

# Comparing protein structures and inferring functions with a novel three-dimensional Yau–Hausdorff method

Kun Tian, Xin Zhao, Yuning Zhang & Stephen Yau

To cite this article: Kun Tian, Xin Zhao, Yuning Zhang & Stephen Yau (2019) Comparing protein structures and inferring functions with a novel three-dimensional Yau–Hausdorff method, Journal of Biomolecular Structure and Dynamics, 37:16, 4151-4160, DOI: [10.1080/07391102.2018.1540359](https://doi.org/10.1080/07391102.2018.1540359)

To link to this article: <https://doi.org/10.1080/07391102.2018.1540359>



Published online: 05 Dec 2018.



Submit your article to this journal [↗](#)



Article views: 93



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



## Comparing protein structures and inferring functions with a novel three-dimensional Yau–Hausdorff method

Kun Tian<sup>a\*</sup>, Xin Zhao<sup>a\*</sup>, Yuning Zhang<sup>b</sup> and Stephen Yau<sup>a</sup>

<sup>a</sup>Department of Mathematical Sciences, Tsinghua University, Beijing, P.R. China; <sup>b</sup>School of Life Sciences, Tsinghua University, Beijing, P.R. China

Communicated by Ramaswamy H. Sarma

### ABSTRACT

Structures and functions of proteins play various essential roles in biological processes. The functions of newly discovered proteins can be predicted by comparing their structures with that of known-functional proteins. Many approaches have been proposed for measuring the protein structure similarity, such as the template-modeling (TM)-score method, GRaphlet (GR)-Align method as well as the commonly used root-mean-square deviation (RMSD) measures. However, the alignment comparisons between the similarity of protein structure cost much time on large dataset, and the accuracy still have room to improve. In this study, we introduce a new three-dimensional (3D) Yau–Hausdorff distance between any two 3D objects. The (3D) Yau–Hausdorff distance can be used in particular to measure the similarity/dissimilarity of two proteins of any size and does not need aligning and superimposing two structures. We apply structural similarity to study function similarity and perform phylogenetic analysis on several datasets. The results show that (3D) Yau–Hausdorff distance could serve as a more precise and effective method to discover biological relationships between proteins than other methods on structure comparison.

### ARTICLE HISTORY

Received 31 July 2018  
Accepted 22 October 2018

### KEYWORDS

Three-dimensional Yau–Hausdorff distance; structure comparison; classification; protein function; phylogenetic analysis

### 1. Introduction


Central problems in protein classification study are how proteins are clustered in relation to each other (Holm & Sander, 1996) and how to compare functional similarity based on protein structures. In order to get a global view of evolutionary distances among multiple proteins, the concept of protein space was introduced (Smith, 1970). In earlier research, protein space was defined based on amino acid sequences of the proteins. Protein space was also known as sequence space (DePristo, Weinreich, & Hartl, 2005). Under this definition, each amino acid in a protein sequence was represented by one dimension (1D), with 20 possibilities in the space. As the length of natural protein sequences had an upper limit, all natural proteins could be put in a finite-dimensional protein space based on sequence (DePristo et al., 2005; Smith, 1970). The evolutionary distance between two different proteins in sequence space was measured by sequence alignment, which was commonly based on substitution matrix such as blocks substitution matrix (BLOSUM) and point-accepted mutation matrix (PAM) (Henikoff & Henikoff, 1992). Substitution matrices were introduced to give a theoretical basis for gene mutation, especially point mutation. If two protein sequences are highly similar, differing in only a few amino acids, sequence alignment could

accurately identify the evolutionary kinship of the associated proteins (Henikoff & Henikoff, 1992; Holm & Sander, 1996). However, sequence alignment can be inaccurate when the evolutionary distances are relatively big and the difference between protein sequences cannot be explained by point mutation. Furthermore, some compensatory mutations can happen during evolution, resulting in proteins retaining their original function even though their sequences have been altered. Under these circumstances, sequence comparison cannot reveal the evolutionary distance between proteins, and a more sophisticated comparison method is required to identify subtle similarities that remain throughout long-term evolution (Holm & Sander, 1996).

