BioAlign – An Accurate Global PPI Network Alignment Algorithm

Supplementary Information

Umair Ayub, Hammad Naveed

1 Parameter Settings of BioAlign

To maximize the alignment quality in terms of semantic similarity and percentage of aligned nodes, we select the protein pairs that have high sequence or structure similarity. A different set of thresholds for 3D structure similarity, global sequence similarity, and local sequence similarity are tested. Higher/strict thresholds increase the semantic similarity but decrease the percentage of aligned nodes and vice versa. After detailed tuning of similarity thresholds, we find that 3D-structure similarity > 0.5, global sequence similarity > 50 (bit-score), and local sequence similarity > 2.0 achieve the best balance between semantic similarity and percentage of aligned nodes. For 3D-structure similarity, thresholds in a range of $0.3 - 0.8$ are tested. For sequence similarity, thresholds in a range of $30 - 80$ are tested. For local sequence similarity, thresholds in a range of $1.0 - 4.0$ are tested.

In the first stage, the protein pairs that have 3D structure similarity higher than 0.5 are aligned. The remaining protein pairs are aligned if they have global sequence similarity greater than 50. In the last step of the first stage, the remaining protein pairs are aligned if they have local sequence similarity greater than 2.0 and aligned length of sequence greater than 35. In the second stage, the protein pairs are aligned on the basis of remote homology. If two proteins have at least one common homolog, they will be considered as a candidate pair for alignment. The protein pair that have maximum common homologs are aligned first. In the last stage, the remaining proteins are aligned on the basis of predicted secondary structure motifs. From the predicted secondary structure, HTH and HLH motifs are extracted. The protein pairs that have a maximum number of motifs are aligned first.

1.1 The Flow Diagram of BioAlign

Fig. 1 presents the flow diagram of the BioAlign algorithm.

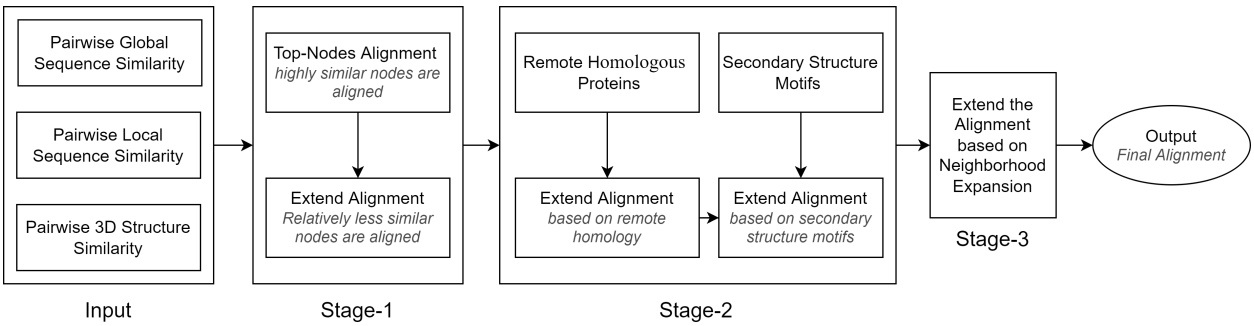

Figure 1: The flow diagram of the BioAlign algorithm. **Input:** Three types of input metrics (global sequence, local sequence and 3D structure similarities) are used. **Stage-1:** The highly similar nodes are aligned first and then the relatively less similar nodes are aligned. **Stage-2:** The remaining nodes are aligned on the basis of remote homologous proteins and predicted secondary structure motifs. **Stage-3:** Seeds are extended using network information (neighbourhood expansion).

1.2 Results of the Different Combinations

BioAlign uses different input metrics (3D structure similarity, global sequence similarity, local sequence similarity, remote homologs, and predicted secondary structure motifs) to align two PPI networks. In global PPI network alignment, the order of the input metrics is influential. Proteins are aligned on the basis of the first metric first and then the remaining proteins are aligned using other scoring metrics. The results of alignments with different orders of inputs are given in Table 2 (columns 3 to 9).

