316	375	2,546	68	1,814	10	786	1
GOSSIP	4,024	572	3,323	90	1,938	10	497	1
YAKUSA	1,471	123	772	24	268	2	24	0
BLAST	23	0	0	0	0	0	0	0

Comparison of all methods on the Astral 1.55 40% non-redundant benchmark. Supporting table for the ROC curve presented in figure 2(b) of the manuscript. (*) $\sharp TP$ designates the number of similarities found within the same SCOP family. (*) $\sharp$ diff. fold is the number of similarities found that belong to different folds of SCOP.