Another protein space representation using natural vectors was proposed by Yau. The original definition of natural vector was proposed in 2011 (Deng, Yu, Liang, He, & Yau, 2011) for DNA sequence and in 2013 (Yu et al., 2013a) for protein sequence. The main idea for natural vector is to represent each protein sequence as an element in  $R^{60}$ , with each dimension being a statistic for the distribution of amino acids in the protein sequence. The biological distance between any two proteins can be represented by the Euclidean distance between the corresponding points in a 60-dimensional Euclidean space (Tian, Zhao, & Yau, 2018; Tian et al., 2015; Yu et al., 2013a). In the previous study, the

CONTACT Stephen Yau  [yau@uic.edu](mailto:yau@uic.edu)

\*These authors contributed equally to this work.

 Supplemental data for this article can be accessed [here](#).

© 2018 Informa UK Limited, trading as Taylor & Francis Group

natural vector was established simply based on sequence information and has been widely used in phylogenetic analysis (Huang et al., 2014; Yu et al., 2013b; Zhao, Tian, He, & Yau, 2017; Zhao, Wan, He, & Yau, 2016). However, besides the distribution of amino acids in a sequence, protein structures contain a large amount of biological information of proteins. As protein structure closely associates with its function, how to compare structures efficiently and effectively becomes a very important research topic. Such a method will be more sophisticated than the method of sequence comparison (Holm & Sander, 1996). Meanwhile, the number of known protein structure is increasing with the rapid development of structural biology. Current techniques used to compare the structures of proteins such as structure alignment methods required long computation time to analyze the experimental results, especially for large protein structures. Therefore, some new methods were proposed in studying protein structure comparison. For example, a scoring function pcSM based on the C-alpha Euclidean metric, secondary structural propensity, surface areas, and an intramolecular energy function parameters has showed the improvement for discriminating a true native structure from an ensemble of candidate structures (Mishra, Rao, Mittal, & Jayaram, 2013). The stoichiometry of amino acids of a given primary sequence together with the Euclidean distance also reveals strong correlation with backbones of folded proteins (Mittal & Jayaram, 2011; Mittal, Jayaram, Shenoy, & Bawa, 2010). The D2N metric has been proposed by combining chemical and physical properties of soluble proteins structural features which could calculate how far a structure is from its native state even without knowing the experimental structure (Mishra, Rana, Mittal, & Jayaram, 2014). Several measures used to compare protein structure were described, such as the root-mean-square deviation (RMSD) measure (Kabsch, 1978), the template-modeling score (TM-score) (Zhang & Skolnick, 2004), the GRaphlet-based Aligner (GR-Align) (Noel & Natasa, 2014), and combinatorial extension (CE) (Prlic et al., 2010). However, these methods have some limitation to obtain accurate results and often consume much time to complete structure comparison so that the computational complexity will be high for large dataset.

Two-dimensional (2D) Yau–Hausdorff distance (Tian et al., 2015) has been proposed to study the sequence comparison. Although it performs well for comparison of DNA and protein sequences, it cannot be used to compare three-dimensional (3D) protein structures. Therefore, we develop a new metric, called 3D Yau–Hausdorff distance, to measure the similarity between protein structures. The (3D) Yau–Hausdorff distance does not require the compared proteins to be aligned before calculation. It can measure the similarity/dissimilarity of protein structures without superimposing them together. The (3D) Yau–Hausdorff is a natural generalization for the minimum 1D Hausdorff distance and takes all possible translation and rotation into full consideration. The complexity of this new method is lower than many other comparison algorithms by descending dimension in calculation without losing information of the structure. Compared with other methods mentioned above, our new approach

could be applied on protein structure dataset more efficiently. These advantages enable it to be a powerful tool for comparing protein structures. Thus, the (3D) Yau–Hausdorff distance can measure the similarity even when proteins are highly dissimilar, filling the gap in which sequence comparison lacks accuracy.

