GRAFENE: graphlet-based alignment-free network approach that integrates 3D structural and sequence data to improve protein structural comparison

Fazle Faisal, Khalique Newaz, Julie Chaney, Jun Li, Scott Emrich, Patricia Clark, and Tijana Milenković*

[1]University of Notre Dame *Corresponding Author (E-mail: tmilenko@nd.edu)

How to efficiently compare protein structures?

- **Early methods** for protein structural comparisons were **sequence-based**.
- Amino acids that are distant in the sequence can be close in the **3-dimensional (3D) structure**.
- 3D contact approaches can complement sequence approaches.
- Traditional 3D contact approaches study 3D structures directly.
- 3D structures can be modeled as **protein structure networks (PSNs)** (see Figure 1).

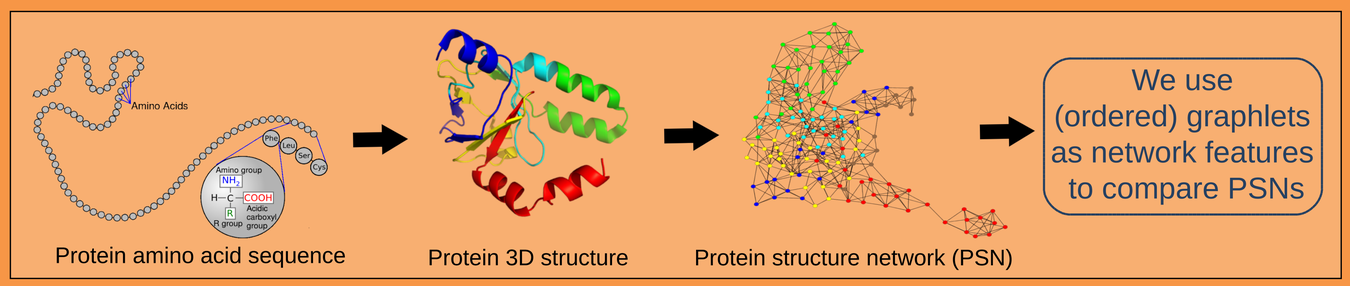

Figure 1: A protein, whose amino acid sequence folds into a 3-dimensional (3D) structure, can be modeled as a protein structure network (PSN). In the PSN, nodes are amino acids and two amino acids are linked by an edge if they are spatially close enough in the 3D structure.

- Network (i.e., PSN) approaches may improve upon traditional 3D contact approaches.
- We cannot use existing PSN approaches to test this, because:
  - They rely on naive measures of network topology.
  - They cannot integrate PSN data with sequence data.

- We address these limitations by:
  - Exploiting well established graphlet measures via a new network approach.
  - Using ordered graphlets to combine the complementary PSN data and sequence data.
- We thoroughly evaluate 24 different approaches for protein comparison (see Figure 2).
- We evaluate the 24 approaches by measuring how well they can distinguish between ~17,000 protein domains that are categorized into ~120 different protein domain groups according to SCOP and CATH databases.

Methodology

- **Graphlets** are equivalence classes of isomorphic connected induced subgraphs [1].
- We use graphlets to study (and in particular, to compare) PSNs [2].
- Existing PSN approaches for protein structural comparison cannot integrate PSN data and sequence data (see above).
- We develop a new PSN approach that is based on a recent notion of **ordered graphlets** (see Figure 3) [3].
- An ordered graphlet is an equivalence class of *labeled* isomorphic connected induced subgraphs; labels account for the (relative rather than absolute) order of amino acid positions in the protein sequence.
- Given a PSN, we count the occurrence of each ordered graphlet to obtain the PSN's **graphlet frequency vector (GFV)**.
- We then apply principal component analysis (PCA) to GFVs of all of the PSNs, in order to compare PSNs (see Figure 4).

