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

Fazle E. Faisal\textsuperscript{1,5,6}, Julie L. Chaney\textsuperscript{2,7}, Jun Li\textsuperscript{3}, Patricia L. Clark\textsuperscript{2,4,6} and Tijana Milenković\textsuperscript{1,5,6,\ast}

\textsuperscript{1}Department of Computer Science and Engineering; \textsuperscript{2}Department of Chemistry and Biochemistry; \textsuperscript{3}Department of Applied and Computational Mathematics and Statistics; \textsuperscript{4}Department of Chemical and Biomolecular Engineering; \textsuperscript{5}Interdisciplinary Center for Network Science and Applications; \textsuperscript{6}ECK Institute for Global Health, University of Notre Dame, Notre Dame, IN 46556, USA; \textsuperscript{7}Siemens Healthineers, Elkhart, IN 46516, USA

To whom correspondence should be addressed.

ABSTRACT

Motivation: Early approaches for protein (structural) classification 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 first be modeled as protein structure networks (PSNs). Then, network approaches can be used to classify the PSNs. Network approaches may improve upon traditional 3D contact approaches. We cannot use existing PSN approaches to test this, because: 1) They rely on naive measures of network topology that cannot capture the complexity of PSNs. 2) They are not robust to PSN size. They cannot integrate 3) multiple PSN measures or 4) PSN data with sequence data, although this could help because the different data types capture complementary biological knowledge.

Results: We address these limitations by: 1) exploiting well-established graphlet measures via a new network approach for protein classification, 2) introducing novel normalized graphlet measures to remove the bias of PSN size, 3) allowing for integrating multiple PSN measures, and 4) using ordered graphlets to combine the complementary ideas of PSN data and sequence data. We classify both synthetic networks and real-world PSNs more accurately and faster than existing network, 3D contact, or sequence approaches. Our approach finds PSN patterns that may be biochemically interesting.

Availability: Our software and the Supplementary Materials are available upon request.

Contact: tmilenko@nd.edu

1 INTRODUCTION

1.1 Motivation and background

Proteins perform important cellular functions. While understanding protein function is clearly important, doing so experimentally is expensive and time-consuming (Ashburner \textit{et al.}, 2000; Kasabov, 2013). Because of this, the functions of many proteins remain unknown (Lee \textit{et al.}, 2007; Kasabov, 2013). Consequently, computational prediction of protein function has received attention, including a strategy called protein structural classification (PC). PC divides proteins into groups based on their similarity with respect to sequence or 3-dimensional (3D) structural patterns (or features), in order to predict functions of unannotated proteins based on functions of annotated proteins from the same group.

Early PC has relied on sequence analyses (Mizianty \textit{et al.}, 2010; Sułkowska \textit{et al.}, 2012). Due to advancements of high-throughput sequencing technologies, rich sequence data is available for many species, and thus, comprehensive sequence pattern searches are possible.

Amino acids that are distant in the linear sequence can be close in 3D structure. Thus, 3D structural analyses can reveal patterns that might not be apparent from the sequence alone. For example, while high sequence similarity between proteins typically indicates their high structural and functional similarity (Lee \textit{et al.}, 2007), proteins with low sequence similarity can still be structurally similar and perform similar function (Krissinel, 2006; Gao and Li, 2009). In this case, 3D structural approaches, unlike sequence approaches, can correctly identify structurally and thus functionally similar proteins. On the other extreme, proteins with high sequence similarity can be structurally dissimilar and perform different functions (Tuinstra \textit{et al.}, 2008; Kosloff and Kolodny, 2008; Clarke and Clark, 2008; Burmann \textit{et al.}, 2012). In this case, 3D structural approaches, unlike sequence approaches, can correctly identify structurally and thus functionally different proteins.

3D structural approaches can be classified into traditional 3D contact approaches and recent network approaches. 3D contact approaches, which typically deal with 3D structural alignment, study 3D structures directly (Holm and Rosenström, 2010; Zhang and Skolnick, 2005). Instead, network approaches first model 3D structures as protein structure networks (PSNs, or contact maps, in which nodes are amino acids and edges link spatially close amino acids (Milenković \textit{et al.}, 2009)) and then compare the PSNs to classify the corresponding proteins. 3D contact approaches produce rigid protein alignments while comparing 3D structures (Malod-Dognin and Pržulj, 2014). Hence, they may not perform well in the task of PC. Also, 3D contact approaches are typically slow. This requires alternative, more flexible and faster PC
approaches. Since PSNs model spatial proximities of amino acids within the protein 3D structure, network analyses of PSNs have a potential to complement and improve upon 3D contact (as well as sequence) approaches in the task of PC.

Network analyses of 3D structures have received attention (Gao and Li, 2009; Emerson and Gothandam, 2013), e.g., when classifying functionally different proteins (Pabuwal and Li, 2008, 2009; Emerson and Gothandam, 2012). However, these approaches have limitations:

1. They rely on naive measures of network topology, which capture the global view of a network but ignore complex local interconnectivities that exist in real-world networks, including PSNs (Milenković et al., 2008; Memisević et al., 2010; Kuchaiev et al., 2011). Hence, a more sensitive measure of local network topology might improve PC.
2. They can bias PC by PSN size: networks of similar topology but different sizes can be misclassified into separate groups by the existing approaches simply because of their size difference. Thus, PC strategies are needed that can avoid the PSN size bias.
3. Because different network measures quantify the same PSN topology from different perspectives (Emerson and Gothandam, 2013; Faisal and Milenković, 2014), and because the existing approaches rely on a single network measure, PC could be biased towards the PSN perspective captured by the given measure. Integration of different and complementary network measures could improve PC.
4. Almost all existing network approaches ignore valuable sequence information (also, the existing sequence approaches ignore valuable PSN information). Combining the complementary ideas of network and sequence measures could improve PC.