In this study, we first tested our method on a benchmark dataset and compared the results with existing methods. The accuracy results, running time, and precision–recall (PR) curves show that the (3D) Yau–Hausdorff method is more accurate and effective than GR-Align, RMSD, TM-score, and CE methods on protein structure comparison. We then used structural similarity measured by the (3D) Yau–Hausdorff distance to discover function similarity masked by sequence divergence. We tried to find proteins with similar structures to lscA, of which homologous protein in fruit fly has a magnetic property. After structure comparison, we got a list of proteins with small (3D) Yau–Hausdorff distance to lscA and three of them worked in the same biological pathway as lscA. Moreover, we used structure comparison instead of sequence alignment as the measure of evolutionary distance in molecular phylogenetic analysis. Choosing cytochrome c and  $\beta$  globin as the molecular clock, we compared sequence alignment, natural vector method, and structure comparison on their performance when serving as the metric to cluster species with homologous proteins into phylogenetic trees. Finally, we drew the conclusion that structure comparison performed by the (3D) Yau–Hausdorff distance could reveal the function similarity hidden by sequence dissimilarity. It can make up the gap where sequence alignment cannot detect evolutionary relationship at a longer evolutionary distance.

## 2. Materials and methods

### 2.1. 3D Yau–Hausdorff distance

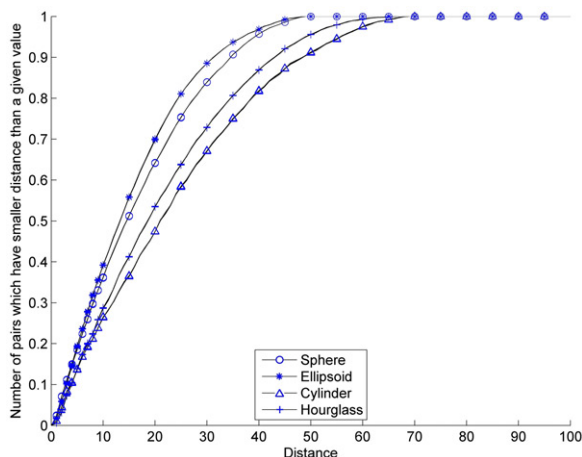
The 3D Yau–Hausdorff distance is used to calculate the similarity between protein structures. Each protein structure is regarded as a 3D point set including the coordinates of backbone chain atoms and side-chain atoms. A detailed description about how to choose the coordinates is shown in the [supplementary material](#). The purpose is to compute the difference between the corresponding point sets of protein structures. We first define the minimum 1D Hausdorff distance of two finite point sets  $A_1$  and  $B_1$  in  $R$  as

$$H^1(A_1, B_1) = \min_{t \in R} h(A_1 + t, B_1), \quad (1)$$

where  $t$  is a real number,  $A_1 + t$  represents the point set containing the sum of any number in  $A_1$ , and  $t$  and  $h$  are the Hausdorff distance

$$h(A_1, B_1) = \max \left\{ \max_{a \in A_1} \min_{b \in B_1} d(a, b), \max_{b \in B_1} \min_{a \in A_1} d(b, a) \right\}, \quad (2)$$

here,  $d(a, b)$  is the Euclidean distance between two points  $a$  and  $b$ , and  $h(A_1 + t, B_1)$  stands for the Hausdorff distance between  $A_1$  and  $B_1$  after shifting  $A_1$  by  $t$ . The (3D) Yau–Hausdorff distance  $D(A, B)$  of two point sets  $A$  and  $B$  in  $R^3$  is then defined in terms of  $H^1$ :



**Figure 1.** The normalized relation curves between the number of distances and distance values on four different geometries sphere, ellipsoid, cylinder, and hourglass. Based on each of the four geometries, 10 collections of point sets are generated. For each collection, the (3D) Yau–Hausdorff distances between each pair of point sets are calculated and used to plot the curve. Thus, we obtain 10 curves (almost overlap completely) for each geometry. Each curve is normalized by dividing the corresponding maximal value number  $R_{\text{total}}$ . The points with  $x = 0, 1, 2, \dots, 10, 15, 20, 25, \dots, 95$  are marked with circles, asterisks, triangles and pluses for sphere, ellipsoid, cylinder and hourglass, respectively. Blue color, sphere; red color, ellipsoid; green color, cylinder; pink color, hourglass.

$$D(A, B) = \max \left\{ \max_{\theta^3} \min_{\varphi^3} H^1 \left( P_x(A^{\theta^3}), P_x(B^{\varphi^3}) \right), \max_{\varphi^3} \min_{\theta^3} H^1 \left( P_x(A^{\theta^3}), P_x(B^{\varphi^3}) \right) \right\}, \quad (3)$$

where  $P_x(A^{\theta^3})$  is a 1D point set representing the projection of  $A$  on the  $x$  axis after being rotated by 3D rotation matrix  $\theta^3$ .

