Supplemented Information for “A global view of the protein universe”

Figure S8 highlights the locations of domains from selected SCOP folds in the domains network of Figure 2. As expected, the members of each fold are closely related to each other, which is reflected in their close proximity in the network. For the most part, folds of the all-alpha, all-beta and alpha+beta SCOP classes appear in the discrete region of the network, each corresponding to its own isolated ‘island’. For example, the globin-like (all-alpha), cupredoxin-like (all-beta), and the DNA clamp (alpha+beta). The immunoglobulin-like beta-sandwich (all-beta) is an exception: minority representatives of this fold appear in an isolated island but the majority is found in the major connected component, albeit somewhat isolated from the rest of the domains.

The major connected component of the network comprises, for the most part, of domains of the alpha/beta SCOP class, such as the NAD(P)-binding Rossman fold, FAD/NAD(P)-binding, TIM alpha/beta barrels, S-adenosyl-L-methionine-dependent methyltransferases and P-loop containing nucleoside triphosphate hydrolases. Typically, the domains of each fold are found in close vicinity to each other in the major component. An exception here is the TIM alpha/beta barrel fold, members of which are found in a number of neighborhoods in the major connected component. A key observation here is that cross-fold similarity is very common in protein space; in particular among domains from the alpha/beta SCOP class.