1.2 Our contributions

We present a new network classification framework that is based on principal component analysis (PCA) and that overcomes the above drawbacks of the existing approaches. Specifically:

1. We use graphlets (Pržulj et al., 2004; Pržulj, 2007), a sensitive measure of local network topology, in hope to improve PC upon the existing network approaches. While graphlets have already been proven in analyses of protein-protein interaction networks (Hulovatyy et al., 2014, 2015; Solava et al., 2012), here we use them in a novel application of PC. Also, we use them within our PCA framework, where none of the existing graphlet methods rely on PCA.
2. We perform graphlet normalization to address the bias of PSN size.
3. We allow for integrating different and complementary network topological measures within our framework.
4. We adopt the idea of ordered graphlets (Malod-Dognin and Pržulj, 2014) to integrate the PSN amino acid interconnectivity with sequence information, in order to improve upon network approaches alone or sequence approaches alone. While ordered graphlets are an existing idea, this idea was introduced only on up to 3-node graphlets. However, using larger graphlets can be beneficial in many real-world contexts (Faisal and Milenković, 2014; Hulovatyy et al., 2014, 2015; Solava et al., 2012). Hence, we extend the existing notion of 3-node ordered graphlets both theoretically and implementation-wise to be able to deal with larger graphlets. We include the implementation of the extended idea of ordered graphlets into our software. Importantly, even when we use only the existing up to 3-node ordered graphlets, our new PCA framework still outperforms the existing approach that is based on the same ordered graphlets (Malod-Dognin and Pržulj, 2014). This validates the PCA framework as a whole.

We study two network types: synthetic networks and real-world PSNs. For each network type, we analyze multiple data sets. In each data set, each network has a known classification label, meaning that we know which network belongs to which class and thus which networks should group together (Sections 2.1 and 2.2). For synthetic networks, we study 20 network approaches (Fig. 1), of which eight are different versions of our proposed graphlet PCA approach (Section 2.3.2) and 12 are existing (non-PCA) approaches (of which four use graphlets and eight do not use graphlets; Section 2.4.1). Given a data set and a network approach, we compute similarity/distance between each pair of networks. We evaluate each approach by measuring how accurately it can group networks from the same class and separate networks from different classes. We measure this by computing the area under precision-recall curve (AUPR) and area under receiver operator characteristic curve (AUROC) (Section 2.3). For real-world networks, in addition to the 20 network approaches, we also study two 3D contact approaches (Section 2.4.2) and a sequence approach (Section 2.4.3), and we perform the same AUPR and AUROC evaluation. (These three approaches cannot be used on the synthetic networks.)

Our key findings are as follows. Over all data sets, our graphlet PCA approach is superior to the existing (non-PCA) graphlet and non-graphlet network approaches. We demonstrate the importance of removing the bias in PSN size, leading to a normalized version of our graphlet PCA approach that is superior to its non-normalized counterpart when controlling for network size. Combining our normalized graphlet PCA approach with sequence data via ordered graphlets results in superior accuracy compared to any network approach alone or sequence approach alone, which confirms that data integration helps. Our network approach outperforms the traditional 3D contact approaches in terms of both accuracy and running time, which confirms the power of network analyses of 3D structures. Our approach reveals PSN patterns that may be biochemically interesting.
2 MATERIALS AND METHODS

2.1 Data
We collect 3D atomic structures of 17,036 proteins from Protein Data Bank (PDB) (Berman et al., 2000). We denote this data set as ProteinPDB (Supplementary Section S1.1). Each protein is comprised of the X, Y, and Z orthogonal Angstrom (Å) co-ordinates of heavy atoms (i.e., carbon, nitrogen, oxygen, and sulfur) of each amino acid within the protein.

Class, Architecture, Topology, Homology (CATH) is a protein domain classification database (Sillitoe et al., 2015; Orengo et al., 1999). A protein is typically composed of one or more domains (a domain refers to a common protein structure), and the purpose of CATH is to annotate these domains. CATH’s classification scheme is hierarchical. Its top hierarchy classifies protein domains into four groups: alpha (α), beta (β), alpha/beta (α+β), and few secondary structures. As annotated by CATH, the 17,036 proteins in ProteinPDB are comprised of 17,884 domains. Among them, 5,324 belong to α, 5,665 belong to β, and 6,895 belong to α+β. None of the domains in ProteinPDB belong to few secondary structures.

Structural Classification of Proteins (SCOP) (Murzin et al., 1995) is another protein domain classification database whose classification scheme is also hierarchical. SCOP’s top hierarchy classifies protein domains into 11 groups: alpha (α), beta (β), alpha plus beta (α+β), coiled coil, membrane, multi-domain, small, low resolution, peptide, and designed. As annotated by SCOP, the 17,036 proteins in ProteinPDB are comprised of 15,762 domains. Among them, 3,065 belong to α, 3,730 belong to β, 4,172 belong to α+β, 219 belong to coiled coil, 235 belong to membrane, and 323 belong to multi-domain. None of the domains in ProteinPDB belong to small, low resolution, peptide, or designed.

2.2 Forming networks
We evaluate the considered approaches in the task of PC on: 1) synthetic networks, i.e., on artificially generated networks for which we know the topology-based ground truth classification, and 2) real-world PSNs, for which we know CATH or SCOP label-based classifications that we hypothesize correlate well with the PSNs’ topology-based characteristics.

Synthetic networks. We generate synthetic networks by using different network models. A good approach should identify networks from the same network model as similar and classify such networks together into the same group (i.e., class), and it should identify networks from different models as dissimilar and separate such networks into different groups (i.e., classes). Specifically, we use three well-established network models: Erdős-Rényi random graphs (ER), geometric random graphs (GEO), and scale-free random graphs (SF) (Milenković et al., 2008; Kuchaiev et al., 2011).

First, we evaluate the considered approaches on synthetic networks of the same size but of different classes (originating from the three network models). To evaluate the robustness of our PC framework to the choice of network size, we repeat this analysis three times, by increasing the size of the considered networks by a factor of 1000 (Table 1), where each set consists of 50 networks per model (totaling to 50 × 3 = 150 networks). The numbers of nodes and edges in these networks are chosen in a way so that the networks closely mimic sizes of real-world PSNs.

Second, we evaluate the considered approaches on networks of different sizes as well as different classes, to check whether the approaches can correctly group networks from the same model despite the networks being of different sizes, as well as that they can correctly separate networks from different models despite the networks being of the same size. To generate a synthetic network set of different sizes, we combine networks from Synthetic-100, Synthetic-500, and Synthetic-1000 together. We denote the combined network set as Synthetic-all (Table 1).