The algorithm to compute the (3D) Yau–Hausdorff distance  $D$  of two protein structures is as follows:

Let  $A = \{a_1, a_2, \dots, a_m\} \subset R^3$ ,  $B = \{b_1, b_2, \dots, b_n\} \subset R^3$  be the corresponding 3D atom coordinate point sets of two protein structures. The values  $D_1$  and  $D_2$  will be more precise if we choose more rotations  $\theta^3$  and  $\varphi^3$  in the calculation theoretically. That means the computational result gets more accurate and stable when the numbers of rotations are large enough. The way for choosing the appropriate rotation times is explained by a control computation example in the Discussion section. In general, the rotation number of each protein structure should be at least 40 and 50 is enough for obtaining stable results. Here, randomly rotate  $A$  50 times by  $\theta_1^3, \theta_2^3, \dots, \theta_{50}^3$ ,  $B$  50 times by  $\varphi_1^3, \varphi_2^3, \dots, \varphi_{50}^3$ , and take  $M = \{\theta_1^3, \theta_2^3, \dots, \theta_{50}^3\}$ ,  $N = \{\varphi_1^3, \varphi_2^3, \dots, \varphi_{50}^3\}$ . For each  $\theta^3 \in M$  compute

$$D_1 = \max_{\theta^3 \in M} \min_{\varphi^3 \in N} H^1 \left( P_x(A^{\theta^3}), P_x(B^{\varphi^3}) \right), \quad (4)$$

similarly,

$$D_2 = \max_{\varphi^3 \in N} \min_{\theta^3 \in M} H^1 \left( P_x(A^{\theta^3}), P_x(B^{\varphi^3}) \right). \quad (5)$$

Take

$$D(A, B) = \max\{D_1, D_2\}. \quad (6)$$

as the final (3D) Yau–Hausdorff distance result of the two protein structures.

**Table 1.** The (3D) Yau–Hausdorff distances between identical and different geometries.

Distance	Spherical	Ellipsoidal	Cylindrical	Hourglass
Spherical	2.8759	12.6317	14.7299	12.5620
Ellipsoidal	12.6317	3.5067	20.6725	15.6916
Cylindrical	14.7299	20.6725	4.7598	8.2008
Hourglass	12.5620	15.6916	8.2008	5.5742

The shadings in Table 1 show the (3D) Yau–Hausdorff distances between identical geometries. These values should be significantly smaller than the distances between different geometries in the forms without shading.

The size of protein structure is uniformly measured by using the unit angstrom. The atomic coordinates in the PDB files used in this study are also measured based on angstrom. Therefore, the unit of (3D) Yau–Hausdorff distance is still angstrom when we calculate the distance between proteins. No matter what the size of protein is, the result is considered reliable if the units are kept consistent. There is no limitation of size difference in protein structure comparison by our metric.

## 2.2. PR curve

The prediction accuracy is measured in terms of area under the PR curve (AUPRC). Given a threshold  $d$  on the pairwise distances between proteins, we compute four values as follows: (1) the true positives ( $TP$ ): the number of distances smaller than  $d$  coming from the same group; (2) the true negatives ( $TN$ ): the number of distances greater than or equal to  $d$  coming from different groups; (3) the false negatives ( $FN$ ): the number of distances greater than or equal to  $d$  coming from the same group; (4) the false positives ( $FP$ ): the number of distances smaller than  $d$  coming from different groups. The PR curve plots the precision rate  $P = TP / (TP + FP)$  as a function of recall rate  $R = TP / (TP + FN)$ , for  $d$  increases from the minimum to the maximum distance. The AUPRC measures the average precision of these pairwise distances. Therefore, the closer the AUPRC is to one, the better the method is applied on classification.