Figures

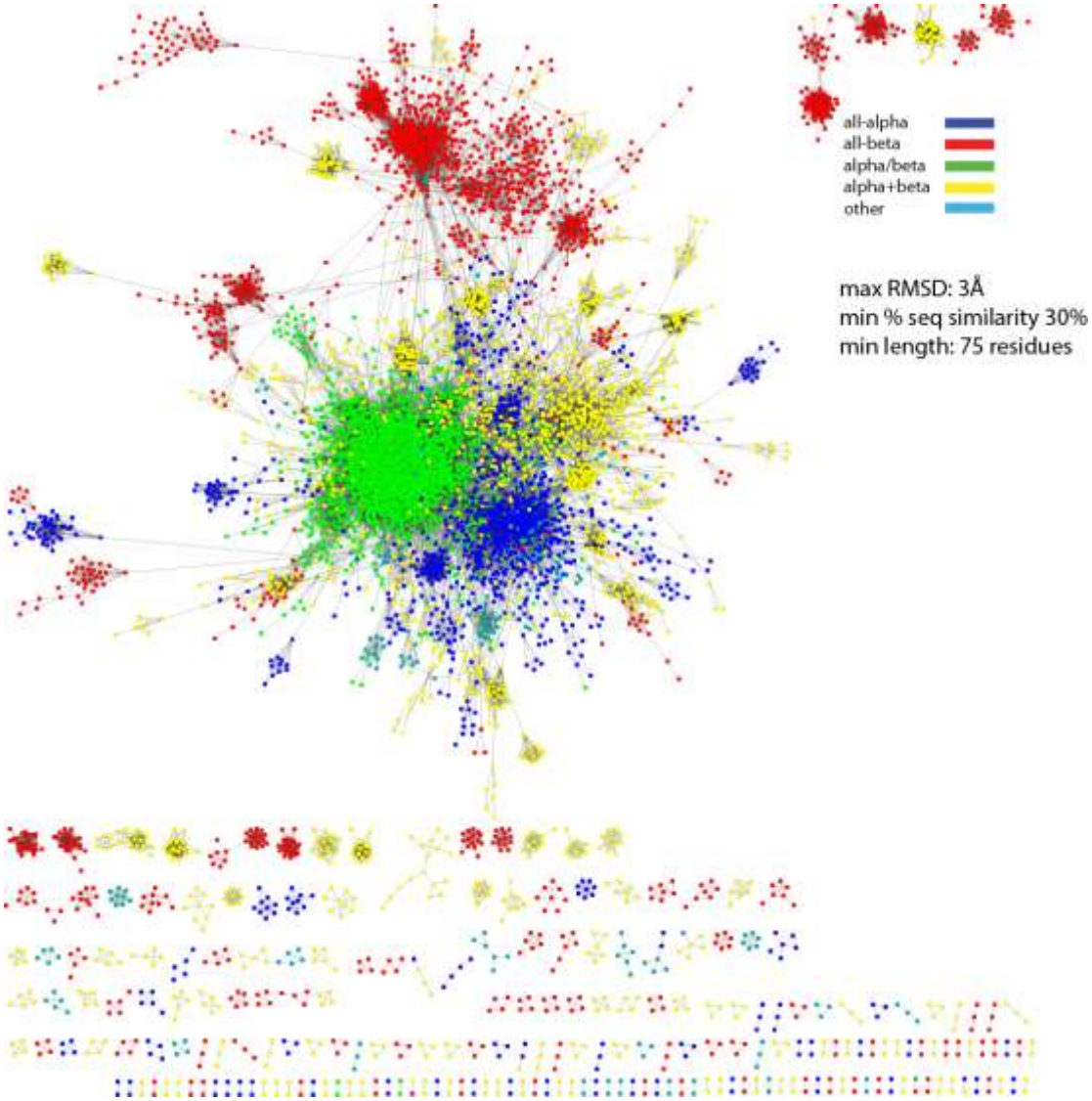

Figure S1: The domain network of protein space, when we connect two domains if they share at least 75 residues with at most 3Å RMSD and at least 30% sequence similarity. The domains are colored by their SCOP class. The network includes 7,982 domains, the remaining 1,728 singletons are omitted for clarity. There are 6,781 domains (i.e., 70% of the dataset) in the largest connected component.

