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).

**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).

**Results**

- **PSN embedding** into **PCA space** may improve upon traditional **3D contact approaches**.
- Amino acids that are distant in the sequence can be close in the **3-dimensional (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.

**References**

[3] Faisal et al. (2017), Scientific Reports, 7, 14890. (This work.)