## 3. Results

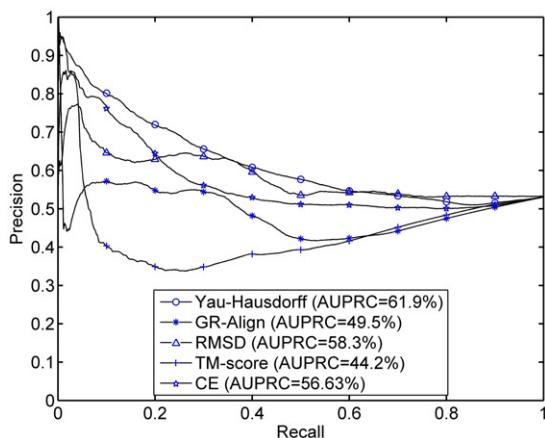
### 3.1. Computation control experiments on the geometries, sphere, ellipsoid, cylinder, and hourglass

Firstly, we set up a control experiment on four simple ideal geometries, sphere, ellipsoid, cylinder, and hourglass like Mittal presented (Mittal & Acharya, 2012) to show the performance of (3D) Yau–Hausdorff distance on different shapes. Take sphere as an example, and the detailed steps are as follows. Fifty concentric shells with a radius from 1 to 50 were constructed. The shape of each shell was simulated with 9900 points forming a geometrical sphere surface and centered at the origin using Matlab software. For each shell, a number of  $R_i$  ( $i = 1, 2, \dots, 50$ ) were generated randomly between 0 and 9.  $R_i$  fixed the number of sets from the given geometrical shell. Each set contained 200 random points in this shell. We then got  $R_1 + R_2 + \dots + R_{50} = R_{\text{total}}$  3D point sets in total. The (3D) Yau–Hausdorff distances between each pair of the  $R_{\text{total}}$  sets were computed. A key aspect to be



**Table 2.** The classification results and running times of the (3D) Yau–Hausdorff method with different rotation numbers, GR-Align method, RMSD method, TM-score method, and CE method.

	GR-Align	RMSD	TM-score	CE	Yau–Hausdorff (rotation 10)	Yau–Hausdorff (rotation 2500)
Accuracy	62.3%	59.2%	61.5%	60.8%	70.8%	81.5%
Running time	2 min	1 h	9 h 20 min	340 h	10 min	4 h 10 min

**Figure 2.** Precision–recall curves and areas under these curves of (3D) Yau–Hausdorff (rotation number 2500) as well as traditional methods performed on the 260 protein domains dataset. Blue color, (3D) Yau–Hausdorff; red color, GR-Align; green color, RMSD; pink color, TM-score; yellow color, CE.

investigated is the distribution of these distance values. We plotted the relation between a given value and the number of distances within the fixed value shown in Figure S1, [supplementary material](#). To view this result more directly, each curve was normalized by dividing the corresponding maximal value number  $R_{total}$ . This process was repeated 10 times with the same  $R_1, R_2, \dots, R_{50}$ , and 10 curves were drawn to compare with each other. The same steps were also done for the other three geometries: ellipsoid, cylinder, and hourglass. As a result, 40 curves are shown in Figure 1. In this figure, the 10 curves of each geometry almost overlap completely, while the curves of different geometries are slightly distinct. In addition, we also calculated the (3D) Yau–Hausdorff distances between identical and different geometries including spherical, ellipsoidal, cylindrical, and hourglass point sets. For each kind of geometry, we generated 10 point sets with the given shape in a  $100 \times 100 \times 100$  cube using Matlab software. Each point set contains 200 random points. The average values of (3D) Yau–Hausdorff distances between identical and different geometries are listed in Table 1. It shows that the (3D) Yau–Hausdorff distances between identical geometries are significantly smaller than different geometries.