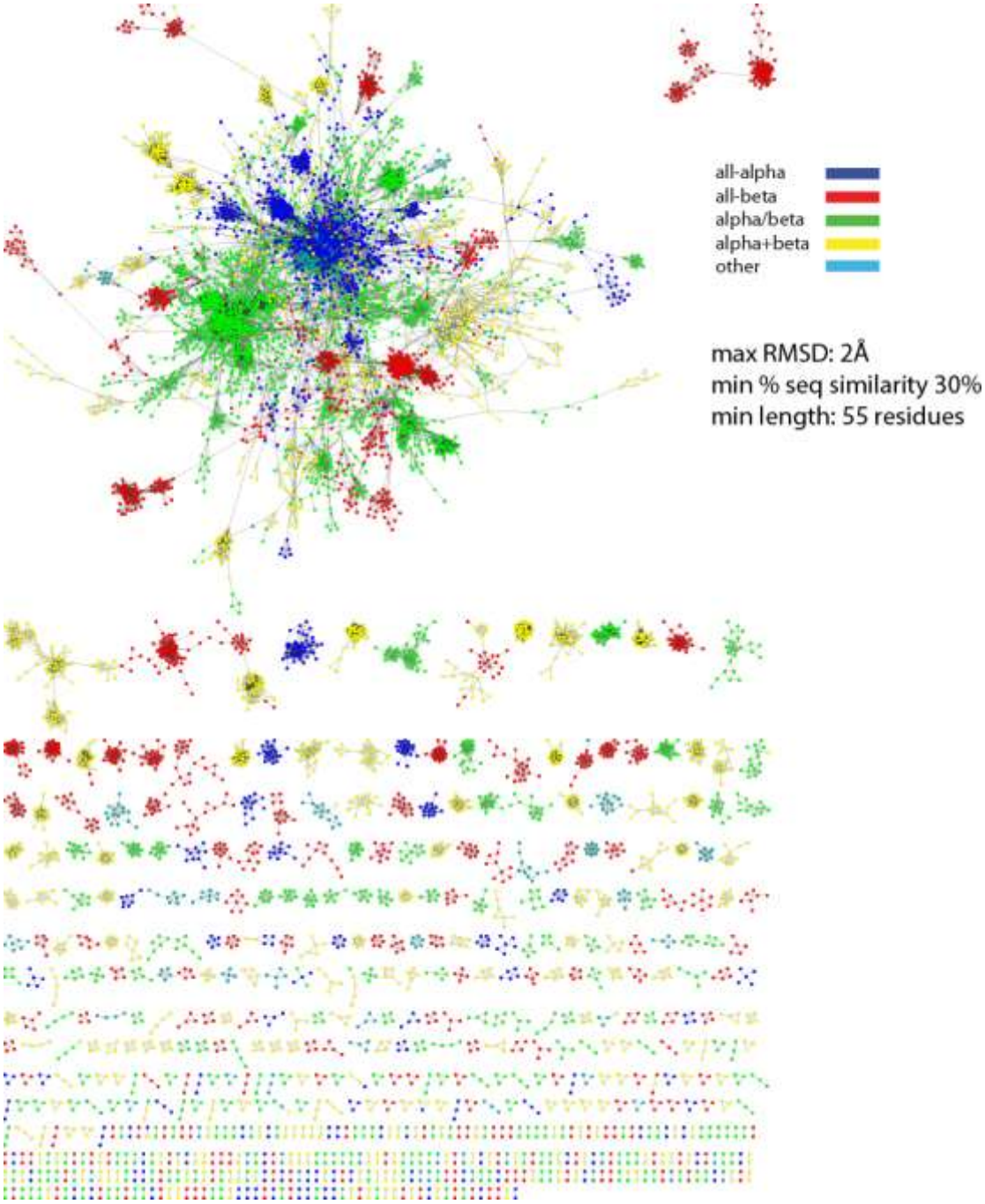

Figure S2: The domain network of protein space, using more strict conditions to relate two domains: if they share at least 55 residues with at most 2Å RMSD and at least 30% sequence similarity. The domains are colored by their SCOP class. The network includes 7,659 domains, the remaining 2,051 singletons are omitted for clarity. There are 4,240 domains in the largest connected component.