Real-world PSNs (with CATH or SCOP classification). Each protein in ProteinPDB (Section 2.1) is composed of 3D co-ordinates of the heavy atoms of its amino acids. Given a protein, we use the 3D co-ordinate information to compute Euclidean distance between each pair of amino acids and then construct a PSN in which nodes represent amino acids and edges connect pairs of amino acids that are sufficiently close (i.e., within a given distance cut-off) in the protein’s 3D structure. We use the suggested distance cut-off of 4 Å (Milenković et al., 2009). Given all resulting PSNs, there are four possible PSN sets with CATH classification: CATH-primary, CATH-α, CATH-β, and CATH-α+β (Table 1 and Supplementary Section S1.2). Similarly, there are six possible PSN sets with SCOP classification: SCOP-primary, SCOP-α, SCOP-β, SCOP-α+β, SCOP-multidomain (Table 1 and Supplementary Section S1.2).
To ensure that PC is not biased by PSN size, we need a data set with PSNs of the same (or at least similar) network size. Hence, focusing on PSNs of α and β classes from the CATH-primary data set, we infer three such same-size PSN data sets, denoted as CATH-95, CATH-99, and CATH-251-265 (Table 1 and Supplementary Section S1.2).

Table 1. Synthetic network and real-world PSN data sets that we use. For the given data set, the second column indicates whether its networks are of the same size or different sizes, and the last three columns indicate the number of its networks as well as their size(s) in terms of the number of nodes and edges.

| Type               | Size       | Number of Networks | Nodes | Edges |
|--------------------|------------|--------------------|-------|-------|
| Synthetic network  | Same       | 150                | 100   | 400   |
|                    | Different  | 450                | 100-1,000 | 400-4,000 |
| Real-world PSN     | Same       | CATH-95            | 24    | 95    |
|                    |            | CATH-99            | 28    | 99    |
|                    |            | CATH-251-265       | 16    | 251-265 | 1,003-1,076 |
|                    | Different  | CATH-primary       | 9,509 | 101-872 | 243-3,849 |
|                    |            | CATH-α             | 2,628 | 101-872 | 320-3,849 |
|                    |            | CATH-β             | 3,085 | 101-559 | 243-2,166 |
|                    |            | CATH-αβ            | 3,796 | 101-759 | 288-3,507 |
|                    |            | SCOP-primary       | 11,451| 101-1,381| 105-5,558 |
|                    |            | SCOP-α             | 1,478 | 101-938 | 147-4,082 |
|                    |            | SCOP-β             | 2,541 | 101-581 | 111-2,113 |
|                    |            | SCOP-αβ            | 3,835 | 101-904 | 105-3,966 |
|                    |            | SCOP-α+β           | 2,879 | 101-696 | 120-3,064 |
|                    |            | SCOP-multidomain   | 318   | 196-1,256| 767-5,558 |

2.3 Our graphlet PCA framework

The novelty of our new PCA framework (Section 2.3.1) comes from using graphlet features in the task of PC (Section 2.3.2; Fig. 1). Yet, our framework is generalizable, as it can use any feature(s).

2.3.1 Our PCA framework

Given a network data set and a feature set (see below), we compute one feature vector per network. We perform PCA (a standard dimension reduction technique) on the resulting feature vectors to compute principal components for each network. We pick the first k principal components, where the value of k is at least two and as low as possible so that the k components account for at least 90% of variation in the data. For every pair of networks \( N_i \) and \( N_j \), we compute their cosine similarity, \( s_{\cos}(N_i, N_j) \), based on the networks’ first k PCs. We convert the similarity into distance as \( d_{\cos}(N_i, N_j) = 1 - s_{\cos}(N_i, N_j) \). We use the PCA-based distances for classification, hypothesizing that same-class networks will be close in the PCA space while networks of different classes will be distant.

Like most of the network approaches from Fig. 1, our approach performs alignment-free network comparison, meaning that it does not need to align nodes between the compared networks before it can quantify their similarity, as alignment-based approaches do (Yaveroglu et al., 2015).

2.3.2 Our graphlet feature sets

Graphlets are small connected induced subgraphs (left side of Fig. 2). They have been proven as sensitive and superior measures of topology in numerous contexts when studying protein-protein interaction networks (Pržulj et al., 2004; Pržulj, 2007; Milenković and Pržulj, 2008; Milenković et al., 2011; Faisal and Milenković, 2014; Malod-Dognin and Pržulj, 2014). Hence, we use graphlets as PSN features for PC, as follows.

Graphlet counts. We count occurrences of each graphlet on up to \( n \) nodes in the given network. To investigate the best choice for \( n \), we use counts for 3-4-node (Fig. 2) and 3-5-node graphlets, resulting in Graphlet-3-4 and Graphlet-3-5 feature sets, respectively. Graphlet counts typically vary by orders of magnitude in real-world networks (Pržulj et al., 2004). Hence, we normalize graphlet counts by taking their logarithms. Here, we do not consider 3-node-only graphlets, because there are only two 3-node graphlets, which may not be suitable for our PCA framework, and also because using up to 4- or 5-node graphlets improves accuracy upon using only 3-node graphlets (Hulováty et al., 2014, 2013; Solava et al., 2012).

Normalization of graphlet counts. Networks with similar topology can have dissimilar graphlet counts simply because of their dissimilar network sizes (Section 3.1.2). To remove the bias of PSN size, we normalize graphlet counts by scaling them between 0 and 1. Formally, given a network, let \( g_1, g_2, \ldots, g_n \) be counts of \( n \) graphlets \( G_1, G_2, \ldots, G_n \), respectively (\( n = 8 \) for 3-4-node graphlets and \( n = 29 \) for 3-5-node graphlets). We normalize count \( g_i \) of graphlet \( G_i \) as \( g_i / \sum_{j=1}^{n} g_j \). We denote the normalized feature sets for Graphlet-3-4 and Graphlet-3-5 as \( \text{NormGraphlet-3-4} \) and \( \text{NormGraphlet-3-5} \), respectively.