### 3.2. Classification analysis of 260 protein domains by (3D) Yau–Hausdorff method

We show the high classification accuracy and effectiveness of the (3D) Yau–Hausdorff method in a structural similarity comparison. The benchmark dataset used in this study consisted of 260 protein domains. This dataset was downloaded from the CATH v4.2.0 database, with a number of residues varying from 44 to 854 and 211 on average. The data can be clustered into two superfamilies. One is C-terminal domain in

DNA helicase RuvA subunit coming from the Alpha class, Orthogonal Bundle Architecture, Helicase, and RuvA Protein fold. The other superfamily Homing endonucleases belongs to the Alpha and Beta class, Roll Architecture, and Endonuclease I-crel fold. By comparing the results and running time using different rotation numbers with that obtained by GR-Align, RMSD, TM-score, and CE, we prove that our method performs better than those methods on protein structure comparison.

We applied the (3D) Yau–Hausdorff method with two rotation numbers 10 and 2500, respectively, to calculate the pairwise distances between any pair of the 260 protein domains and got the distance matrix. The 1-nearest neighbor accuracy rate (1-NN) is an effective way of assessing the quality of score function methods, by counting the number of proteins that are from the same class with their nearest neighbors in the reference classification. The accuracies were also computed by GR-Align, RMSD, TM-score, and CE methods. The programs were downloaded from <http://bio-nets.doc.ic.ac.uk/home/software/gralign/> and <https://zhanglab.ccmb.med.umich.edu/TM-score/>. All the programs were done on a PC with a configuration of 2.40GHz and 8Gb RAM. Table 2 shows the results and running times by these approaches. The best accuracy was performed by the (3D) Yau–Hausdorff method with rotation number 2500 (81.5%). Although the running time of it was longer than GR-Align and RMSD methods, the accuracy rate was much higher than GR-Align (62.3%), TM-score (61.5%), CE (60.8%), and RMSD (59.2%). In order to accelerate the running speed, we tested the result of the (3D) Yau–Hausdorff method with rotation number 10, and the result was still better than the other methods. GR-Align presented a lower accuracy in spite of it having the fastest computation time. We also drew the PR curves of these approaches shown in Figure 2. In this figure, we can see that the AUPRC presented by the (3D) Yau–Hausdorff method (rotation number 2500) achieves higher than the other four.

The (3D) Yau–Hausdorff method performs better than GR-Align, RMSD, TM-score, and CE methods because it takes all possible translations and rotations into consideration to achieve the best match of two protein structures. The information of the structures is not lost during the computing process, and the distance can precisely measure the difference between protein structures. This method could complete a global structure comparison task based on 3D coordinates, no matter how long the residues are and where the residues locate in the corresponding sequences. Although the classical (3D) Hausdorff distance under rigid motion can give an accurate distance between two protein structures, it cannot be implemented due to its high computational complexity. Our method not only completes the task but also has a lower complexity than the (2D) minimum Hausdorff distance, which could help us save much time in protein structure comparison.

**Table 3.** The (3D) Yau–Hausdorff distances between IscA and three proteins IscU, SufA, and OsCnfU-1A domain I.

(3D) Yau–Hausdorff distance	IscU	SufA	OsCnfU-1A domain I
IscA	2.365	1.233	0.909

### 3.3. Identification of functional similarity via structural similarity

Based on the principle that protein function is closely related to its structure, we hope that after structure comparison, we would be able to identify the functional similarity between proteins with similar structures. Firstly, we started with a recently reported protein having a magnetic property (Bilder, Ding, & Newcomer, 2004). The protein was named MagR and as a newly identified protein in fruit fly with a never-reported magnetic-sensing ability. It was highly conserved during evolution, and the structure of its homologous protein in *Escherichia coli*, IscA, was already solved in 2004 (PDB ID: 1R94) (Bilder et al., 2004).

With the purpose of finding more protein candidates with the magnetic property, we screened a set of proteins by calculating its (3D) Yau–Hausdorff distance with IscA. We selected 109 proteins from different species by a keyword search in the PDB database. The keywords used were the description of properties similar to IscA, such as metal binding and ion transport. The PDB IDs of 109 proteins are listed in the [supplementary material](#). The first step was to screen a small number of proteins with a similar structure as IscA among these 109 proteins. We retained 17 proteins having (3D) Yau–Hausdorff distance with IscA less than 3 Å which means they have similar structures with IscA based on a large number of examples. Thus, we focused on the properties of the 17 proteins. We found that seven of them are related to iron–sulfur cluster and three of them, though lacking in sequence similarity with IscA, work with IscA in the same biological process. The (3D) Yau–Hausdorff distances between IscA and these three proteins are shown in [Table 3](#). The distance between IscA and OsCnfU-1A domain I is the smallest. The other two distances are also less than 3 Å which means their structures have more similarity with IscA than most of the 109 selected proteins.