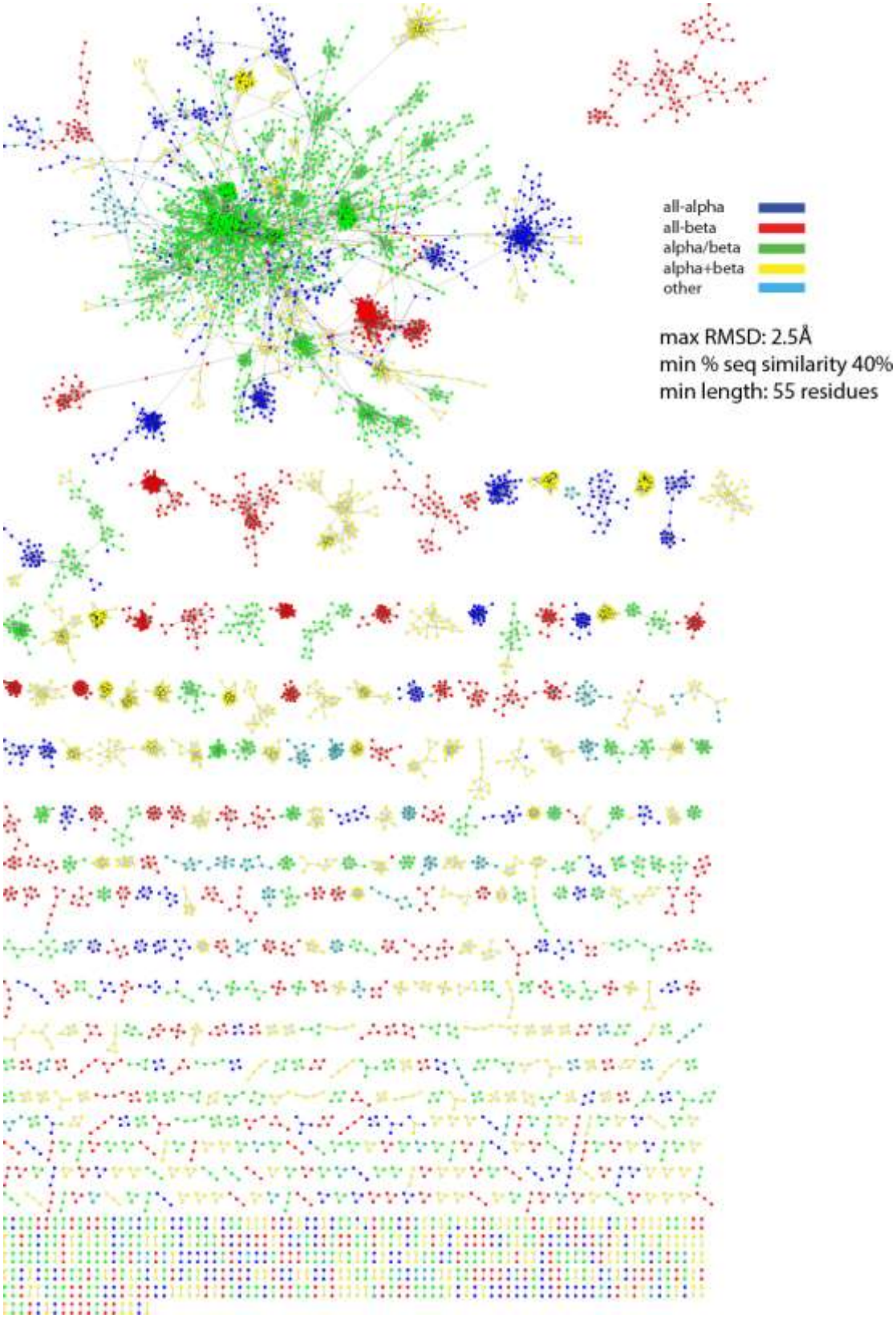

Figure S3: The domain network of protein space, using even more strict conditions to relate two domains: if they share at least 55 residues with at most 2.5Å RMSD and at least 40% sequence similarity. The domains are colored by their SCOP class. The network includes 7,058 domains, the remaining 2,652 singletons are omitted for clarity. There are 2,642 domains in the largest connected component.