Integration of graphlets with protein sequences: ordered graphlet counts. While amino acids appear in a particular order in the sequence, graphlets were originally not designed to capture this node order information. For example, nodes in graphlet \( G_1 \) can appear in three different orders (Fig. 2, top), but \( G_1 \) cannot differentiate between them. To take advantage of both network and sequence data, ordered graphlets were recently proposed (Malod-Dognin and Pržulj, 2014), which embed the relative order of nodes onto graphlets. For example, the three different orders of graphlet \( G_1 \) were formulated as three different ordered graphlets: \( O_1, O_2, \) and \( O_3 \) (Fig. 2, top). This way, Malod-Dognin and Pržulj (2014) defined all four possible 3-node ordered graphlets for all two possible 3-node “regular” (i.e., original non-ordered) graphlets.

We denote the feature set consisting of the existing four counts for 3-node ordered graphlets as \( \text{OrderedGraphlet-3} \) and our normalized counterpart of this feature set as \( \text{NormOrderedGraphlet-3} \) (normalization is done in the same way as explained above). Unlike for regular (non-ordered) graphlets, we do consider 3-node-only ordered graphlets within our PCA approach. We do this to fairly compare our alignment-free PCA approach with the existing alignment-based non-PCA approach by Malod-Dognin and Pržulj (2014) that can support only 3-node ordered graphlets.

Faisal et al.
To benefit from larger graphlets, we extend this idea to include within our PCA approach all 38 possible 4-node ordered graphlets for all six possible 4-node regular graphlets on top of the existing four 3-node ordered graphlets. We denote the resulting feature set consisting of 42 ordered graphlet counts for 3-4-node graphlets (Fig. 2) as OrderedGraphlet-3-4 and its normalized counterpart as NormOrderedGraphlet-3-4.

Inclusion of ordered graphlets on five nodes would cause the number of graphlets to grow significantly (e.g., graphlet \( G_9 \) can be formulated as 60 different ordered graphlets). Since using too many features often causes overfitting, which can eventually lead to increased misclassification (Aggarwal, 2015), we do not consider 5-node ordered graphlets.

Ordered graphlet counts do not vary by orders of magnitude in our data as regular graphlet counts do, so we do not take their logarithms.

2.4 Existing approaches

We use 15 existing network (Section 2.4.1), 3D contact (Section 2.4.2), and sequence (Section 2.4.3) approaches in the task of PC (Fig. 1).

2.4.1 Existing network approaches

Existing approaches of this type that we use for PC (not all of which were proposed for PC but can be adapted to it) can be classified into graphlet and non-graphlet approaches. None of them use PCA as we do.

Existing graphlet approaches. These include graphlet degree distribution agreement (GDDA) (Pržulj, 2007), relative graphlet frequency distance (RGFD) (Pržulj et al., 2004), graphlet correlation distance (GCD) (Yaveroglu et al., 2014), and GR-align (Malod-Dognin and Pržulj, 2014). Among them, GDDA,
RGFD, and GCD can compare any type of networks, while GR-align has been specifically designed to compare PSNs. GDDA, RGFD, and GCD are alignment-free network comparison approaches, while GR-align is an alignment-based approach. In particular, it was the GR-align study (Malod-Dognin and Pržulj, 2014) that introduced the idea of 3-node-only ordered graphlets, which we partly base our approach on.

Note that an alternative graphlet approach was used in the context of PSNs (Vacic et al., 2010), but it was used to predict functional residues (where residues correspond to nodes in PSNs) and not for PSN comparison. As such, this approach is out of the scope of our study.

**Existing non-graphlet approaches.** Several PSN measures have already been used for PC: average degree, average distance, maximum distance, average closeness centrality, average clustering coefficient, intra-hub connectivity, and assortativity (Supplementary Section S1.3) (Pabuwal and Li, 2008, 2009; Gao and Li, 2009; Emerson and Gothandam, 2012, 2013). For each measure, for each pair of networks, we compute Euclidean distance between the networks’ feature values (because all features are 1-dimensional, here we cannot use cosine similarity as for our approach).

We combine the seven measures into an eight feature set, Existing-all, to investigate whether the integration of different and complementary topological measures helps PC. We use Existing-all within our PCA framework (Section 2.3.1). This way, we can fairly compare our graphlet features and the existing non-graphlet features within the same framework.

**2.4.2 Existing 3D contact approaches** These include DaliLite (Holm and Rosenström, 2010) and TM-align (Zhang and Skolnick, 2005), both of which are alignment-based.

**2.4.3 Existing sequence approaches** Mizianty et al. (2010) used protein length and amino acid propensities to define a sequence feature set, which outperformed methodologies of 11 other studies (Mizianty et al., 2010). Thus, we use this feature set, denoted as Sequence, within our PCA framework (Section 2.3.1). This way, we can fairly compare network and sequence features within the same framework.

**2.5 Evaluation of classification accuracy**
Given a set of objects (proteins or networks) with known classification labels, for a good approach, the distance between objects of the same class should be small, while the distance between objects of different classes should be large. To evaluate this, for each approach, we compute its AUPR and AUROC scores (Supplementary Section S1.4).

### 3 RESULTS

#### 3.1 Classification of synthetic networks

Unlike for real-world PSNs, for synthetic networks, we cannot evaluate 3D contact and sequence approaches, as they require 3D contact and sequence information, respectively, which synthetic networks do not contain. Thus, we can only apply network approaches to synthetic networks, with the exception of ordered graphlet approaches, including GR-align, that require some node order, which again synthetic networks do not have. We evaluate the remaining (15) network approaches on synthetic networks. For these networks, the topology-based ground truth classification is known.

First, we evaluate the network approaches (i.e., their existing versions that are non-normalized in terms of network size; Sections 2.3.2 and 2.4.1) on synthetic networks of the same size (Section 3.1.1). Second, we evaluate whether the current non-normalized versions of the network approaches can successfully cope with synthetic networks of different sizes (Section 3.1.2). We find that our graphlet PCA approach overall outperforms the existing network approaches, including the existing graphlet (non-PCA) approaches. Therefore, in all subsequent tests, we focus on the graphlet PCA methodology. Yet, the accuracy of the graphlet PCA approach (as well as every other approach) drops when analyzing networks of different sizes compared to analyzing networks of the same size, meaning that some level of misclassification arises due to the networks having different sizes. This indicates a need for devising a normalized version of the graphlet PCA approach. Thus, third, we develop such a normalized approach (Section 3.2.3), and as we show, normalization indeed improves PC (Section 3.1.3). Forth, we summarize our key findings resulting from analyzing the synthetic network data (Section 3.1.4).

