Journal of proteome research

Article

pubs.acs.org/jpr

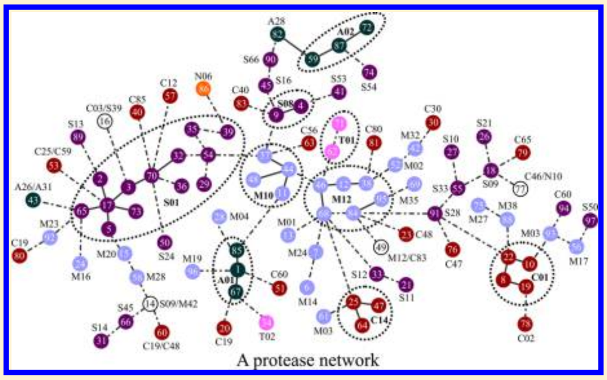
Visualizing and Clustering Protein Similarity Networks: Sequences, Structures, and Functions

Te-Lun Mai, Geng-Ming Hu, and Chi-Ming Chen*

Department of Physics, National Taiwan Normal University, Taipei, Taiwan

S Supporting Information

ABSTRACT: Research in the recent decade has demonstrated the usefulness of protein network knowledge in furthering the study of molecular evolution of proteins, understanding the robustness of cells to perturbation, and annotating new protein functions. In this study, we aimed to provide a general clustering approach to visualize the sequence−structure−function relationship of protein networks, and investigate possible causes for inconsistency in the protein classifications based on sequences, structures, and functions. Such visualization of protein networks could facilitate our understanding of the overall relationship among proteins and help researchers comprehend various protein databases. As a demonstration, we clustered 1437 enzymes by their sequences and structures using the minimum span clustering (MSC) method. The general structure of this protein network was delineated at two clustering resolutions, and the second level MSC clustering was found to be highly similar to existing enzyme classifications. The clustering of these enzymes based on sequence, structure, and function information is consistent with each other. For proteases, the Jaccard's similarity coefficient is 0.86 between sequence and function classifications, 0.82 between sequence and structure classifications, and 0.78 between structure and function classifications. From our clustering results, we discussed possible examples of divergent evolution and convergent evolution of enzymes. Our clustering approach provides a panoramic view of the sequence−structure−function network of proteins, helps visualize the relation between related proteins intuitively, and is useful in predicting the structure and function of newly determined protein sequences.



KEYWORDS: *protein similarity networks, sequence−structure−function relationship, sequence similarity, structure similarity*

1. INTRODUCTION

With the explosive number of newly discovered protein sequences deposited into databases in the postgenome era, it has become a challenging task to determine their structures and to characterize their functions efficiently for these protein sequences. Moreover, to reach a tight control of cellular processes, proteins often act in association with other proteins in a dynamic way.[1] Therefore, it is desirable to develop powerful methods for the study of protein networks.

Proteases, kinases, and phosphatases play essential roles in various biological and pathological processes.[2] Proteases encompass 50% industrially used enzymes.[3] They are involved in general metabolism through protein modification, such as food protein digestion, tissue protein mobilization, and zymogen processing, as well as cellular metabolism by proteasomes, which are large protease complexes that degrade unneeded or damaged proteins as a major mechanism for cellular regulation.[4] Unlike proteases performing irreversible proteolysis, kinases phosphorylate proteins and phosphatases dephosphorylate proteins reversibly. Phosphorylation or dephosphorylation results in a change in the structure of modified proteins, causing them to become activated or deactivated. Through the combined action of kinases and phosphatases, the activity of an enzyme can be reversibly altered.[5] Many intracellular or pericellular proteases are regulated by phosphorylation or dephosphorylation. For the phosphorylation and dephosphorylation of enzymes to serve a regulatory function, these two processes must in turn to be regulated by controlling kinases and phosphatases.

It is now recognized that, beyond nonspecific degradative functions, proteases also carry out highly selective cleavage of specific substrates, which regulates the activity of many proteins, modulates protein−protein interactions, creates new bioactive molecules, and contributes to the processing of cellular regulation. So far, more than 125 examples of kinases (including all the major branches of the kinome) that undergo regulated processing by one or more proteases have been observed, suggesting frequent direct interactions between proteases and kinases/phosphatases.[2d] Such bidirectional protease−kinase/phosphatase interactions play important roles in many cellular processes, such as apoptosis, transmembrane signaling, and cell migration. An example for the protease−kinase interplay in cell proliferation is demonstrated

Received: November 12, 2015
Published: June 7, 2016

DOI: 10.1021/acs.jproteome.5b01031
J. Proteome Res. 2016, 15, 2123−2131

by the experimental finding that proteolysis drives cell cycle progression by regulating the activity of cyclin-dependent kinases.[6] Dysregulation of protease-kinase/phosphatase interactions is relevant in various stages of cancer progression. Therefore, understanding the interplay between proteases, kinases, and phosphatases will offer new opportunities for cancer treatments, and has important clinical applications.

One easy-to-use technique to analyze diverse protein databases is the construction of protein similarity networks (PSNs), in which the interrelationships between proteins are described as a collection of independent pairwise alignments between sequences and structures. By incorporating function-related information, this technique provides a fast and easy way to compute the framework for intuitively observing the sequence−structure−function relationship among a large set of evolutionarily related proteins.[7] Several recent investigations have utilized PSNs to extract useful bioinformatics information for specific enzyme families, such as the bacterial protein-tyrosine kinase family,[8] the polymerase and histidinol phosphatase family,[9] and the C25 cysteine protease family.[10] Nevertheless, a panoramic view of the combined protease/kinase/phosphatase network is desirable.

It is a common practice to construct PSNs using sequence homology.[7a,8,11] However, we are aware of three issues in using a single E value ($-\log_{10} E$) from BLAST (Basic Local Alignment Search Tool) as the similarity measure to cluster protein sequences. First, the clustering of protein sequences often relies on an artificial threshold E value, and the threshold is different for different protein networks. Second, the range of $-\log_{10} E$ is between $-\infty$ and ∞ and its value is larger for closer pairs, which is not suitable to define a distance measure for network graphing. Third, although the consistency between structure similarity and this local homology measure is acceptable, there are considerable exceptions (as shown in Table S1) that will be further discussed in the Results and Discussion.

