Studying reuse patterns in the protein universe

We study the global nature of protein space, and study reuse patterns within it. To do so, we represent all similarities among a set of representative structures as a graph where edges connect proteins that share significantly sized segments of similar sequence and structure. This graph offers a way to organize protein space, and examine how the definition of “evolutionary relatedness” influences it. At excessively strict thresholds the graph “falls apart”; for very lax thresholds, there are paths between virtually all nodes. Interestingly, at intermediate thresholds the graph has two regions: "discrete" versus “continuous.” The discrete region consists of isolated islands, each generally corresponding to a fold; the continuous region is dominated by domains with alternating alpha and beta elements.

Considering such a graph for two representative sets of ECOD domains and PDB chains, we study reuse patterns in protein space. Reuse (described by the edges in the graph) has a clear evolutionary advantage over 'design from scratch', where most newly-formed segments are not even foldable. The best characterized form of sequence reuse is structural domains, where the shared parts are of 100 amino acids on average. To systematically explore reuse in proteins, we develop a DP algorithm that derives the most reused non-overlapping segments of a domain/protein from the set of its alignments to other domains/proteins. This allows us to automatically identify shared ‘themes’, segments of 35 residues or more that are similar in sequence and structure. We show that reuse prevails at all levels, and that it increases with the decrease in length of the themes. In this respect, domains are just one of many forms of reuse in proteins, i.e., a special case of themes. The observed behavior is consistent with evolution by divergence, duplication, and mutations, consolidating the suggestion that proteins have evolved from ancestral amino acid segments. Indeed, some of our themes could be the descendants of these ancestral segments.

I'll be describing new work with Sergey Nepomnyachiy and Nir Ben-Tal (to be published still), but also mention older projects: CyToStruct: augmenting the network visualization of cytoscape with the power of molecular viewers, Nepomnyachiy, Ben-Tal, Kolodny, Structure (2015) and Global view of the protein universe, Nepomnyachiy, Ben-Tal,Kolodny, PNAS (2014).