#### 3.1.1 Evaluation of non-normalized network features

Here, we evaluate non-normalized versions of our graphlet PCA approach (i.e., Graphlet-3-4 and Graphlet-3-5; Section 2.3.2), existing graphlet approaches (i.e., GDDA, RGFD, and GCD; Section 2.4.1), and existing non-graphlet approaches (i.e., average degree, average distance, maximum distance, average closeness centrality, average clustering coefficient, intra-hub connectivity, assortativity, and Existing-all; Section 2.4.1). We evaluate the approaches on synthetic network data of the same size (i.e., Synthetic-100, Synthetic-500, and Synthetic-1000; Section 2.2).

For each data set, both non-normalized versions of our graphlet PCA approach outperform the existing graphlet and non-graphlet approaches, as the former two achieve 100% accuracy (Table 2 and Supplementary Table S1). Some of the existing methods also achieve 100% accuracy on some of the data sets, but only one (RGFD) does so on all three data sets and is thus comparable to our approach. However, as we show in Section 3.2.6, RGFD loses its comparable performance in other tests.

Both our graphlet PCA approach and the existing graphlet (non-PCA) approaches outperform the existing non-graphlet approaches (Table 2 and Supplementary Table S1). This confirms the power of the local graphlets over the global network measures that have traditionally been used for PC.

Combining the seven existing non-graphlet features into Existing-all and using Existing-all in our PCA framework improves the accuracy of each individual non-graphlet feature. This confirms that feature integration helps, which is why we have developed our framework to allow for this in the first place. Also, Existing-all is comparable to our graphlet PCA approach (Table 2 and Supplementary Table S1). However, as we show in Section 3.1.3, Existing-all loses its comparable performance in other tests.

#### 3.1.2 Network size affects classification via non-normalized features

To test whether the non-normalized versions of our graphlet PCA approach, existing graphlet approaches, and existing non-graphlet approaches, all of which are non-normalized, are robust to the size of...
Table 2. Classification accuracy with respect to AUPRs (expressed as percentages) on synthetic networks. Results for non-normalized approaches are highlighted in 1) light gray for network data of the same size (Section 3.1.1) and 2) dark gray for network data of different sizes (Section 3.1.2). Results for normalized approaches (Section 3.1.3) are not highlighted. Given a network data set (within a column), the AUPR of the best approach is shown in bold. For equivalent results with respect to AUROCs, see Supplementary Table S1.

| Approach          | Synthetic-100 | Synthetic-500 | Synthetic-1000 | Synthetic-All |
|-------------------|---------------|---------------|----------------|---------------|
| Graphlet-3-4      | 100.00        | 100.00        | 100.00         | 81.76         |
| Graphlet-3-5      | 100.00        | 100.00        | 100.00         | 83.28         |
| NormGraphlet-3-4  | 100.00        | 100.00        | 100.00         | 94.37         |
| NormGraphlet-3-5  | 100.00        | 100.00        | 100.00         | 99.86         |
| GDEA              | 97.36         | 100.00        | 99.99          | 91.46         |
| RGFD              | 100.00        | 100.00        | 100.00         | 98.55         |
| GCD               | 89.26         | 100.00        | 100.00         | 86.27         |
| Average degree    | 79.76         | 79.76         | 79.76          | 68.77         |
| Average distance  | 82.47         | 98.12         | 99.60          | 57.10         |
| Maximum distance  | 68.82         | 84.32         | 93.08          | 46.11         |
| Average closeness centrality | 86.10 | 88.46 | 85.33 | 48.41 |
| Average clustering coefficient | 98.93 | 99.68 | 99.25 | 79.37 |
| Intra-hub connectivity | 70.88 | 69.11 | 69.31 | 66.61 |
| Assortativity     | 82.79         | 92.27         | 91.73          | 81.98         |
| Existing-all      | 100.00        | 100.00        | 100.00         | 85.92         |

3.1.3 Normalization of graphlet features improves classification. Motivated by this network size-related bias of all non-normalized network features, we propose a normalized version of the best of all features, namely our graphlet PCA features (Section 2.3.2). We validate our normalized graphlet PCA features as follows. When we apply them to Synthetic-100, Synthetic-500, and Synthetic-1000, we hope to preserve the maximum (100%) accuracy for the three network data sets of the same size while improving the accuracy for Synthetic-all that contains networks of different sizes. Indeed, this is exactly what we observe (Table 2 and Supplementary Table S1). Now the best of our graphlet PCA approaches (i.e., NormGraphlet-3-5) outperforms each of the three existing graphlet (non-PCA) approaches, even though all of these approaches are based on graphlets. This shows the usefulness of our PCA framework as a whole over the existing graphlet methodologies. Also, now NormGraphlet-3-5 outperforms the non-graphlet Existing-all approach under the same PCA framework, which confirms the power of graphlets.

3.1.4 Summary of results for synthetic networks. Our (non-normalized) graphlet PCA features overall outperform the existing graphlet (non-PCA) approaches, which in turn outperform the existing non-graphlet approaches. Our normalized graphlet PCA features further improve upon their non-normalized counterparts (and thus upon the existing approaches). NormGraphlet-3-5 is the most accurate approach.

3.2 Classification of PSNs

In our analysis of real-world PSNs, for which CATH- or SCOP-label-based (rather than topology-based as in Section 3.1) ground truth classification is known, first, we evaluate the considered approaches (i.e., their existing versions that are non-normalized in terms of network size; Sections 2.3.2, 2.4.1, 2.4.2, and 2.4.3) on PSNs of the same size (Section 3.2.1). Second, we test the approaches on PSNs of different sizes (Section 3.2.2). In both tests, overall, our graphlet PCA approach is superior to the existing approaches. Yet, the accuracy of all approaches drops when analyzing PSNs of different sizes compared to analyzing PSNs of the same size. Therefore, third, we test whether graphlet normalization can improve PC. Indeed, this is what we observe (Section 3.2.3). Fourth, in order to investigate whether the integration of network topology with protein sequences can improve PC, we test our ordered graphlet PCA approach (Section 3.2.4). Fifth, we compare the considered approaches in terms of their running times (Section 3.2.5). Sixth, we summarize our key findings resulting from analyzing the PSN data (Section 3.2.6).