The aims of the present work are to construct a general approach for clustering protein sequences/structures and visualizing the relationship among proteins. In this study, new distance measures of protein sequence/structure similarity were proposed for clustering and visualizing protein networks. Further, we combined the minimum span clustering (MSC)[12] and the minimum spanning tree (MST)[13] methods to visualize the sequence−structure−function relationship of enzymes at two MSC resolutions. The clustering accuracy of MSC was previously shown to outperform hierarchical clustering (HC), Markov clustering (MCL), K-means, and affinity propagation (AP) for membrane proteins.[14] The MSC classifications based on sequence/structure information on enzymes were compared with their functional classifications in existing databases. For proteases considered in this study, they belong to 67 MEROPS families and are divided into 72 clusters in the second level MSC. The Jaccard's similarity between these two classifications is 0.86, which is considerably higher than the maximum value of 0.53 between MEROPS and various MCL classifications. Finally, we demonstrated the applicability of our method in predicting the structure and function of newly determined sequences.

2. METHODS

2.1. Data set preparation

In this study, the sequences and structures of proteases, kinases, and phosphatases were downloaded from the protein data bank (PDB). These protein sequences were chosen based on the Enzyme Commission's (EC) classification of enzymes[15] and the MEROPS classification of proteases.[16] As this is our first attempt to cluster the network of proteases, kinases, and phosphatases and to visualize their network structure using MSC and MST, we limited our data set to those nonredundant proteins with high quality structure. These sequences were sifted with the protein sequence culling server, PISCES,[17] by criteria of sequence length (between 40 and 10000), sequence identity (pairwise sequence identity less than 95%) and structural quality (X-ray crystal resolution better than 3 Å and the traditional crystallographic R-factor better than 0.3). The sifted data set contains 536 proteases, 882 kinases, and 199 phosphatases from 183 organisms, among which 40% of kinases, 25% of proteases, and 26% of phosphatases are from *Homo sapiens*. To avoid overrepresentation of human kinases (only about 7% of kinases in the database KinBase are human kinases), we kept the percentage of human kinases to 25% by randomly removing 180 human kinases from our data set. In total, 536 proteases, 702 kinases, and 199 phosphatases were considered in this study, and each enzyme was represented by a single chain. Detailed information regarding enzyme sequences in our data set is available in the Supporting Information Table S2. An additional set of 4464 protease sequences were selected randomly from Uniprot to generalize the test of MSC on clustering large protein networks. In addition, we randomly selected 300 enzyme sequences from Uniprot to test the applicability of our method in functional prediction for novel protein sequences.

2.2. Calculating the distance/similarity matrix

The BLAST E value, a parameter describing the number of hits one expected to see just by chance when searching the database of a particular size,[7a,11a,18] was computed using the general scoring matrix BLOSUM62 with default parameters. Since a lower E value infers a more significant match, the symmetrized similarity between sequences i and j is expressed as $s_{i,j} = \mathrm{e}^{-\sqrt{E_{i,j}E_{j,i}}}$. The similarity pattern of sequence i in the data set is denoted as $\vec{s_i} = \{s_{i,j}\}$, $j = 1, ..., 1437$. We further define the overall similarity between sequences i and j as the cosine measure of their similarity patterns, i.e., $\tilde{s}_{i,j} \equiv \vec{s_i} \cdot \vec{s_j} / |\vec{s_i}||\vec{s_j}|$. In general, our overall sequence similarity measure is similar to the neighborhood correlation, which also scores the similarity of two sequences by including all sequence pairs but is defined to be the correlation coefficient of their neighborhoods.[7b] Finally, we define the distance between sequences i and j in the distance matrix $\{d_{i,j}\}$ of our data set as

$$d_{i,j} = -\ln(\tilde{s}_{i,j}) \qquad (1)$$

This defined sequence distance is mainly contributed from those small E value pairs in the overall sequence similarity pattern, since large E values have little meaning.

The similarity between the tertiary structures of enzymes was calculated using TM-align, a protein structure alignment algorithm based on the TM-score (scaled by the average length of template proteins).[19] More information about TM-align is available in the Supporting Information. The transformation between similarity (measured by TM-score) and distance for tertiary structures was also defined as eq 1.

2.3. Network clustering and visualization

The MSC method was used to cluster the networks of enzyme sequences or structures from their distance matrix by minimizing both the average intracluster distance of each cluster and the overall connected distance of the network.[12] The general structure of the enzyme network was visualized using the MST method to minimize the length sum of all edges in the tree diagram. The constructed MST diagram is therefore also a connected tree diagram of MSC clusters. Here we briefly describe the three steps in the MSC procedure with a flowchart as shown in Figure 1, and provide an example of implementing MSC in the Supporting Information.

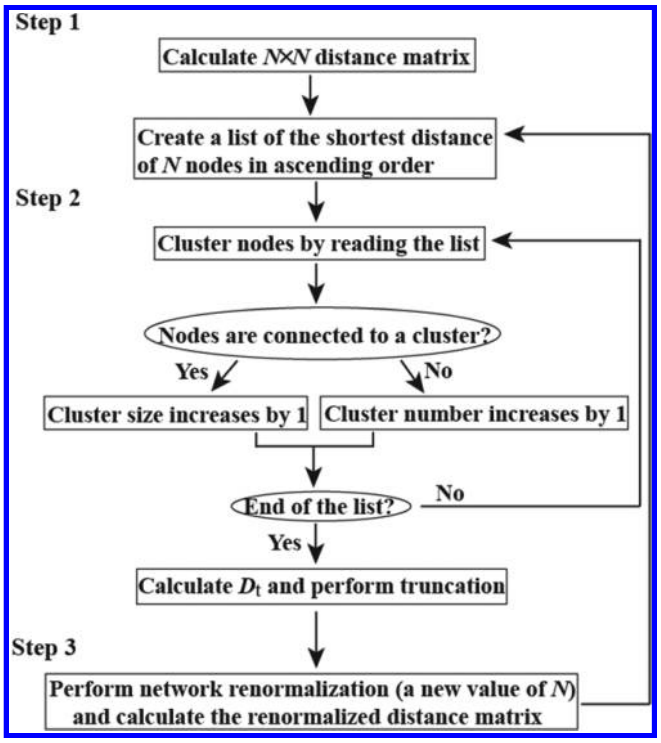

Step 1
Calculate N×N distance matrix
Create a list of the shortest distance of N nodes in ascending order
Step 2
Cluster nodes by reading the list
Nodes are connected to a cluster?
Yes
No
Cluster size increases by 1
Cluster number increases by 1
End of the list?
No
Yes
Calculate D_t and perform truncation
Step 3
Perform network renormalization (a new value of N) and calculate the renormalized distance matrix

Figure 1. Flowchart of the MSC network clustering algorithm.

