Network approach integrates 3D structural and sequence data to improve protein structural comparison

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

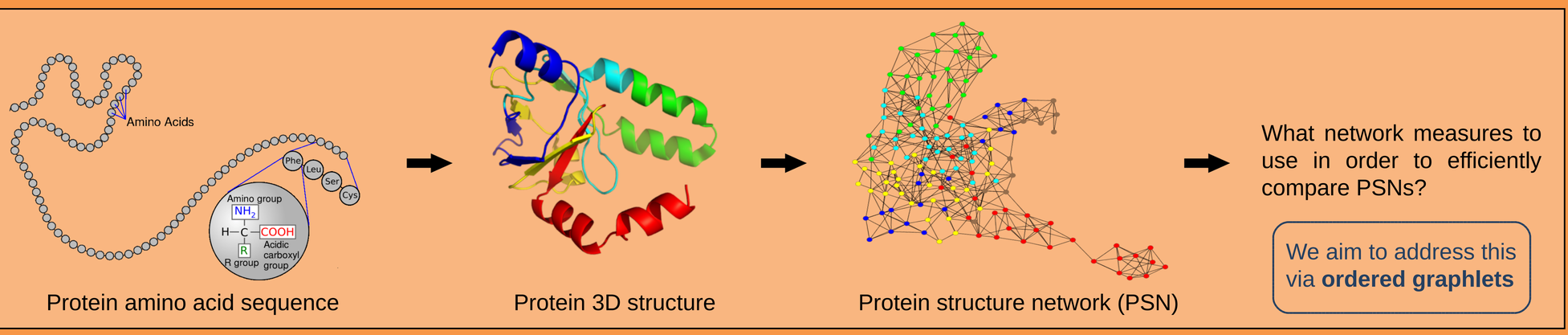
Department of Computer Science and Engineering, Department of Chemistry and Biochemistry,
Department of Applied and Computational Mathematics and Statistics, University of Notre Dame
Interdisciplinary Center for Network Science and Applications; ECK Institute for Global Health
***Corresponding Author (E-mail: tmilenko@nd.edu)**

How to efficiently compare protein structures?

- Initial protein structural comparisons were sequence-based.
- Since 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.

Instead, 3D structures can be modeled as protein structure networks (PSNs) (Figure 1).

- Then, network approaches can compare proteins by comparing their PSNs.
- Network (or 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, although this could help because the different data types capture complementary biological knowledge.
- 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 of protein comparison (Figure 1).
- We use SCOP and CATH protein domain categorization databases to categorize ~17,000 protein domains and use these labeled domains to evaluate all 24 approaches.



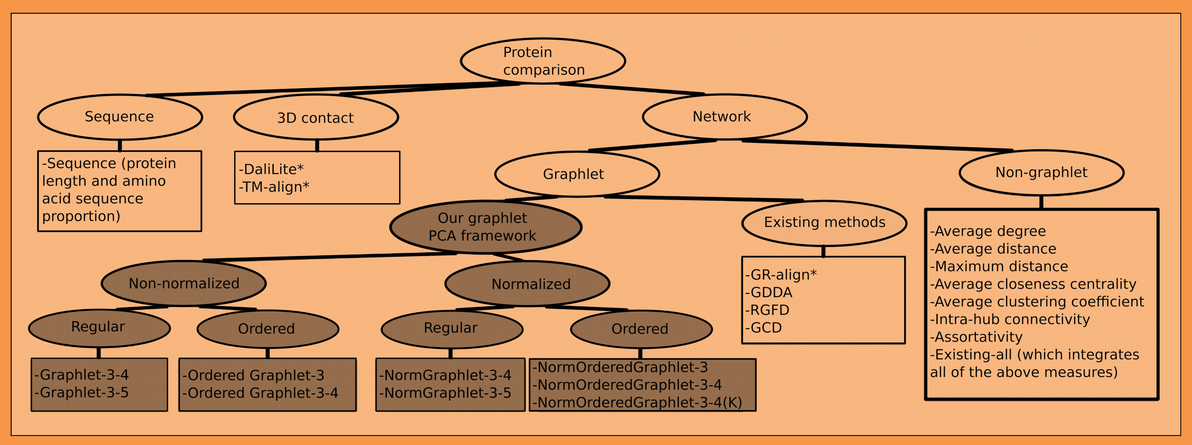


Figure 2: Categorization of the 24 approaches (listed in squares) that we evaluate. Different versions of our proposed graphlet approach are colored in gray.