3.2.1 Evaluation of non-normalized features. Here, we benchmark the non-normalized versions of our PCA graphlet approach, existing graphlet (non-PCA) approaches, existing non-graphlet approaches, and 3D contact approaches on all PSN data sets for which networks within the given set are of the same size, i.e., on CATH-95, CATH-99, and CATH-251-265 (Section 2.2).

For each PSN set, just as for the synthetic networks, the non-normalized versions of our graphlet PCA approach (Graphlet-3-4 and Graphlet-3-5) are superior to the existing graphlet, non-graphlet, and 3D contact approaches, except one (RGFD) that is comparable to our approach (Table 3 and Supplementary Tables S2-S3). Yet, as we show below, RGFD loses its comparable performance in other tests. Again,
combining the seven existing non-graphlet measures into Existing-all typically improves the accuracy of each individual feature (Table 3). Existing-all is comparable to the two non-normalized versions of our graphlet PCA approach. However, as we show below, Existing-all loses its comparable performance in other tests.

3.2.2 Network size affects classification via non-normalized features. Next, we evaluate the same non-normalized approaches as in Section 3.2.1 on all 10 sets of PSNs of different sizes (i.e., CATH-primary, CATH-α, CATH-β, CATH-α/β, SCOP-primary, SCOP-α, SCOP-β, SCOP-α/β, SCOP-multidomain; Section 2.2).

We again observe a decline in accuracy for each approach, which confirms the bias of network size. Nonetheless, the non-normalized versions of our graphlet PCA approach remain superior or comparable to all existing methods (Table 3 and Supplementary Tables S2-S3).

3.2.3 Normalization of graphlet features improves classification. Motivated by the above network size-related bias of all considered non-normalized approaches, we propose a normalized version of the best of those approaches, namely of the graphlet PCA features. When we apply each of the normalized features to the PSN data, we hope to ideally improve or at least preserve the accuracy on the PSN data sets of the same network size (i.e., CATH-95, CATH-99, and CATH-251-265) while improving the accuracy for the 10 sets of PSNs of different sizes, compared to the accuracy of the features’ non-normalized counterparts. Indeed, this is exactly what we observe (Table 3 and Supplementary Tables S2-S3).

3.2.4 Integration of network and sequence data via ordered graphlets. The versions of our PCA approach that are based on regular (non-ordered, as considered thus far) graphlets, already perform much better than the sequence approach (Table 3 and Supplementary Tables S2-S3). Integration of network data with sequence data may further improve the accuracy compared to only network and only sequence approaches. We test this by using ordered graphlets to impose the sequence-based order of amino acids onto nodes in regular graphlets (Fig. 2 and Section 2.3.2).

Considering only non-normalized graphlet features, ordered graphlets (i.e., OrderedGraphlet-3 and OrderedGraphlet-3-4) improve upon their regular graphlet counterparts for PSNs of the same size as well as of different sizes (Table 3). Considering also normalized graphlet features, NormOrderedGraphlet-3 to 4 leads to better accuracy compared to its non-normalized counterpart, though NormOrderedGraphlet-3 does not improve upon its non-normalized counterpart (Table 3).

The fact that within our PCA framework ordered graphlets beat regular graphlets alone as well as the sequence approach alone confirms that PSN data and sequence data are complementary and should thus be integrated. We consider this to be one of our key contributions.

Table 3. Classification accuracy with respect to AUPRs (expressed as percentages) on real-world PSN data sets of the same network size as well as of different network sizes. For the latter, the PSN data sets corresponding to the top hierarchical classes of CATH and SCOP are shown. For equivalent results for other levels of CATH and SCOP hierarchy, see Supplementary Table S2. Results for non-normalized approaches are highlighted in 1) light gray for network data of the same size (Section 3.2.1) and 2) dark gray for network data of different sizes (Section 3.2.2). Results for normalized approaches (Section 3.2.3) are not highlighted. Given a network data set (within a given column), the AUPR of the best approach is shown in bold. For equivalent results with respect to AUROCs, see Supplementary Table S3.

| Approach                        | PSNs of the same size | PSNs of different sizes |
|---------------------------------|-----------------------|-------------------------|
|                                 | CATH-95  | CATH-99  | CATH-251-265 | CATH- primary | SCOP- primary |
| Graphlet-3to4                   | 82.28    | 92.05    | 92.35        | 47.47        | 37.85        |
| Graphlet-3to5                   | 83.31    | 92.78    | 92.89        | 47.86        | 37.29        |
| OrderedGraphlet-3               | 90.99    | 95.93    | 91.02        | 45.75        | 35.33        |
| OrderedGraphlet-3to4            | 96.69    | 91.88    | 97.20        | 46.16        | 33.60        |
| NormGraphlet-3to4               | 96.03    | 100.00   | 95.28        | 52.57        | 36.89        |
| NormGraphlet-3to5               | 94.11    | 99.73    | 97.67        | 53.24        | 37.51        |
| NormOrderedGraphlet-3           | 83.54    | 96.49    | 93.51        | 54.20        | 33.36        |
| NormOrderedGraphlet-3to4        | 97.59    | 96.58    | 98.74        | 65.66        | 44.50        |
| GDDA                            | 77.65    | 80.78    | 71.46        | 42.77        | 32.09        |
| RGFD                            | 87.87    | 89.49    | 94.00        | 53.75        | 38.86        |
| GCD                             | 71.70    | 74.92    | 77.23        | 42.61        | 31.78        |
| GR-align                        | 76.25    | 65.03    | 70.25        | 39.07        | 28.55        |
| Average degree                  | 48.22    | 50.21    | 61.22        | 42.81        | 30.85        |
| Average distance                | 64.49    | 60.22    | 51.60        | 35.13        | 29.75        |
| Maximum distance                | 62.39    | 73.49    | 54.89        | 35.07        | 29.67        |
| Average closeness centrality    | 62.73    | 60.94    | 45.73        | 34.67        | 27.84        |
| Average clustering coefficient  | 87.01    | 72.10    | 89.96        | 44.15        | 30.77        |
| Intra-hub connectivity          | 54.94    | 72.34    | 63.76        | 34.79        | 29.21        |
| Assortativity                   | 76.97    | 85.34    | 93.31        | 37.41        | 26.99        |
| Existing-all                    | 82.14    | 91.56    | 92.48        | 47.69        | 35.68        |
| TM-align                        | 53.38    | 69.12    | 58.96        | 39.84        | 31.64        |
| DaliLite                        | 50.93    | 62.02    | 45.79        | 37.30        | 25.78        |
| Sequence                        | 70.23    | 62.14    | 54.48        | 40.45        | 29.02        |
Here, we note that 3-node-only ordered graphlets were used for protein 3D structural alignment within the GR-align approach. We adopt the existing idea of 3-node ordered graphlets but we do so within our new alignment-free PCA framework as opposed to the existing alignment-based GR-align approach. Also, we extend this idea into larger, 3-4-node ordered graphlets. Importantly, when we consider 3-node-only ordered graphlets within our PCA framework, which makes the comparison with GR-align as fair as possible, our PCA approach is superior to GR-align (Table 3). This is further supported when we compare the accuracy rankings of the different methods over all PSN sets (Table 4). When we also consider larger ordered graphlets, this further improves the performance of our PCA approach (Table 3). GR-align is also slower than our approach. For example, it is 25 times slower than the fairly comparable 3-node-only ordered graphlet version of our PCA approach (Section 3.2.5 and Table 4). These results validate our PCA framework as a whole. Note that when one’s goal is not just to quantify the level of similarity between networks but also to map (align) nodes between the networks, using an alignment-free approach such as our graphlet PCA framework is inappropriate, and instead, an alignment-based approach such as GR-align needs to be used. For further discussion on alignment-free versus alignment-based approaches, see Yaveroglu et al. (2015).