Step 1. Starting with the $N \times N$ distance matrix for a network of N nodes, we identify the closest neighbor of each node and record their distances in a list of ascending order. For a network of N nodes, instead of dealing with a distance matrix of N^2 elements, MSC only processes N distances in the list in the next (clustering) step.

Step 2. The first cluster is constructed by starting from the shortest node pair, then including additional pairs from the list in the order of increasing distance. For the added distance, if one of the two nodes is involved in one of the above constructed clusters, the size of this cluster increases but the number of clusters remains the same. If both nodes of the distance are not involved in the above constructed clusters, a new cluster is identified and the number of clusters increases. All clusters of the network are found when all distances in the list are considered. Since several clusters contain outliers that are remotely related to most cluster members, we further calculate the threshold distance D_t (as described in the next paragraph) between clusters and truncate all links longer than D_t. The identified clusters after truncation in the first run are referred as the first level clustering, which has the highest resolution.

Step 3. Clusters constructed in step 2 are considered as renormalized nodes, and the average distance matrix of these clusters is calculated for all intercluster node pairs between two clusters. The network consisting of these renormalized components is further clustered by steps 1 and 2, and higher levels of clustering with lower resolutions are constructed.

To determine the threshold distance in step 2, as shown in Figure 2, we compared the length distribution of edges in the MST diagram and that in the nontruncated MSC clusters. Both MST and MSC deal with the minimum span of the network; MST constructs a connected network of the minimum span, while MSC constructs a group of minimum spanned subnetworks. The MSC edge distribution (dotted line) contains only intracluster edges, while the MST edge distribution (solid line) contains both intra- and intercluster edges. Thus, the comparison in the edge distributions (as a function of edge length) of MST and MSC gives a meaningful threshold distance (D_t) between clusters. Since our data set contains only a subset of proteins, outliers may be clustered with other remotely related sequences, and therefore MSC removes outliers by truncating links longer than D_t. As shown in the inset, the difference between MST and MSC edge distributions gives the length distribution of intercluster edges, which is vanishing small at sequence distances less than 0.4 but is significant at sequence distances greater than 0.4. Therefore, this characteristic sequence distance of 0.4 was set to be the threshold distance in step 2 for the cluster boundary of the sequence similarity network of proteases/kinases/phosphatases in the first level MSC. The comparison of our clustering results on various values of D_t in Figure S1 suggests a drastic change for D_t less than 0.3 due to the breaking of a significant amount of intracluster links.

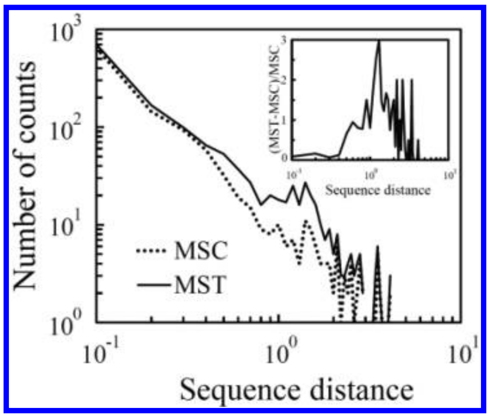


Figure 2. Statistical distributions of edge lengths in the MST diagram and in the first level MSC clusters (untruncated). The inset shows the relative difference of these two distributions, which gives the length distribution of intercluster edges.

2.4. Similarity measure for sets of clustering results

Consider two sets of clustering results, A = {A_1, A_2, ..., A_m} and B = {B_1, B_2, ..., B_n}, of the same data set. The similarity matrix of A and B can be expressed as

$$\mathbf{S}_{\mathbf{A},\mathbf{B}} = \begin{bmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mn} \end{bmatrix} \quad (2)$$

where the matrix element $\boldsymbol{S}_{\mathrm{ij}} = p/q$ is the Jaccard's similarity coefficient with p being the size of intersection and q being the size of the union of cluster sets A_i and B_j. In this study, the similarity of clustering results **A** and **B** is defined as $\mathrm{Sim}(\mathbf{A}, \mathbf{B}) = \Sigma_{i\leq m,\ j\leq n}S_{ij}/\max(m,n)$, and it has been shown that $0 < \mathrm{Sim}(\mathbf{A}, \mathbf{B}) \leq 1$ and $\mathrm{Sim}(\mathbf{A}, \mathbf{A}) = 1$.

3. RESULTS AND DISCUSSION

In previous studies, the BLAST E value of protein pairs has been widely used as the sequence distance in recognizing remote protein homology for structural and functional annotations of newly determined proteins, in spite of the mentioned drawbacks in the Introduction. Here we first discuss examples that a simple E value falsely predicts the similarity between proteins by comparing their TM scores, and justify $d_{i,j}$ (defined in section 2.2) as a better distance measure. The TM score is a metric for measuring the structural similarity of two proteins; a score below 0.17 corresponds to randomly chosen unrelated proteins, while a score above 0.5 assumes generally the same fold in SCOP/CATH.[19] Overall, by sifting protein pairs in our data set, 87% (147/169) pairs have a TM score greater than 0.5 for the criteria of $d_{i,j} < 0.16$ and $E > 1.0$ (small $d_{i,j}$ but large E value), while only 27% (4/15) pairs have a score larger than 0.5 for $d_{i,j} > 0.92$ ($s_{i,j} < 0.4$) and $E < 10^{-10}$ (large $d_{i,j}$ but small E value), indicating a higher consistency between structural similarity and small $d_{i,j}$ values. Table S1 selectively lists the sequence, structural, and functional properties of protein pairs having inconsistent d_{ij} and E values. Here the structural similarity of a protein pair is evaluated by the TM score and the posterior probability to find them in the same fold family using the Fold and Topology definition from SCOP.[20] In Figure S2, the structural similarity of these protein pairs is further demonstrated by the overlapping structures shown in (a) for kinases 3g2f-A (green) and 2iwi-A (red), in (b) for phosphatases 4n0g-A (green) and 1txo-A (red), and in (c) for proteases 1hne-E (green) and 1dle-A (red). In addition, without introducing any free parameter for clustering and visualizing the protein networks, MSC clustering results are shown to be consistent with the classification of existing databases.