The first column presents the results of stage-1 that use 3D structure, global and local sequence similarities. We align the highly similar nodes in stage-1 that results in accurate but incomplete alignment. To complete the alignment we test topological and non-topological metrics. The second column of Table 2 presents the results of stage-1 in combination with topology, while the third column presents the results of stage-1 with non-topological metrics (remote homologs and predicted secondary structure motifs). The last column presents the results of a default version of BioAlign that utilizes all biological information sources and topology of the network. From the results, we conclude that the combination of biological and topological metrics best optimize the results in terms of AFS and coverage. We also test a variant of BioAlign that incorporates topology after top-nodes alignment (phase-1 of stage-1). The average AFS is slightly reduced but the incompleteness level of the alignments is dropped significantly.

2 Topological Results

This section presents the results of BioAlign and existing aligners in terms of ICS, EC and SSS. The results of BioAlign are better than SANA, MONACO, Twadn, and BEAMS. The difference between the results of these aligners is notable. The results of BioAlign are inferior as compared to PROPER, MAGNA++, NETAL, SAlign, HubAlign and ModuleAlign. MAGNA and NETAL outperform all the existing aligners. MAGNA optimizes the align-

Table 1: The results of the different combinations of inputs.

Pairs	Eval	Results of the Stages of BioAlign								
		Com.1	Com.2	Com.3	Com.4	Com.5	Com.6	Com.7	Com.8	Com.9
MH	AFS_{MF}	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.75
	AFS_{BP}	0.67	0.67	0.67	0.66	0.66	0.66	0.66	0.67	0.64
	$Nodes_{MF}$	88	88	88	88	88	88	89	88	88
	$Nodes_{BP}$	92	92	92	92	92	92	90	92	92
MY	AFS_{MF}	0.51	0.46	0.47	0.46	0.47	0.47	0.46	0.46	0.51
	AFS_{BP}	0.35	0.31	0.32	0.31	0.32	0.32	0.32	0.31	0.35
	$Nodes_{MF}$	56	71	71	70	71	70	70	73	35
	$Nodes_{BP}$	65	85	85	85	84	84	84	88	42
YH	AFS_{MF}	0.60	0.53	0.55	0.54	0.55	0.56	0.54	0.55	0.55
	AFS_{BP}	0.45	0.39	0.42	0.41	0.42	0.42	0.41	0.41	0.40
	$Nodes_{MF}$	52	62	63	63	62	61	63	63	48
	$Nodes_{BP}$	60	73	73	73	72	72	72	74	56
MF	AFS_{MF}	0.68	0.66	0.67	0.66	0.67	0.68	0.66	0.67	0.64
	AFS_{BP}	0.50	0.48	0.49	0.49	0.49	0.49	0.48	0.49	0.46
	$Nodes_{MF}$	76	77	78	79	77	75	77	78	71
	$Nodes_{BP}$	81	82	83	83	83	80	82	83	75
MW	AFS_{MF}	0.63	0.58	0.61	0.61	0.61	0.62	0.60	0.62	0.60
	AFS_{BP}	0.46	0.43	0.45	0.44	0.44	0.45	0.45	0.45	0.43
	$Nodes_{MF}$	69	69	73	72	72	72	73	73	60
	$Nodes_{BP}$	63	64	67	67	67	65	66	68	43
Avg.	AFS_{MF}	0.64	0.60	0.62	0.61	0.62	0.62	0.61	0.62	0.61
	AFS_{BP}	0.48	0.45	0.47	0.46	0.47	0.47	0.46	0.47	0.46
	$Nodes_{MF}$	68	73	75	75	74	73	75	75	61
	$Nodes_{BP}$	72	79	80	80	79	78	79	81	64