Another of our key contributions is that even our regular graphlet PCA approaches and especially their normalized and ordered counterparts are superior to traditional 3D contact approaches, despite the fact that both types of approaches (network vs. 3D contact) use 3D structural information. This highlights the usefulness of network analyses of protein structures. This is especially true given that our network approaches are also faster than the 3D contact approaches, as we discuss next.

### 3.2.5 Running time comparison

All alignment-free network approaches are comparable in terms of running time to each other as well as to the sequence approach, they are followed by the only alignment-based network approach (GR-align), and all of them are significantly faster than the 3D contact approaches (Table 4).

### 3.2.6 Summary of results for PSNs

The non-normalized versions of our graphlet PCA approach are superior to the existing graphlet (non-PCA), non-graphlet, 3D contact, and sequence approaches. By normalizing the graphlet features, we improve upon the non-normalized features and thus upon the existing methods (Table 3 and Supplementary Tables S2-S3). By imposing sequence order onto nodes via ordered graphlets, we further improve the accuracy. NormOrderedGraphlet-3-4 is superior to all considered methods in terms of its accuracy ranking over all considered PSN sets, and its ranking is statistically significantly better than the ranking of any other method (Table 4). This further validates our graphlet PCA framework for PC.

### Table 4.

Summary of classification accuracy and running time of the considered approaches. The 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 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 over each of the other approaches in terms of their ranks, using paired t-test. The running times of the approaches are shown when classifying proteins from the CATH-α set. The running times for the other data sets are qualitatively the same. For visual representation of the results, see Supplementary Fig. S1 and S2.

| Approach                  | AUPR Rank | AUPR p-value | AUROC Rank | AUROC p-value | Running time (hrs) |
|---------------------------|-----------|--------------|------------|---------------|-------------------|
| Graphlet-3-4              | 7.4       | 0.011        | 9.5        | 0.006         | 0.43              |
| Graphlet-3-5              | 7.9       | 0.001        | 9.5        | 0.004         | 0.49              |
| OrderedGraphlet-3         | 6.2       | 0.05         | 9.1        | 0.016         | 0.38              |
| OrderedGraphlet-3-4       | 6.5       | 0.025        | 7.8        | 0.037         | 2.39              |
| NormGraphlet-3-4          | 6.8       | 0.005        | 7.2        | 0.011         | 0.44              |
| NormGraphlet-3-5          | 7.2       | 0.003        | 5.8        | 0.087         | 0.51              |
| NormOrderedGraphlet-3     | 9.5       | 1e-04        | 9.0        | 0.002         | 0.39              |
| NormOrderedGraphlet-3-4   | 3.3       | -            | 3.9        | -             | 2.41              |
| GDDA                      | 16.3      | 2e-07        | 16.7       | 6e-07         | 0.54              |
| RGFID                     | 8.5       | 3e-06        | 8.8        | 9e-04         | 0.49              |
| GCD                       | 16.1      | 4e-09        | 16.1       | 1e-06         | 1.32              |
| GR-align                  | 7.5       | 0.069        | 8.8        | 0.057         | 9.49              |
| Average degree            | 18.0      | 4e-09        | 15.2       | 1e-05         | 0.39              |
| Average distance          | 14.4      | 5e-05        | 15.5       | 2e-04         | 0.48              |
| Maximum distance          | 16.3      | 3e-07        | 15.8       | 4e-06         | 0.49              |
| Average closeness centrality | 17.5    | 2e-06        | 15.5       | 1e-04         | 0.48              |
| Average clustering coefficient | 15.8   | 4e-08        | 13.6       | 6e-06         | 0.56              |
| Intra-hub connectivity     | 15.4      | 6e-06        | 14.1       | 2e-04         | 0.64              |
| Assortativity             | 19.1      | 3e-08        | 18.2       | 1e-06         | 0.46              |
| Existing-all              | 10.0      | 3e-04        | 9.0        | 7e-03         | 1.01              |
| DaliLite                  | 11.7      | 2e-03        | 9.8        | 0.023         | 2021.41            |
| TM-align                  | 21.0      | 9e-10        | 21.3       | 4e-09         | 168.32             |
| Sequence                  | 13.5      | 5e-05        | 15.6       | 1e-05         | 0.24              |
3.3 Application of graphlet PCA features in revealing biochemically interesting PSN patterns

We aim to identify graphlet patterns that lead to successful distinction of different CATH or SCOP classes from the PSN data, focusing as an illustration on the PSN sets containing networks of the same size (i.e., on CATH-95, CATH-99, and CATH-251-265) from $\alpha$ or $\beta$ protein domain classes. Such graphlets that are significantly (Mann-Whitney U test; $p < 0.05$) represented in $\alpha$ but not in $\beta$, or vice versa, could be linked to the functionality of the given domain class.