3.1. Sequence−structure relationship of enzymes

The clustering of enzymes' similarity networks was carried out by MSC, which decomposed the sequence network into 267 clusters and 228 outliers, as listed in Table S2. Based on the EC numbers of these enzymes, the MSC clusters contain 128 kinase clusters, 42 phosphatase clusters, and 97 protease clusters. To verify the validity of our clustering result, we first investigated the sequence−structure relationship of the predicted MSC clusters. Protein structure similarity is often measured by TM score, and it has been found that the posterior probability from various data sets has a similar rapid phase transition at a TM score about 0.5, suggesting that protein pairs with a TM-score > 0.5 are mostly in the same fold while those with a TM-score < 0.5 are mainly not in the same fold.[21] Figure S3 shows the cumulative distributions of TM scores for intracluster protein pairs. The cumulative distributions sharply increase at a TM-score = 0.5 for all three enzyme categories. The percentage of intracluster protein pairs with TM-score > 0.5 is 95%, 97%, and 98% for proteases, kinases, and phosphatases, respectively. Alternatively, the posterior probability of intracluster protein pairs to be in the same fold family is 95.6%, 95.2%, and 97.9% for proteases, kinases, and phosphatases, respectively.

In Figure 3, we displayed the tertiary structure similarity matrix of 1437 enzymes, in which network nodes were ordered according to the first level MSC sequence clustering (outliers, kinases, phosphatase, and proteases in descending order). In this depiction, each matrix element is the TM-score between two sequences. The apparent block diagonalization pattern in Figure 3 suggests a strong correlation between the sequence and tertiary structure of three enzyme categories. The observed off-diagonal signals in Figure 3 indicate similar tertiary structures between proteins in related clusters, and these clusters merge together in the second level MSC. For example, as listed in Table S2, clusters 2, 3, 5, 16, 17, 29, 30, 32, 35, 36, 39, 54, 65, 70, and 73 are in the PA clan of MEROPS, and the corresponding off-diagonal elements have a high TM score. Since the classification of proteases into clans in MEROPS is mainly based on the similarity of their tertiary structures, this observation suggests a strong sequence−structure relationship of enzymes. Although recent work demonstrates that only three mutations are enough to induce a significantly different folding structure for some proteins, it is found in our analysis that natural proteins generally conserve the sequence−structure relationship. This result suggests that Paracelsus protein pairs are not seen to occur naturally and that comparative modeling of proteins is still a valid approach in predicting protein structure.[22] Similar patterns are also observed in Figure S4, whose matrix elements display the posterior probability of protein pairs to be in the same fold family.

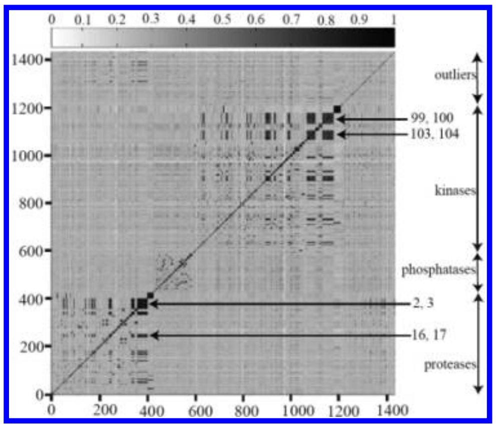


Figure 3. Tertiary structure similarity matrix of 1437 enzymes, ordered according to the first level sequence-based MSC results. The bar shows the TM score between two sequences with a grayscale from white (0) to black (1). The structure similarity is very low in the diagonal block for sequences at the upper right corner, which are isolated nodes in our clustering result.

3.2. Sequence−function relationship of enzymes

To further demonstrate the validity of the MSC results, we investigated the sequence−function relationship of predicted clusters. Figure 4 shows the network structure of 267 enzyme clusters, as visualized by their MST diagrams. In Figure 4(a), the function of each enzyme sequence was colored according to its EC number. Protein sequences in the three enzyme