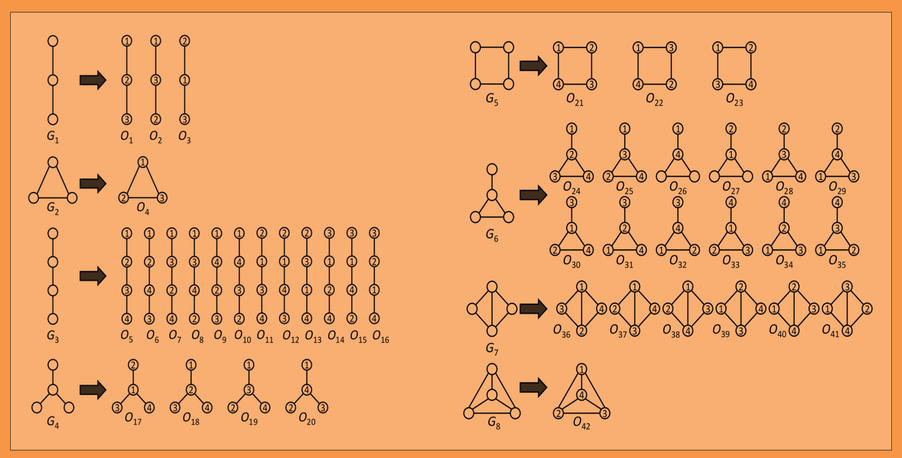

Figure 3: All eight unordered 3-4-node graphlets (G_1,G_2,..,G_8; on the left of the arrows) and their corresponding 42 ordered graphlets (O_1,O_2,..,O_{42}; on the right of the arrows).

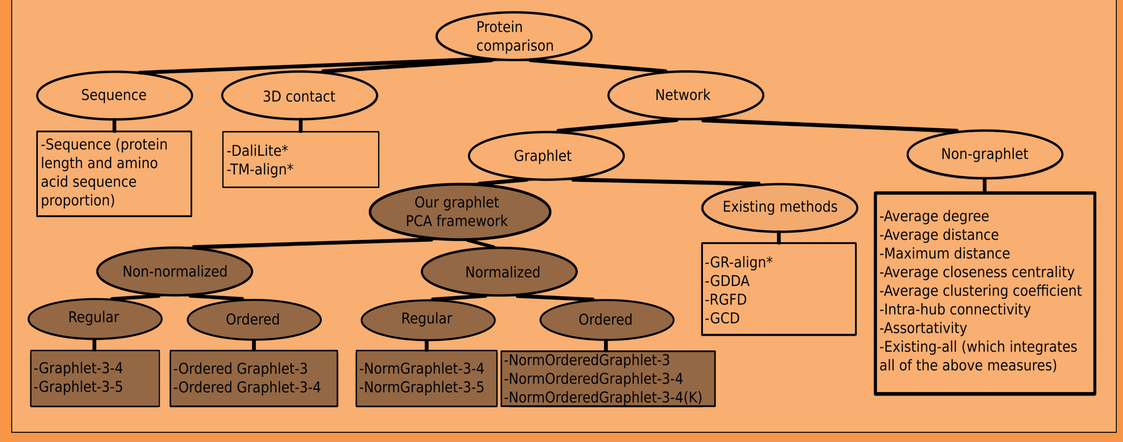

Figure 2: Categorization of the 24 approaches (listed in squares) that we evaluate. Different versions of our proposed graphlet approach are colored in gray.