For the 3-5-node regular graphlet feature set (i.e., Graphlet-3-5), graphlets represented in $\alpha$ tend to be denser than graphlets represented in $\beta$ (Fig. 3). For example, all of the complete (clique) graphlets (i.e., $G_2$, $G_5$, $G_{29}$, which are the densest graphlets) are represented in $\alpha$. On the other hand, all of the path-like graphlets (i.e., $G_1$, $G_3$, $G_9$, which are the sparsest graphlets) are represented in $\beta$.

For the 3-4-node ordered graphlet feature set (i.e., OrderedGraphlet-3-4), in ordered graphlets represented in $\alpha$ (e.g., $O_1$), there is typically a node order-respecting path through the graphlet, unlike in most of ordered graphlets represented in $\beta$ (e.g., $O_2$ and $O_3$) (Supplementary Fig. S3).

Linking the identified domain class-specific PSN patterns to their potential biochemical meaning is our future interest.

---

4 CONCLUSIONS

We present a general computational framework for network classification, which can use any feature(s). We demonstrate the effectiveness of our framework in the context of PC, in particular the power of using graphlets as state-of-the-art network features. Specifically, we use ordered graphlets to integrate via network analysis complementary protein 3D structural data and sequence data, which improves upon the existing network (graphlet or non-graphlet), 3D contact, and sequence approaches. In the process, we address the network size bias of the existing approaches.

---

FUNDING

This work was supported by the National Science Foundation (CAREER CCF-1452795 and CCF-1319469), Air Force Young Investigator Program (AFOSR YIP), National Institutes of Health (1R21AI111286-01A1 and R01 GM074807), and Clare Boothe Luce Graduate Research Fellowship.

REFERENCES

Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.
Ashburner, M., et al. (2000). Gene ontology: tool for the unification of biology. Nature Genetics, 25(1), 25–29.
Berman, H. M., et al. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235–242.
Burmann, B. M., et al. (2012). An $\alpha$ helix to $\beta$ barrel domain switch transforms the transcription factor RfaH into a translation factor. Cell, 150(2), 291–303.
Clarke, T. F. and Clark, P. L. (2008). Rare codons cluster. PLoS One, 3(10), e3412.
Emerson, I. A. and Gotthandam, K. M. (2012). Network analysis of transmembrane protein structures. Physica A, 391, 905–916.
Emerson, I. A. and Gotthandam, K. M. (2013). Residue centrality in alpha helical polytopic transmembrane protein structures. Journal of Theoretical Biology, 309, 78–87.
Faisal, F. E. and Milenković, T. (2014). Dynamic networks reveal key players in aging. Bioinformatics, 30(12), 1721–1729.
Gao, J. and Li, Z. (2009). Conserved network properties of helical membrane protein structures and its implication for improving membrane protein homology modeling at the twilight zone. Journal of Computer-Aided Molecular Design, 23, 755–763.
Network-based integrative protein classification

Hulovatyy, Y., et al. (2014). Revealing missing parts of the interactome via link prediction. PLOS ONE, 9(3), e90073.

Kosloff, M. and Kolodny, R. (2008). Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins, 71(2), 891–902.

Kuchaiev, O., et al. (2011). GraphCrunch 2: Software tool for network modeling, alignment and clustering. BMC Bioinformatics, 12(24).

Lee, D., et al. (2007). Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology, 8(12), 995–1005.

Malod-Dognin, N. and Pržulj, N. (2014). GR-Align: fast and flexible alignment of protein 3D structures using graphlet degree similarity. Bioinformatics, 30(9), 1259–65.

Milenković, T., and Pržulj, N. (2008). Uncovering biological network function via graphlet degree signatures. Cancer Informatics, 6, 257–273.

Milenković, T., et al. (2008). GraphCrunch: a tool for large network analyses. BMC Bioinformatics, 9(70).

Milenković, T., et al. (2009). Optimized null model for protein structure networks. PLoS ONE, 4(6), e5967.

Milenković, T., et al. (2011). Dominating biological networks. PLoS ONE, 6(8), e23016.

Mizianty, M. J., et al. (2010). Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics, 26(18), i489–i496.

Muzzin, A. G., et al. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, 247(4), 536–540.

Orengo, C. A., et al. (1999). The CATH database provides insights into protein structure/function relationships. Nucleic Acids Research, 27(1), 275–279.

Pabuwal, V. and Li, Z. (2008). Network pattern of residue packing in helical membrane proteins and its application in membrane protein structure prediction. Protein Engineering, Design and Selection, 21(1), 55–64.

Pabuwal, V. and Li, Z. (2009). Comparative analysis of the packing topology of structurally important residues in helical membrane and soluble proteins. Protein Engineering, Design and Selection, 22(2), 67–73.

Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. Bioinformatics, 23, e177–e183.

Pržulj, N., et al. (2004). Modeling interactome: Scale-free or geometric? Bioinformatics, 20(18), 3508–3515.

Silinove, I., et al. (2015). CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Research, 43(D1), D376–D381.

Solava, R., et al. (2012). Graphlet-based edge clustering reveals pathogen-interacting proteins. Bioinformatics, 18(28), i480–i486.

Sułkowska, J. I., et al. (2012). Genomics-aided structure prediction. Proceedings of the National Academy of Sciences, 109(26), 10340–45.

Tuinstra, R. L., et al. (2008). Interconversion between two unrelated protein folds in the lymphotactin native state. Proceedings of the National Academy of Sciences, 105(13), 5057–62.

Vacic, V., et al. (2010). Graphlet Kernels for Prediction of Functional Residues in Protein Structures. Journal of Computational Biology, 17(1), 55–72.

Yaveroglu, O. N., et al. (2014). Revealing the Hidden Language of Complex Networks. Scientific Reports, 4, 4547.

Yaveroglu, O. N., et al. (2015). Proper evaluation of alignment-free network comparison methods. Bioinformatics, 31(16), 2697–2704.

Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research, 33(7), 2302–09.