DOI: 10.1021/acs.jproteome.5b01031
J. Proteome Res. 2016, 15, 2123−2131

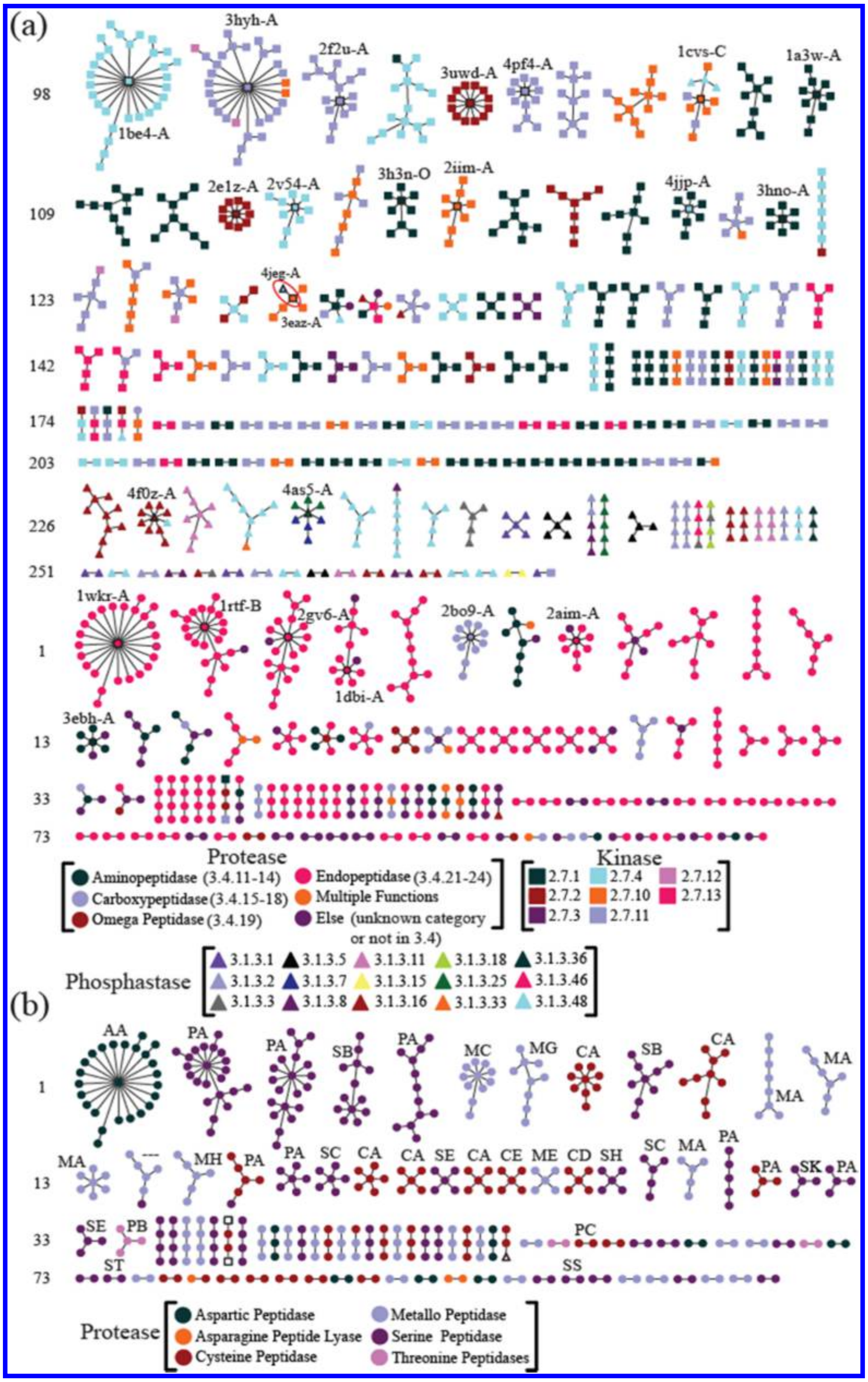


Figure 4. MST diagrams of enzyme clusters from the first level MSC. Each node is colored according to its function specified in the EC classification (267 clusters) (a) or in the MEROPS classification (97 protease clusters) (b), as described in the legend. These clusters are arranged from left to right in descending order of cluster size. We labeled the PDB ID for hubs with more than 5 connections in (a). In (b), we labeled the associated clan for the largest 34 clusters. Isolated nodes are not shown. The red circle in cluster 127 of (a) points out an example of mixing enzymes of different categories.

categories were represented by different shapes: circles for proteases, squares for kinases, and triangles for phosphatases. In general, the MSC result of 1437 enzymes in our data set is consistent with their functional classification of EC, as can be seen from the colors and shapes of nodes in the predicted MSC clusters. Only ten clusters mix nodes of different enzyme categories, such as 4jeg-A, the Src Homology 2 (SH2) domain of a protein tyrosine phosphatase, and 3eaz-A, the SH2 domain of a protein tyrosine kinase, in cluster 127. Such inconsistency results from the fact that the sequence of few enzyme chains in PDB is just a fragment, which disappears after removing fragment sequences. In general, our results agree well with the EC functional classification of enzymes. Several examples of divergent evolution were observed in Figure 4. For the divergent evolution observed in cluster 99, its hub (3hyh-A) and three neighboring nodes (3poz-A, 3ugc-A, and 4pdp-A) are highly similar in sequence (average distance 0.004, and E value ranging from 10^{-41} to 10^{-12}) and structure (average TM score 0.76, and 99% average probability to be in the same fold family). The superposition of their tertiary structures shown in Figure 5 suggests that these four kinase domains basically have the same geometrical features in three-dimensional space. All these kinase folds have a small N-terminal lobe and a large C-terminal lobe. The small lobe consists of five-stranded β-sheets and a helix. The large lobe is composed of 6 α-helices and a few short 3_{10} helices. Nevertheless, their functions are different from each other (EC numbers 2.7.11.1, 2.7.10.1, 2.7.10.2, and 2.7.12.1). For example, 3hyh-A is the protein kinase domain of yeast AMP-activated protein kinase Snf1, which is a master metabolic regulator for several intracellular systems, including the cellular uptake of glucose, the β-oxidation of fatty acids, and the biogenesis of glucose transporter type 4 and mitochondria.[23] It is an important target for drug development against diabetes, obesity, and other diseases. On the other hand, 3poz-A is the kinase domain of human epidermal growth factor 2 (HER2), which plays a role in the regulation of cell proliferation, differentiation, and migration.[24] Aberrant signaling of HER2 promotes cell proliferation and opposes apoptosis, and therefore a tight regulation of HER2 is helpful to prevent uncontrolled cell growth from occurring. Both protein kinases are activated through an allosteric mechanism.[25] More information related to possible examples of divergent evolution of enzymes is available in Table S3.

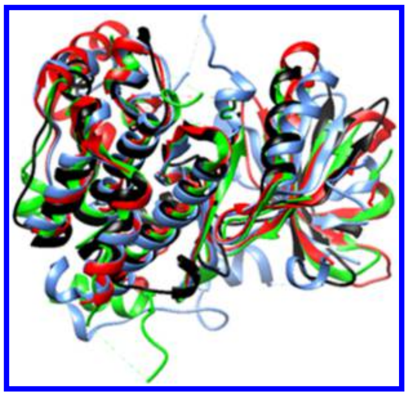


Figure 5. Superposition of the tertiary structures of kinases involved in an example of divergent evolution: 3hyh-A (hub, black), 3poz-A (blue), 3ugc-A (red), and 4pdp-A (green), as shown in cluster 99 of Figure 4(a).

In addition to the EC classification of enzymes, another known enzyme database is MEROPS, which classifies proteases in a hierarchical manner and manually assigns proteins to families and clans. In Figure 4(b), we colored nodes in the protease clusters (MSC clusters 1−97) according to the MEROPS classification. Six colors were used to represent the six catalytic types of proteases in MEROPS, including aspartic peptidases, asparagine peptide lyases, cysteine peptidases, metallo peptidases, serine peptidases, and threonine peptidases. Moreover, for the largest 34 clusters of proteases, their associated clan was also labeled if all of their members belong to the same clan. Among these clusters, two contain mixed enzymes, two contain proteases of mixed catalytic types, and one contains proteases of mixed clans. It is found that our MSC result of proteases seems to be more consistent with the MEROPS classification than the EC classification. However, as shown in clusters 1−97 in Figure 4(a), most cases of color inconsistency contain the color purple (proteases of unknown EC number). We believe that both EC and MEROPS are equally good at classification if the EC numbers of those proteases colored in purple were provided.