Results

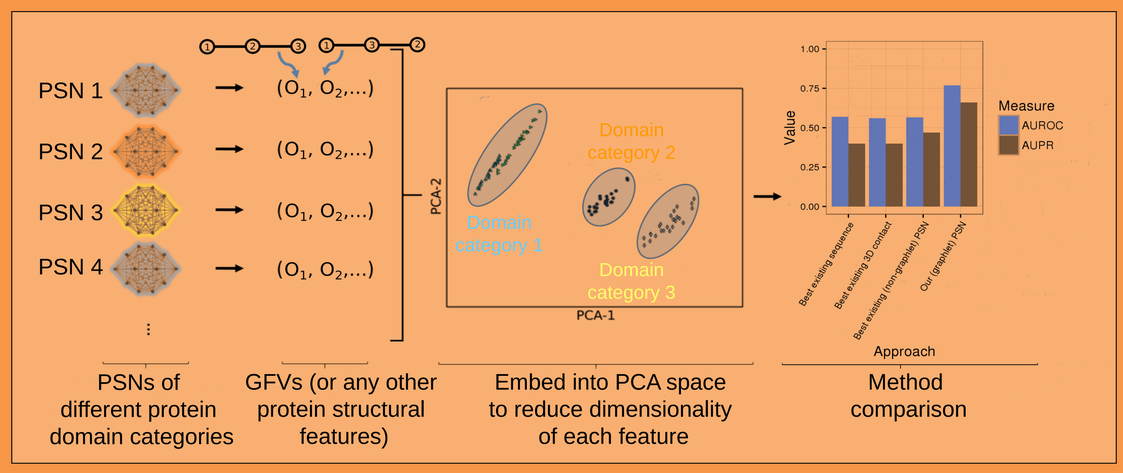

Figure 4: Our PCA framework for protein structural comparison.

Approach	Rank-based			
	AUPR		AUROC	
	Avg rank	*p*-value	Avg rank	*p*-value
Graphlet-3-4	9.91	2.94e-11	12.80	3.81e-15
Graphlet-3-5	12.34	1.61e-14	12.94	3.46e-15
NormGraphlet-3-4	10.89	2.25e-18	10.14	5.83e-15
NormGraphlet-3-5	10.25	5.87e-16	9.26	4.57e-14
OrderedGraphlet-3	11.03	1.92e-13	13.09	2.61e-14
OrderedGraphlet-3-4	7.91	1.49e-14	8.91	8.79e-10
NormOrderedGraphlet-3	10.77	3.49e-13	10.48	7.33e-11
NormOrderedGraphlet-3-4	4.31	5.23e-07	5.14	1.28e-06
NormOrderedGraphlet-3-4(K)	**1.83**	-	**1.97**	-
GDDA	16.17	1.66e-15	17.37	1.67e-15
RGFD	11.29	1.55e-13	11.60	2.01e-12
GCD	15.71	4.21e-16	15.43	2.93e-13
GR-Align	4.43	2.10e-03	6.68	8.91e-06
Average degree	18.85	3.32e-20	16.00	8.27e-16
Average distance	17.66	4.04e-19	16.91	1.19e-16
Maximum distance	16.03	3.18e-17	14.83	1.14e-14
Average closeness centrality	16.31	9.70e-18	15.51	2.31e-14
Average clustering coefficient	18.6	6.65e-22	15.54	4.83e-16
Intra-hub connectivity	14.37	8.99e-12	15.80	2.33e-16
Assortativity	21.00	2.29e-24	19.00	3.53e-17
Existing-all	10.14	6.49e-15	9.57	3.28e-11
DaliLite	9.14	1.84e-06	6.29	6.21e-04
TM-align	18.23	5.38e-15	20.09	3.05e-19
AAComposition	12.80	6.40e-12	14.63	1.16e-13

Figure 5: Summary of method accuracy. Accuracy of the given approach is shown with respect to its average ranking compared to all considered approaches across all considered PSN sets, and the results are shown based on AUPR as well as AUROC. The ranking of each method is expressed as follows. For the given PSN set, we determine which approach results in the highest accuracy (rank 1), the second highest accuracy (rank 2), etc. Then, we average the rankings of the given method over all PSN sets. So, the lower the average rank, the better the method.Since NormOrderedGraphlet-3-4(K) has the best average rank with respect to both AUPR and AUROC (shown in bold), we compute the statistical significance of the improvement of NormOrderedGraphlet-3-4(K) over each of the other approaches in terms of their ranks.

References

[1] Hulovatyy and Milenković (2015), B*ioinformatics*, 31, 171-180.
[2] Malod-Dognin and Pržulj (2014), *Bioinformatics*, 30, 1259–65.
[3] Faisal et al. (2017), *Scientific Reports,* 7, 14890. **(This work.)**

FUNDING: NIH 1R01GM120733-01A1, 1R21AI111286-01A1, and R01GM074807; AFOSR YIP FA9550-16-1-0147; Clare Boothe Luce Graduate Research Fellowship.