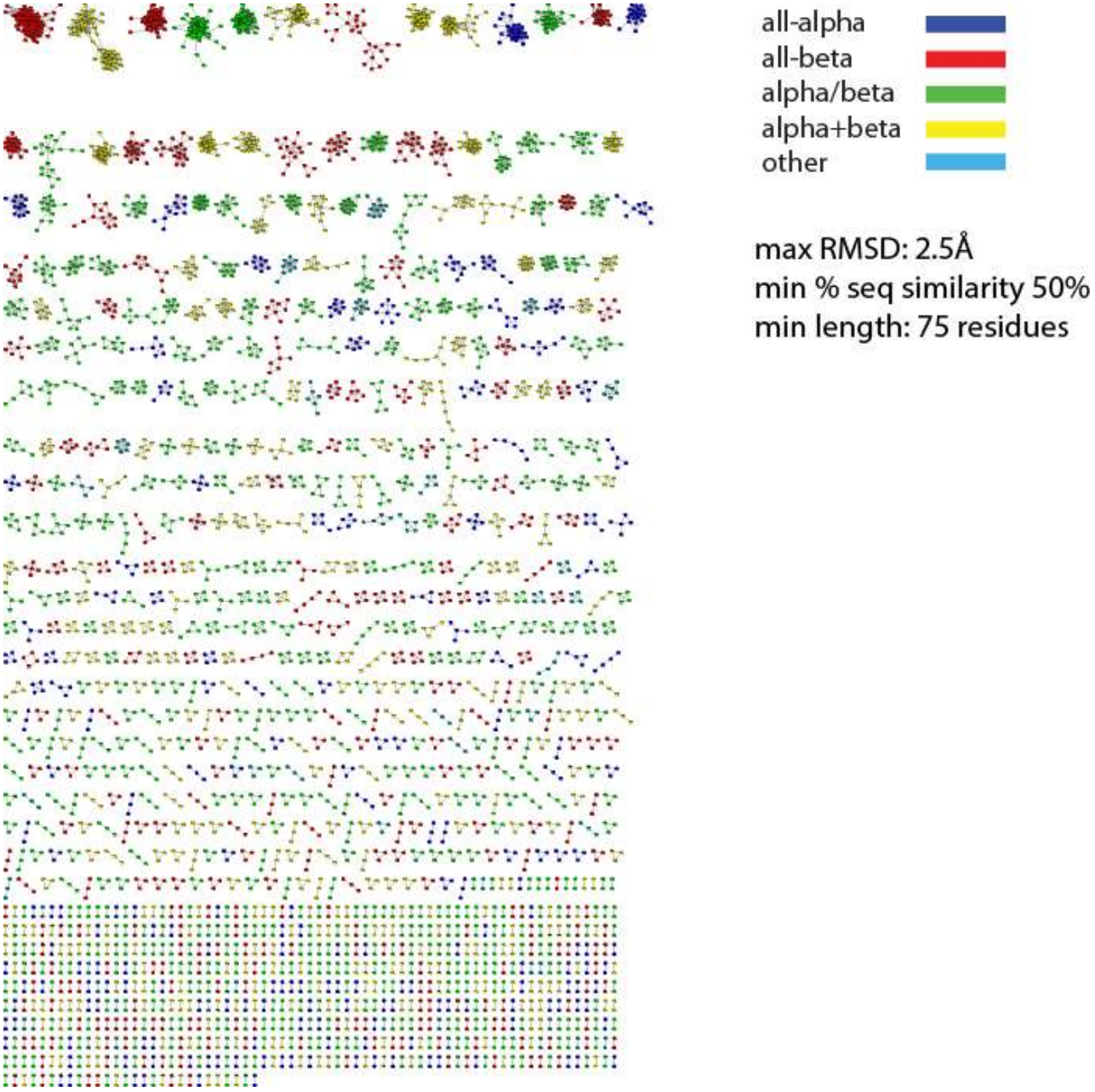

Figure S4: The domain network of protein space, using even more strict conditions to relate two domains: if they share at least 75 residues with at most 2.5Å RMSD and at least 50% sequence similarity. The domains are colored by their SCOP class. The network includes 4,509 domains, the remaining 5,201 singletons are omitted for clarity. We see that the network disintegrated into small connected components (the largest has 99 domains)

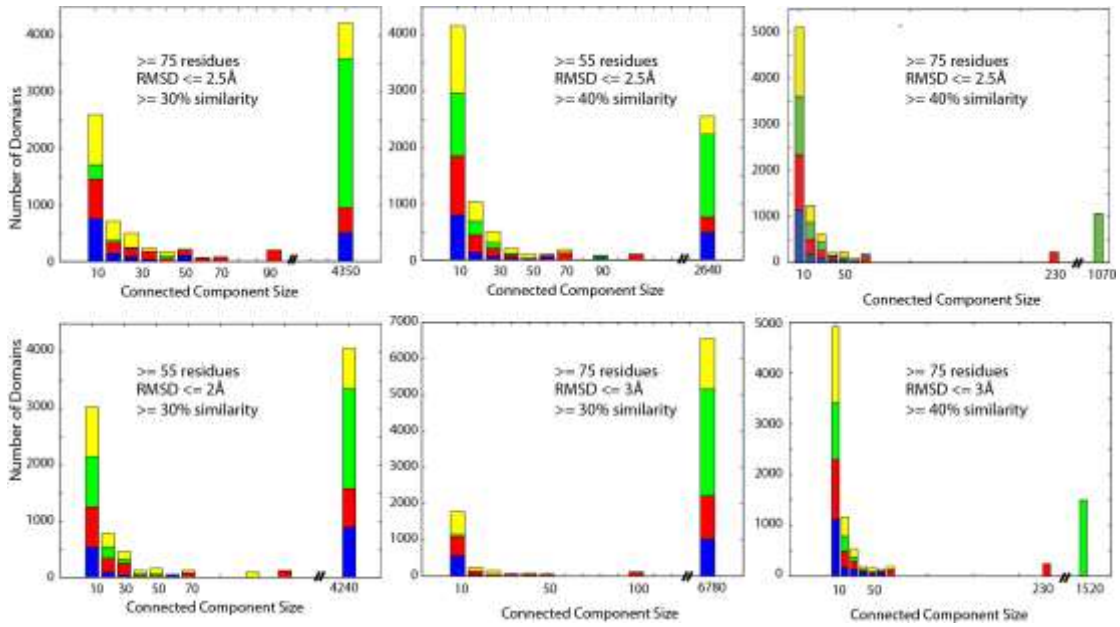

Figure S5: Distribution of cluster sizes, colored by the SCOP class of the domains, for domain networks using different size, RMSD, and percent sequence similarity thresholds. As expected, more strict thresholds results in a more disconnected network, placing more domains in smaller connected components. However, in all the cases shown here, we see that the single, largest, connected component is dominated by the (green) alpha/beta domains.