We have used the MSC method to cluster protease networks of size ranging from 536 to 5000 sequences in order to further validate the applicability of MSC on clustering large protein networks. The threshold distance of these protease networks is found to be around 0.5. Since the clustering result is not sensitive to a small change of the threshold distance as suggested by Figure S1, the threshold distance is set to be 0.5 for protease networks of various sizes. In general, we found that the value of D_t is not sensitive to network size, but differs for different systems. For every intracluster edge predicted by MSC, we checked if both proteases linked by the edge also belong to the same protease family in MEROPS, and calculated the precision, recall, and F_1-score of the first level MSC classifications as shown in Table 1. As defined in classification tasks, precision = TP/(TP + FP), recall = TP/(TP + FN), and F_1-score = 2·precision·recall/(precision + recall), where TP, FP, and FN stand for true positive, false positive, and false negative, respectively. The results in Table 1 suggest that MSC is applicable to networks of various sizes and the F_1-score of MSC classification increases with network size due to the improved prediction precision for large networks.

Table 1. Precision, Recall, and F_1-Score of MSC Classifications of Protease Networks of Size Ranging from 536 to 5000 Sequences

network size	precision	recall	F_1-score
536	0.958	0.988	0.973
1000	0.969	0.980	0.974
2000	0.978	0.978	0.978
3000	0.980	0.980	0.980
4000	0.981	0.983	0.982
5000	0.981	0.987	0.984

To further compare our clustering result of proteases with the MEROPS classification, we attempted to find a suitable MSC resolution that is comparable to the family classification of MEROPS. Since MEROPS families usually consist of several first level MSC clusters, the MSC resolution was set to the second level for a direct comparison. The MST diagram in Figure 6 delineates the general network structure of proteases, in which a node represents a first level MSC cluster. The sequence distance between two clusters is the average distance of their cluster members. In the second level MSC, these nodes were furthered clustered into groups, and the threshold distance was found to be 0.3. Figure 6 clearly shows a high consistency between the classifications of MEROPS (67

DOI: 10.1021/acs.jproteome.5b01031
J. Proteome Res. 2016, 15, 2123−2131

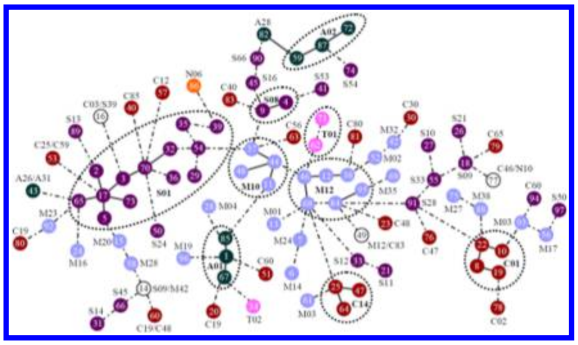


Figure 6. MST diagram of protease clusters. Each node represents a cluster in the 1st level MSC, and is colored as in Figure 4(b). The 2nd level MSC clusters these nodes into groups, and the intragroup and intergroup links are represented by solid lines and dot-dashed lines, respectively. Dashed circles enclose clusters in the same MEROPS family (as denoted).

families) and the second level MSC (72 groups). The Jaccard's similarity coefficient is 0.86 for these two classifications, which is considerably higher than the maximum value of Jaccard's similarity between the MEROPS' classification and the sequence classifications from various other clustering algorithms (0.53 for MCL, 0.78 for HC, and 0.54 for *K*-means as described in the Supporting Information). For comparison, the average similarity coefficient is 0.31 for ten random assignments of 67 protease clusters with the MEROPS classification. As shown in Table S4, there is some discrepancy between the second level MSC and MEROPS classifications in MEROPS families S01, M12, T01, and A02. For example, MSC clusters 36 and 70 are both in the S01 family of MEROPS, but they are in different MSC groups. The MSC clustering seems more reasonable than MEROPS considering their low average TM score of 0.20 (0.8% average probability to be in the same fold family) and a large average sequence distance of 19.8. It also seems reasonable that MSC clusters 59 and 82 are in different MEROPS families but are in the same MSC group, as they have a high average TM score of 0.65 (92% average probability to be in the same fold family) and a small average sequence distance of 0.26. The above comparison demonstrates that the sequence network and functional network of proteases correlate strongly, but are slightly different.

In the MEROPS classification of proteases, a family is a set of homologous proteases with a significant similarity in their functions, and a clan contains one or more families that show evolutionary evidence by their similar tertiary structures. For example, proteases in the clan PA exhibit a double β-barrel fold, proteases in the clan SB have a parallel β sheet structure, and proteases in the clan SC display an α,β-hydrolase fold. It was observed that proteases in these three clans have the same catalytic triad (serine, histidine, and aspartic acid) at their active site in spite of the dissimilarity in their sequences and structures.[26] Such an observation demonstrates a possible convergent evolution in proteases. A recent study systematically evaluated simple active site features from all serine, cysteine, and threonine proteases of independent lineage, and identified several convergently related cysteine proteases (e.g., 1euv-A of MSC cluster 23 and 1g2i-A of MSC cluster 63) and serine proteases (e.g., 2ic8-A of MSC cluster 74 and 1zrs-A of MSC cluster 90).[26b] This result is consistent with our clustering in Figure 6, which shows a large sequence difference for clusters 23 and 63, as well as clusters 74 and 90. To further study the convergent evolution of proteases, we searched for proteases with the same function but homologically different sequences. For simplicity, a protease's function was characterized by its EC number, as the fourth EC digit describes the specificity of the enzyme reaction by defining the specific reaction substrate/product or the cofactors used. In Table S5, we showed the convergent evolution observed for serine and cysteine proteases. For example, in MSC clusters 21, 33, and 89, serine proteases of EC number 3.4.16.4 have the same function, but their average values in sequence distance, *E* value, and TM score are respectively 23.5, 440.5, and 0.49. In MSC clusters 20, 40, and 57, cysteine proteases of EC number 3.4.19.12 have an average value of 55.1, 3064.5, and 0.25 in the sequence distance, *E* value, and TM score. Figure 7 schematically shows the superposition of the tertiary structures of 1xp4-A, 3pte-A, and 1w79-A (a) as well as that of 3mhs-A, 3c0r-A, and 1cmx-A (b). Although these proteins have very low sequence homology and belong to different sequence clusters, their core structures (the average TM score is 0.37 for entire proteins and is 0.56 for the overlapping cores in Figure 7) and functions are similar to each other. Since these nonhomologous sequences fold to form similar core structures and functions, we considered them as examples of convergent evolution. An enlargement of the core area of Figure 7(a) is shown in Figure S5. In Figure S6, we show the structural alignment of protease sequences involved in these two examples of convergent evolution. Aligned residues are located in the core area. Although these protease sequences are very different from each other and their sequence alignments are poor, the structural similarity in their core area is crucial for their functional similarity.