The three protein structures of IscU (PDB ID: 1WFZ), SufA (PDB ID: 2D2A), and OsCnfU-1A domain I (PDB ID: 2JNV) are shown in [Figure 3](#). IscU was reported to work as a scaffold for iron–sulfur cluster assembly together with IscA, and accordingly, both of them got their name (Ollagnier-de-Choudens, Sanakis, & Fontecave, 2004). SufA is a paralogous protein of IscA and works as the scaffold in biosynthesis of iron–sulfur cluster as well. OsCnfU-1A domain I works in *Oryza sativa* chloroplasts as the scaffold on which iron–sulfur cluster assembles (Saio et al., 2007). Therefore, these three proteins work in the same biological pathway, though from different species. Moreover, we performed protein blast to compare the sequence similarity of these three proteins to IscA. All of them have an alignment Expect value larger than 0.57. This means that they do not show a high sequence similarity with IscA. However, they still resemble each other in biological functions, which could be identified by structure comparison using the (3D) Yau–Hausdorff distance. Structure comparison by the (3D) Yau–Hausdorff distance can serve as

a sophisticated method to uncover biological relationships between proteins. We can reasonably deduce that other protein candidates, which have a small (3D) Yau–Hausdorff distance with IscA, might work similar as IscA and some of the candidates might have the same magnetic-sensing property as MagR, IscA's homologous protein in fruit fly.

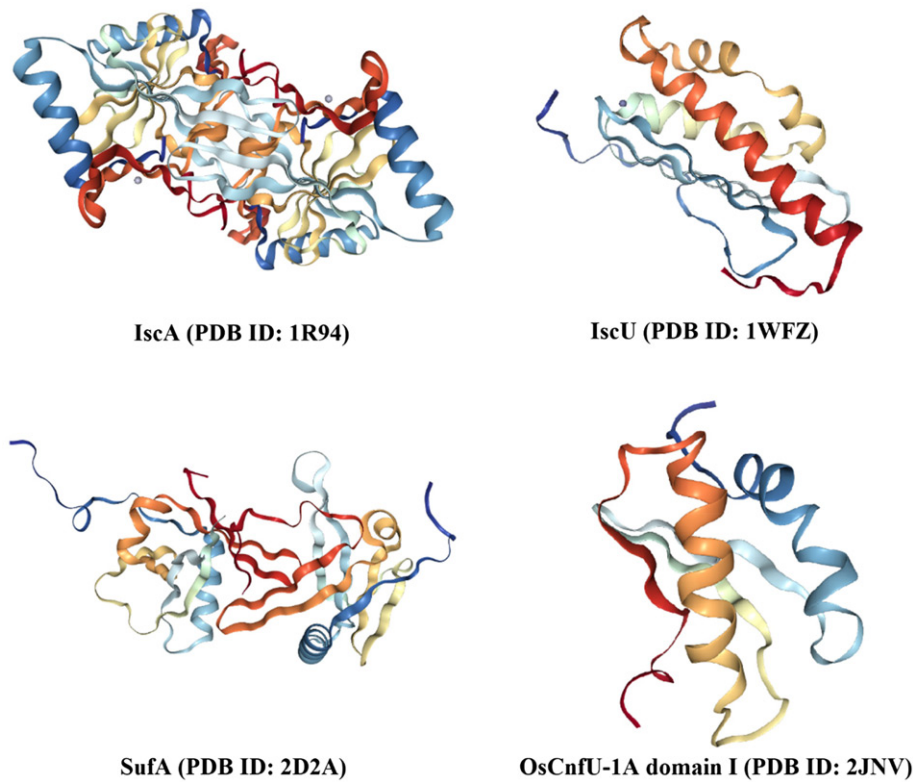
### 3.4. Molecular phylogenetic analysis choosing cytochrome c as the molecular barcode