Figure 1: Illustration of how 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. We can extract network features from the respective PSNs of the proteins and use these features to structurally compare them.

Methods

Graphlets are small connected induced subgraphs. They have been proven as sensitive and superior measures of topology in numerous contexts when studying protein-protein interaction networks and protein structure networks [1,2].

Existing PSN approaches for protein comparison have drawbacks (as outlined above). To overcome these drawbacks, we take a relatively current notion of **ordered graphlets** (Figure 3) to take advantage of both the network data and the sequence data [3]. We define an ordered graphlet as an equivalence class of isomorphic connected subgraphs; equivalence is taken with respect to the relative order of **amino acid positions** (in the protein sequence), without considering the amino acids' actual positions.

We count ordered graphlets in the PSNs to obtain PSN-specific graphlet frequency vectors (GFVs). We then apply principal component analysis (PCA) framework on the GFVs, in order to compare PSNs (Figure 4).

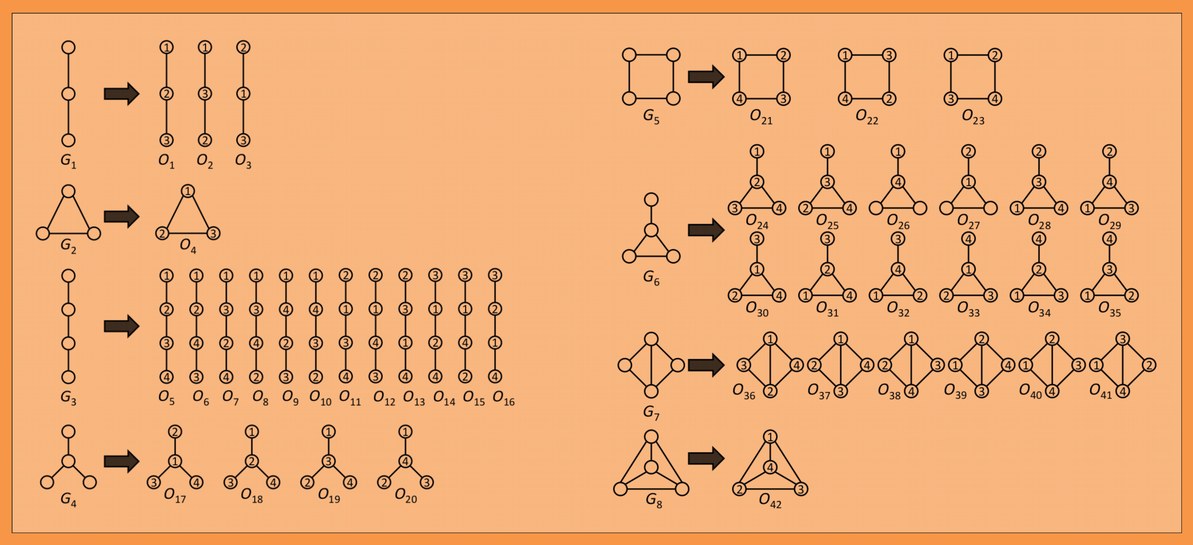


Figure 3: All possible eight regular (unordered) 3-4-node graphlets (G_1, G_2, .., G_8; on the left of arrow) and their corresponding 42 ordered graphlets (O_1, O_2, .., O_{42}; on the right of arrow).