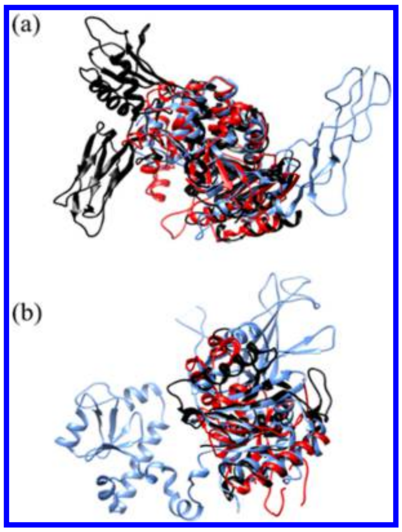


Figure 7. Superposition of the tertiary structures of proteases involved in two examples of convergent evolution: (a) 1xp4-A, 3pte-A, and 1w79-A, and (b) 3mhs-A, 3c0r-A, and 1cmx-A.

3.3. Sequence−structure−function relationship of enzymes

Our clustering results have shown high similarity between the second level MSC and existing databases for enzymes. Here we examined the second level MSC results of proteases to further delineate their sequence−structure−function relationship. The structure distance was calculated using the TM-score and eq 1. Figure S7 shows the comparison of the MEROPS classification with the MST diagram of 110 clusters of protease structures in

DOI: 10.1021/acs.jproteome.5b01031

the second level MSC ($D_t = 0.3$), suggesting a high consistency (the Jaccard's similarity coefficient is 0.78) between the functional and structural classifications. Meanwhile, the Jaccard's similarity coefficient is 0.82 between the sequence and structure classifications and is 0.86 between function and sequence classifications in the second level MSC. These values are much larger than 0.3 in the case of random assignments. The observed strong sequence−structure−function relationship for proteases provides evidence for the concept for proteins, "sequence determines structure determines function", that has been proposed for many decades.

3.4. Prediction and visualization of proteins' functions

In the era of genomics, the number of amino-acid sequences with unknown function grows exponentially. Bioinformatics tools can be useful in deciphering the role of these novel sequences in the cell or organism by comparing these sequences with those of proteins with known structure and function. To demonstrate the applicability of our method in this regard, we randomly selected 300 sequences (100 sequences for each enzyme category) from the Uniprot database (not included in our data set), and predicted their function by finding out their associated MSC cluster. Since the distance matrix in eq 1 varies with data sets, we investigated the test sequences one by one to minimize the perturbation to the distance matrix. In this test, the precision of functional prediction, which was calculated based on the functional categories in Figure 4, is 100% (80/80) for proteases, 97% (89/92) for kinases, and 91% (67/74) for phosphatases. Among 300 test sequences, 20 proteases, 8 kinases, and 26 phosphatases were identified as outliers, and their function cannot be annotated using the present data set. Table S6 shows the functional prediction of 300 test sequences.

Another application of this study is to provide an interface for intuitively understanding complex databases. For example, a simplification of the MEROPS database is presented in Figure 6, and further exploration of a specific clan or family can be linked to the corresponding clusters in Figure 4(b). The understanding of a novel sequence in its structure and function can also be achieved at various resolutions. Currently we are constructing an interactive web interface for viewing proteases, which could serve as a complementary tool to the MEROPS database.

4. CONCLUSION

In this study, we have successfully achieved our aims in constructing a general approach for clustering proteases, kinases, and phosphatases and visualizing their sequence−structure−function relationship. Our results are consistent with the EC and MEROPS classifications of these enzymes. We found high consistency between the second level MSC classifications in sequences/structures with existing functional classifications of MEROPS and EC. The observed strong sequence−structure−function relationship for the three enzyme categories provides some evidence for the proposed concept for proteins "sequence determines structure determines function", and supports the computational approach for the structure and function prediction of proteins using their sequence information. From our clustering results, we discussed possible examples of divergent evolution and convergent evolution in the network of proteases, kinases, and phosphatases. Our study not only estimates the consistency level for the sequence−structure−function relationship of enzymes, but also points out possible inconsistency due to convergent/divergent evolutions.

With the exponentially growing number of sequenced genomes, our method has the advantage of being accurate and efficient in predicting the structure and function of newly determined protein sequences. Furthermore, as demonstrated in this study, a panoramic view of the sequence−structure−function network of proteins is feasible using the present approach. We proposed to construct an interactive web interface for understanding the protein networks intuitively and exploring the networks at various resolutions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b01031.

Supplementary description of implemented methods in this study, supplementary figures discussed in the paper, and supplementary tables of the clustering results and analyses (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: cchen@phy.ntnu.edu.tw. Tel:+886277346039.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by the Ministry of Science and Technology of Taiwan under grant no. MOST102-2112-M-003-010-MY3.

■ ABBREVIATIONS

MSC, minimum span clustering; PSN, protein similarity network; BLAST, basic local alignment search tool; MST, minimum spanning tree; HC, hierarchical clustering; MCL, Markov clustering; AP, affinity propagation; PDB, protein data bank; EC, enzyme commission; SH2, Src homology 2; HER2, human epidermal growth factor 2

■ REFERENCES

(1) Alberts, B. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell* **1998**, *92* (3), 291−294.

(2) (a) Blume-Jensen, P.; Hunter, T. Oncogenic kinase signalling. *Nature* **2001**, *411* (6835), 355−365. (b) Forrest, A. R.; Ravasi, T.; Taylor, D.; Huber, T.; Hume, D. A.; Grimmond, S.; Group, R. G.; Members, G. S. L. Phosphoregulators: protein kinases and protein phosphatases of mouse. *Genome Res.* **2003**, *13* (6B), 1443−54. (c) Lopez-Otin, C.; Bond, J. S. Proteases: Multifunctional Enzymes in Life and Disease. *J. Biol. Chem.* **2008**, *283* (45), 30433−30437. (d) Lopez-Otin, C.; Hunter, T. The regulatory crosstalk between kinases and proteases in cancer. *Nat. Rev. Cancer* **2010**, *10* (4), 278−92.