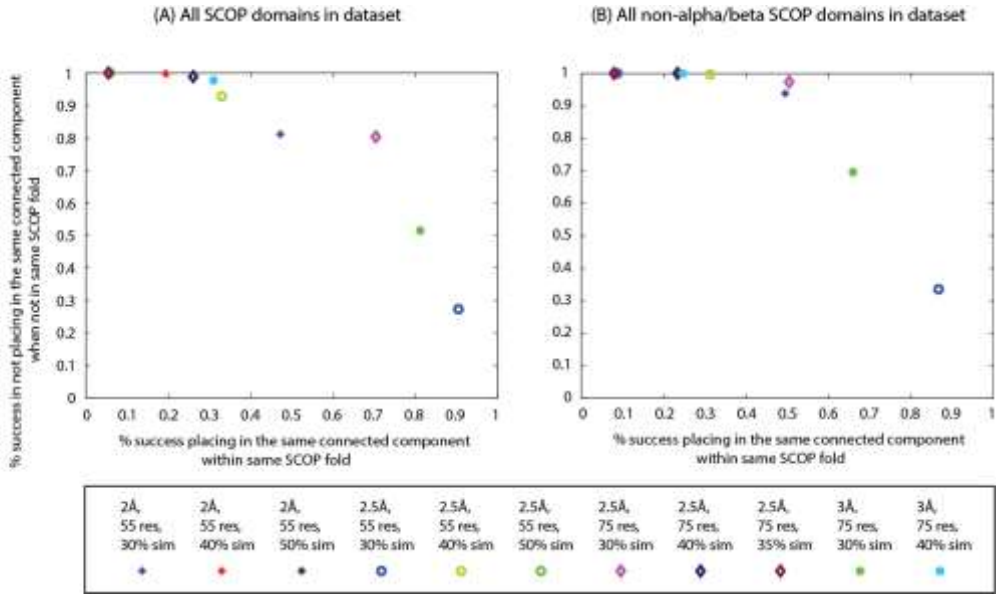

Figure S6: The percent success in placing domain pairs with the same SCOP fold in the same connected component vs. the percent success in not placing pairs with different SCOP folds in the same connected component. The left panel describes these values for various threshold combinations and all the domains in the dataset, the right panel for the 61% of the domains that are non-alpha/beta.

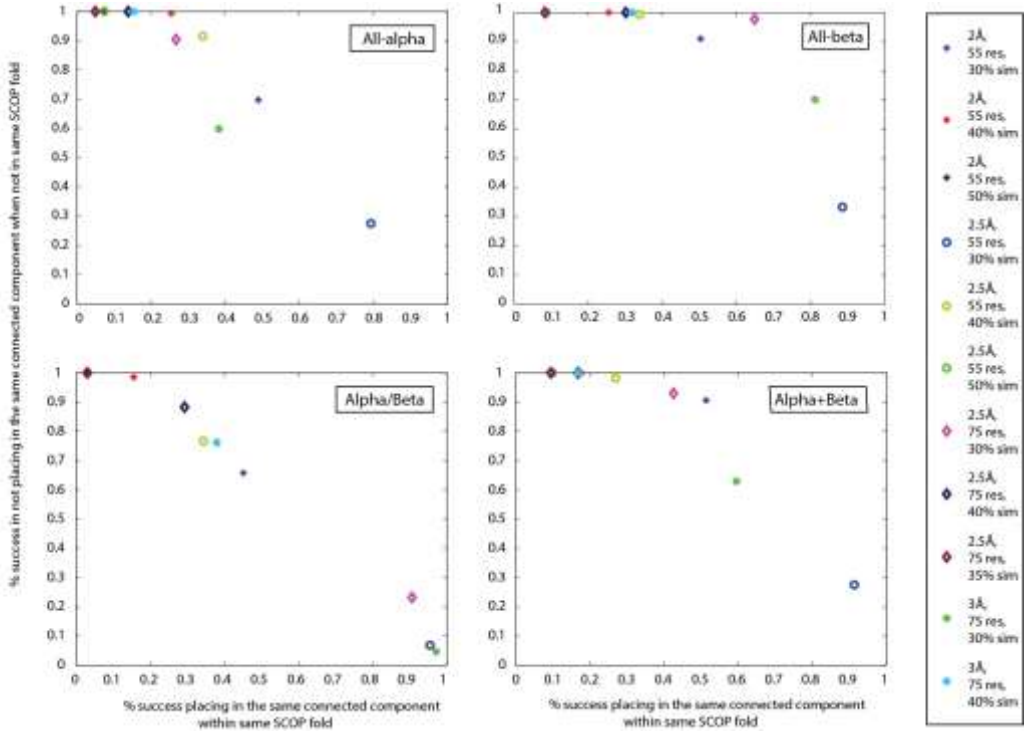

Figure S7: The percent success in placing domain pairs with the same SCOP fold in the same connected component vs. the percent success in not placing pairs with different SCOP folds in the same connected component. In this analysis we consider each of the four major SCOP classes separately. We see that there are threshold combinations, e.g., 75 residues, 2.5Å RMSD, 30% sequence similarity (magenta diamond), that capture the fold-level connectivity, and achieve high values along both the *x* and *y* axes. On the other hand, none of the threshold combinations that we considered, successfully separates the alpha/beta domains into folds: either the domains are in the same connected component as other domains from their fold, but also from other folds, or, the domains are not with domains from other folds, but also not with ones from their own fold.