Com.1: 3D-Structure + Global-Sequence + Local-Sequence (Stage-1)
Com.2: 3D-Structure + Global-Sequence + Local-Sequence + Topology (Neighbours-Extension)
Com.3: 3D-Structure + Global-Sequence + Local-Sequence + Remote-Homology + Secondary-Structure
Com.4: 3D-Structure + Global-Sequence + Local-Sequence + Secondary-Structure + Remote-Homology
Com.5: Global-Sequence + Local-Sequence + 3D-Structure + Remote-Homology + Secondary-Structure
Com.6: Local-Sequence + Global-Sequence + 3D Structure + Remote-Homology + Secondary-Structure
Com.7: Global-Sequence + 3D-Structure + Local-Sequence + Remote-Homology + Secondary-Structure
Com.8: Global-Sequence + 3D-Structure + Local-Sequence + Remote-Homology + Secondary-Structure + Topology
Com.9: Global-Sequence + 3D-Structure + Local-Sequence (Top-Alignment only) + Topology

Table 2: The results of the different Aligners in terms of ICS, EC and SSS.

Pairs	Eval	Results of the Stages of BioAlign											
		BA	BA2	SA	PR	HA	MA	NE	SAN	MAG	MON	TW	BE
MH	ICS	0.16	0.17	0.23	0.26	0.20	0.17	0.70	0.02	0.71	0.13	0.21	0.19
	EC	0.11	0.22	0.45	0.15	0.70	0.45	0.61	0.01	0.33	0.14	0.09	0.14
	SSS	0.07	0.08	0.18	0.10	0.19	0.14	0.48	0.01	0.29	0.07	0.07	0.09
MY	ICS	0.17	0.32	0.22	0.29	0.24	0.22	0.66	0.02	0.80	0.08	0.06	0.06
	EC	0.16	0.15	0.61	0.53	0.68	0.37	0.60	0.01	0.34	0.03	0.01	0.02
	SSS	0.09	0.12	0.19	0.23	0.22	0.16	0.45	0.01	0.32	0.02	0.01	0.01
YH	ICS	0.06	0.13	0.08	0.20	0.13	0.20	0.31	0.01	0.24	0.03	0.02	0.03
	EC	0.06	0.11	0.09	0.26	0.17	0.32	0.41	0.01	0.14	0.03	0.01	0.01
	SSS	0.03	0.06	0.05	0.13	0.08	0.14	0.21	0.01	0.11	0.01	0.01	0.01
MF	ICS	0.09	0.22	0.28	0.54	0.34	0.31	0.81	0.01	0.78	0.14	0.09	0.13
	EC	0.05	0.18	0.54	0.38	0.65	0.54	0.57	0.01	0.49	0.08	0.02	0.05
	SSS	0.03	0.11	0.23	0.19	0.28	0.22	0.50	0.01	0.43	0.05	0.01	0.04
MW	ICS	0.05	0.13	0.19	0.24	0.22	0.19	0.65	0.01	0.68	0.06	0.07	0.03
	EC	0.07	0.21	0.47	0.52	0.63	0.51	0.60	0.01	0.49	0.06	0.02	0.01
	SSS	0.03	0.09	0.15	0.20	0.20	0.18	0.45	0.01	0.40	0.03	0.02	0.01
Avg.	ICS	0.11	0.19	0.20	0.31	0.23	0.22	0.63	0.01	0.64	0.09	0.09	0.08
	EC	0.09	0.17	0.43	0.37	0.57	0.44	0.56	0.01	0.36	0.07	0.03	0.05
	SSS	0.05	0.09	0.16	0.17	0.19	0.17	0.42	0.01	0.31	0.04	0.02	0.03

BA:Stage1+Topology, BA2:Top-Alignment+Topology, SA: SAlign, PR: PROPER, HA: HubAlign, MA:ModuleAlign, NE:NETAL, SAN:SANA, MAG:MAGNAA++, MON:MONACO, TW:Twadn, BE:BEAMS

ments using topological measures while NETAL incorporates 100% topological information that result in best performance of these aligners in terms of topological measures. HubAlign and ModuleAlign incorporates 40-60% topology while aligning the PPI networks. PROPER incorporates topology in a stage wise manners. All these aligners use a high amount of topological information that results in better alignments in terms of topological measures. From Tables 3 and 4 (of the main article), we can see that all these aligners show poor performance in terms of biological similarity and coverage.