(3) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5* (12), 993−996.

(4) Peters, J. M. Proteasomes - Protein-Degradation Machines of the Cell. *Trends Biochem. Sci.* **1994**, *19* (9), 377−382.

(5) Krebs, E. G.; Beavo, J. A. Phosphorylation-Dephosphorylation of Enzymes. *Annu. Rev. Biochem.* **1979**, *48*, 923−959.

(6) King, R. W.; Deshaies, R. J.; Peters, J. M.; Kirschner, M. W. How proteolysis drives the cell cycle. *Science* **1996**, *274* (5293), 1652−1659.

(7) (a) Atkinson, H. J.; Morris, J. H.; Ferrin, T. E.; Babbitt, P. C. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLoS One* **2009**, *4* (2), 14. (b) Song, N.; Joseph, J. M.; Davis, G. B.; Durand, D. Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput. Biol.* **2008**, *4* (5), e100006310.1371/journal.-pcbi.1000063 (c) Zhang, Y.; Zagnitko, O.; Rodionova, I.; Osterman, A.; Godzik, A. The FGGY Carbohydrate Kinase Family: Insights into the Evolution of Functional Specificities. *PLoS Comput. Biol.* **2011**, *7* (12), e100231810.1371/journal.pcbi.1002318

(8) Shi, L.; Ji, B.; Kolar-Znika, L.; Boskovic, A.; Jadeau, F.; Combet, C.; Grangeasse, C.; Franjevic, D.; Talla, E.; Mijakovic, I. Evolution of bacterial protein-tyrosine kinases and their relaxed specificity toward substrates. *Genome Biol. Evol.* **2014**, *6* (4), 800−17.

(9) Ghodge, S. V.; Fedorov, A. A.; Fedorov, E. V.; Hillerich, B.; Seidel, R.; Almo, S. C.; Raushel, F. M. Structural and Mechanistic Characterization of l-Histidinol Phosphate Phosphatase from the Polymerase and Histidinol Phosphatase Family of Proteins. *Biochemistry* **2013**, *52* (6), 1101−1112.

(10) Cross, K. J.; Huq, N. L.; Reynolds, E. C. A bio-informatics study of the c25 cysteine protease family. *Open J. Genet.* **2013**, *2*, 18.

(11) (a) Apeltsin, L.; Morris, J. H.; Babbitt, P. C.; Ferrin, T. E. Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution. *Bioinformatics* **2011**, *27* (3), 326−333. (b) Enright, A. J.; Van Dongen, S.; Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* **2002**, *30* (7), 1575−1584. (c) Paccanaro, A.; Casbon, J. A.; Saqi, M. A. Spectral clustering of protein sequences. *Nucleic Acids Res.* **2006**, *34* (5), 1571−1580. (d) Uberto, R.; Moomaw, E. W. Protein Similarity Networks Reveal Relationships among Sequence, Structure, and Function within the Cupin Superfamily. *PLoS One* **2013**, *8* (9), 10. (e) Wittkop, T.; Baumbach, J.; Lobo, F. P.; Rahmann, S. Large scale clustering of protein sequences with FORCE - A layout based heuristic for weighted cluster editing. *BMC Bioinf.* **2007**, *8*, 12.

(12) Chang, Y. F.; Chen, C. M. Classification and Visualization of the Social Science Network by the Minimum Span Clustering Method. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62* (12), 2404−2413.

(13) Kruskal, J. B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.* **1956**, *7* (1), 48−50.

(14) Hu, G.-M.; Mai, T.-L.; Chen, C.-M. Clustering and visualizing similarity networks of membrane proteins. *Proteins: Struct., Funct., Genet.* **2015**, *83* (8), 1450−1461.

(15) Bairoch, A. The ENZYME database in 2000. *Nucleic acids research* **2000**, *28* (1), 304−305.

(16) Rawlings, N. D.; Waller, M.; Barrett, A. J.; Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **2014**, *42*, gkt953.

(17) Wang, G.; Dunbrack, R. L. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **2005**, *33* (suppl 2), W94−W98.

(18) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **1997**, *25* (17), 3389−3402.

(19) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **2005**, *33* (7), 2302−2309.

(20) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics* **2010**, *26* (7), 889−895.

(21) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26* (7), 889−895.

(22) Sadowski, M. I.; Jones, D. T. The sequence−structure relationship and protein function prediction. *Curr. Opin. Struct. Biol.* **2009**, *19* (3), 357−362.

(23) (a) Ojuka, E. O. Role of calcium and AMP kinase in the regulation of mitochondrial biogenesis and GLUT4 levels in muscle. *Proc. Nutr. Soc.* **2004**, *63* (02), 275−278. (b) Winder, W. W. Energy-sensing and signaling by AMP-activated protein kinase in skeletal muscle. *J. Appl. Physiol.* **2001**, *91* (3), 1017−1028. (c) Bergeron, R.; Russell, R. R.; Young, L. H.; Ren, J.-M.; Marcucci, M.; Lee, A.; Shulman, G. I. Effect of AMPK activation on muscle glucose metabolism in conscious rats. *American Journal of Physiology - Endocrinology and Metabolism* **1999**, *276* (5), E938−E944.

(24) Schlessinger, J. Common and Distinct Elements in Cellular Signaling via EGF and FGF Receptors. *Science* **2004**, *306* (5701), 1506−1507.

(25) (a) Aertgeerts, K.; Skene, R.; Yano, J.; Sang, B.-C.; Zou, H.; Snell, G.; Jennings, A.; Iwamoto, K.; Habuka, N.; Hirokawa, A.; Ishikawa, T.; Tanaka, T.; Miki, H.; Ohta, Y.; Sogabe, S. Structural Analysis of the Mechanism of Inhibition and Allosteric Activation of the Kinase Domain of HER2 Protein. *J. Biol. Chem.* **2011**, *286* (21), 18756−18765. (b) Rudolph, M. J.; Amodeo, G. A.; Bai, Y.; Tong, L. Crystal structure of the protein kinase domain of yeast AMP-activated protein kinase Snf1. *Biochem. Biophys. Res. Commun.* **2005**, *337* (4), 1224−1228.

(26) (a) Polgár, L. The catalytic triad of serine peptidases. *Cell. Mol. Life Sci.* **2005**, *62* (19−20), 2161−2172. (b) Buller, A. R.; Townsend, C. A. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (8), E653−E661.