Results

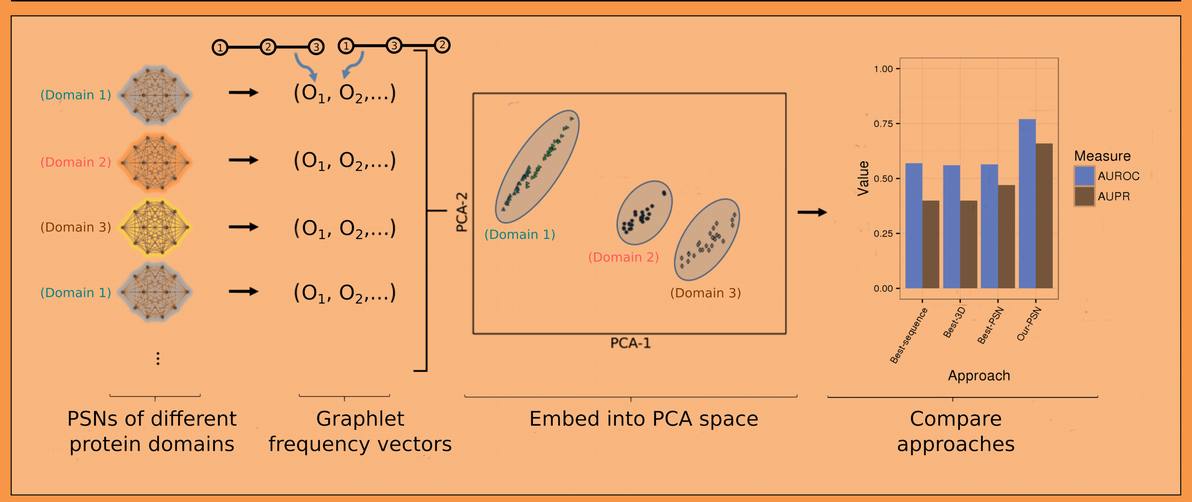


Figure 4: Comparison of PSNs in our PCA framework

Approach	AUPR		AUROC		Running
	Rank	*p*-value	Rank	*p*-value	time (hrs)
Graphlet-3-4	8.38	9.42e-05	10.50	0.000147	0.43
Graphlet-3-5	9.00	4.81e-06	10.40	8.74e-05	0.49
OrderedGraphlet-3	7.15	0.00225	9.92	0.000692	0.38
OrderedGraphlet-3-4	7.31	0.00143	8.69	0.0018	2.39
NormGraphlet-3-4	7.77	3.57e-05	8.15	0.000156	0.44
NormGraphlet-3-5	8.15	5.04e-05	6.69	0.00124	0.51
NormOrderedGraphlet-3	10.50	4.33e-05	9.92	0.000135	0.39
NormOrderedGraphlet-3-4	4.31	0.000999	4.92	0.00127	2.41
NormOrderedGraphlet-3-4(K)	**1.69**	-	**2.08**	-	2.41
GDDA	17.30	6.16e-09	17.70	2.57e-08	0.54
RGFD	9.46	6.84e-06	9.85	1.39e-05	0.49
GCD	17.10	1.21e-09	17.10	1.51e-08	1.32
GR-align	8.31	0.00705	9.69	0.00423	9.49
Average degree	18.90	2.32e-10	16.20	2.02e-07	0.39
Average distance	15.40	9.54e-07	16.50	3.59e-06	0.48
Maximum distance	17.30	1.58e-09	16.90	4.95e-08	0.49
Average closeness centrality	18.50	2.18e-08	16.50	3.08e-07	0.48
Average clustering coefficient	16.80	5.01e-08	14.50	3.55e-07	0.56
Intra-hub connectivity	16.40	2.84e-08	15.10	1.14e-06	0.64
Assortativity	20.10	1.79e-08	19.20	1.48e-07	0.46
Existing-all	10.90	1.33e-06	10.00	3.05e-05	1.01
DaliLite	12.70	3.27e-05	10.60	0.00192	2021.41
TM-align	22.00	1.85e-12	22.30	5.75e-12	168.32
Sequence	14.50	1.44e-06	16.60	2.1e-08	0.24

Figure 5: Summary of method accuracy and running times. Accuracy of the given approach is shown with respect to its average ranking compared to all considered approaches across all considered real-world 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, using paired t-test. Running times of the approaches are shown when comparing proteins from the CATH-α set. Running times for the other data sets are qualitatively the same.

References

[1] Malod-Dognin and Pržulj (2014), *Bioinformatics* 30, 1259–65.
[2] Faisal. & Milenković (2014), B*ioinformatics* 30, 1721–1729.
[2] Faisal et al. (2016), *arXiv:1605.07247 [q-bio.MN]*. **(This work.)**

FUNDING: NIH 1R01GM120733-01A1, 1R21AI111286-01A1, and R01GM074807; AFOSR YIP FA9550-16-1-0147; Clare Boothe Luce Graduate Research Fellowship; NSF CAREER CCF-1452795, CCF-1319469.