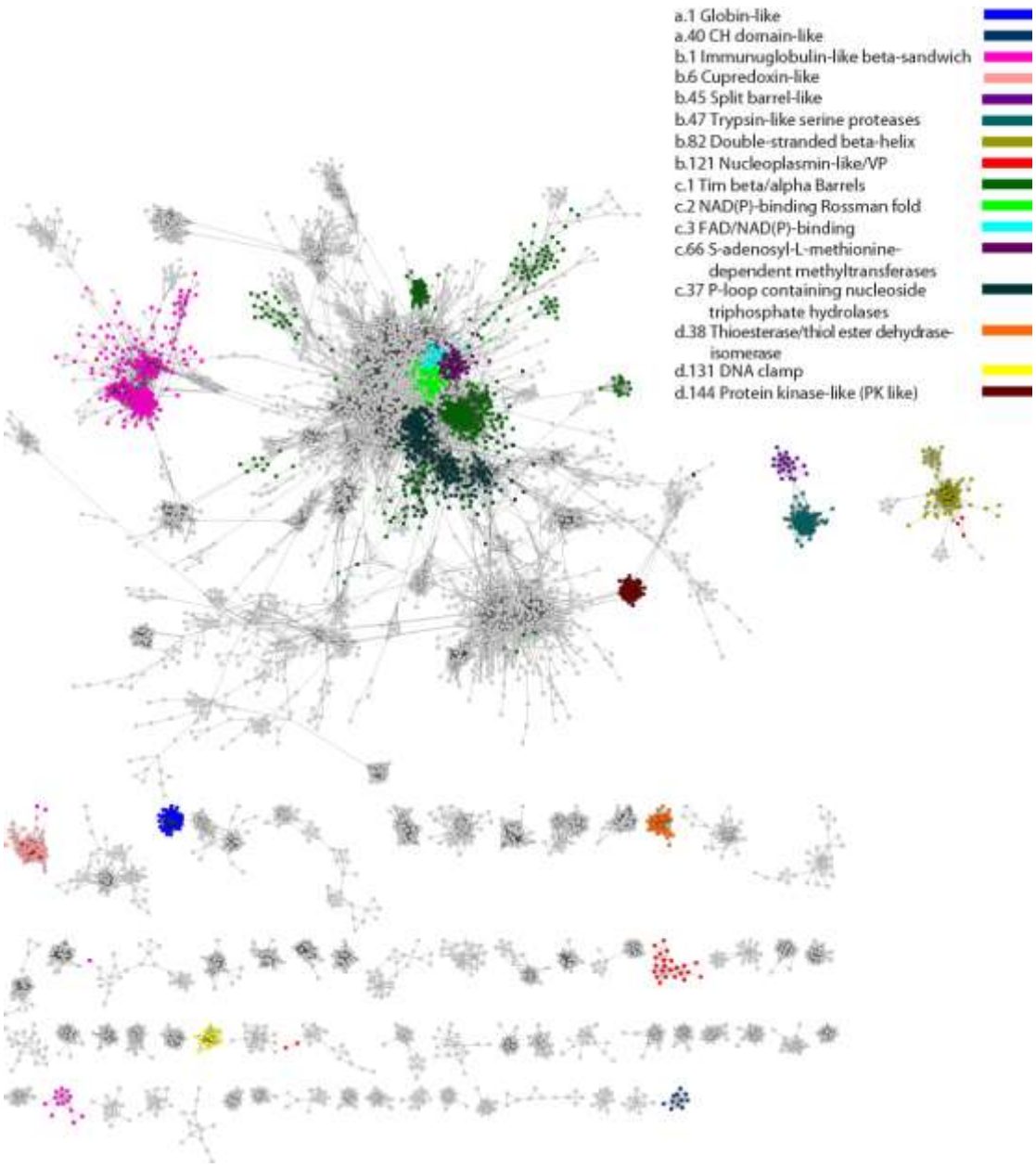

Figure S8: Several well-known SCOP folds in the domain network (shown in Figure 2 of the main manuscript). As expected, the domains of each of the folds generally lie together in the domain network. In some of the cases, the connected component agrees well with the fold characterization. We also see that the "usual suspects": the TIM barrels and the Rossman Folds, lie in the continous region of protein space.