We also want to see the performance of structure comparison in clustering homologous proteins from different species into phylogenetic trees. However, limited by the number of protein structures solved by now, we could not choose proteins encoded by recognized barcoding genes such as cytochrome c oxidase I (COI) or cytochrome b (Hebert, Ratnasingham, & de Waard, 2003). Considering the number of existing structures solved and the extent of conservatism during evolution, we chose cytochrome c as our first molecular marker (Wilson & Sarich, 1969). After using the (3D) Yau–Hausdorff distance to compare structures of cytochrome c in nine species, we got the distance matrix and constructed the UPGMA phylogenetic tree shown in [Figure 4](#). For comparison, we also used sequence alignment and natural vector method to generate UPGMA phylogenetic trees of the same nine species. For sequence alignment method, we used BLOSUM matrix as the substitution matrix.

The nine species chosen varied from bacteria to mammals, which is a quite big span in evolutionary distance. From the result, we could see that there are consistent as well as discrepancies among three phylogenetic trees generated by three differently defined distances. Mouse and horse were clustered into a branch in all three methods. Two kinds of tuna were clustered into the same branch in the (3D) Yau–Hausdorff and sequence alignment methods. In the phylogenetic tree generated by (3D) Yau–Hausdorff method, species in the same branch belonged to the same classes but four mammals were separated into two remote branches. Compared to the tree generated by sequence alignment, in which bacteria fish and mammal were in a bottom-up order, the (3D) Yau–Hausdorff tree showed poor species hierarchy. However, sequence alignment method had its own mistake that cattle was clustered close to bacteria in a low evolution level. That might be caused by the large evolutionary distances among the species chosen. Therefore, in general, phylogenetic analysis based on structure comparison calculated by the (3D) Yau–Hausdorff distance choosing cytochrome c as the molecular barcode could realize the basic cluster of species. Compared to sequence alignment method, each of them has its own disadvantages. Structure comparison could classify species into clusters well, and sequence alignment is better at displaying evolutionary hierarchy.

### 3.5. Molecular phylogenetic analysis choosing $\beta$ globin as the molecular barcode

We also selected  $\beta$  globin which is a subunit of hemoglobin as another molecular marker. After using the (3D)



**Figure 3.** Protein structures of IscA (PDB ID: 1R94), IscU (PDB ID: 1WFZ), SufA (PDB ID: 2D2A), and OsCnfU-1A domain I (PDB ID: 2JNV). IscU, SufA, and OsCnfU-1A domain I have similar structures to IscA, which are proved by small (3D) Yau–Hausdorff distances. Though they lack similarity in amino acid sequences, these three proteins were all reported to work in the same biological process.

Yau–Hausdorff distance to compare structures of  $\beta$  globin in nine species, we got the distance matrix and constructed the UPGMA phylogenetic tree shown in Figure 5. Sequence alignment and natural vector method were used again for comparison. In sequence alignment method, we used BLOSUM matrix as the substitution matrix.

The nine species chosen this time did not vary a lot, from fish to mammal. Thus, under this circumstance, sequence alignment worked well. In its phylogenetic tree, two species belonged to *Perciformes* order which were clustered under the same branch; cattle and goat from *Bovidae* family were clustered into the same branch; wolf and dog from *Canidae* family were clustered into the same branch, and the evolutionary hierarchy performed well. The tree of (3D) Yau–Hausdorff structure comparison was highly similar to the tree of sequence alignment. The only differences in all lay in details. Wolf and dog were close but not in the same branch. Cattle and goat were closed but not in the same branch either. We could see that when it comes to analysis among closely related species, structure comparison could cluster species generally well but it does not have a high resolution. In order to have a high resolution, we need to use many 3D structures of the same protein. However, even using only one protein structure for structure comparison, one could uncover relationship between distant species. Structure comparison and sequence comparison might complement each other in phylogenetic analysis. The high resolution of sequence alignment could identify kinship between closely related species well, and structure comparison could uncover relationship between distant species.

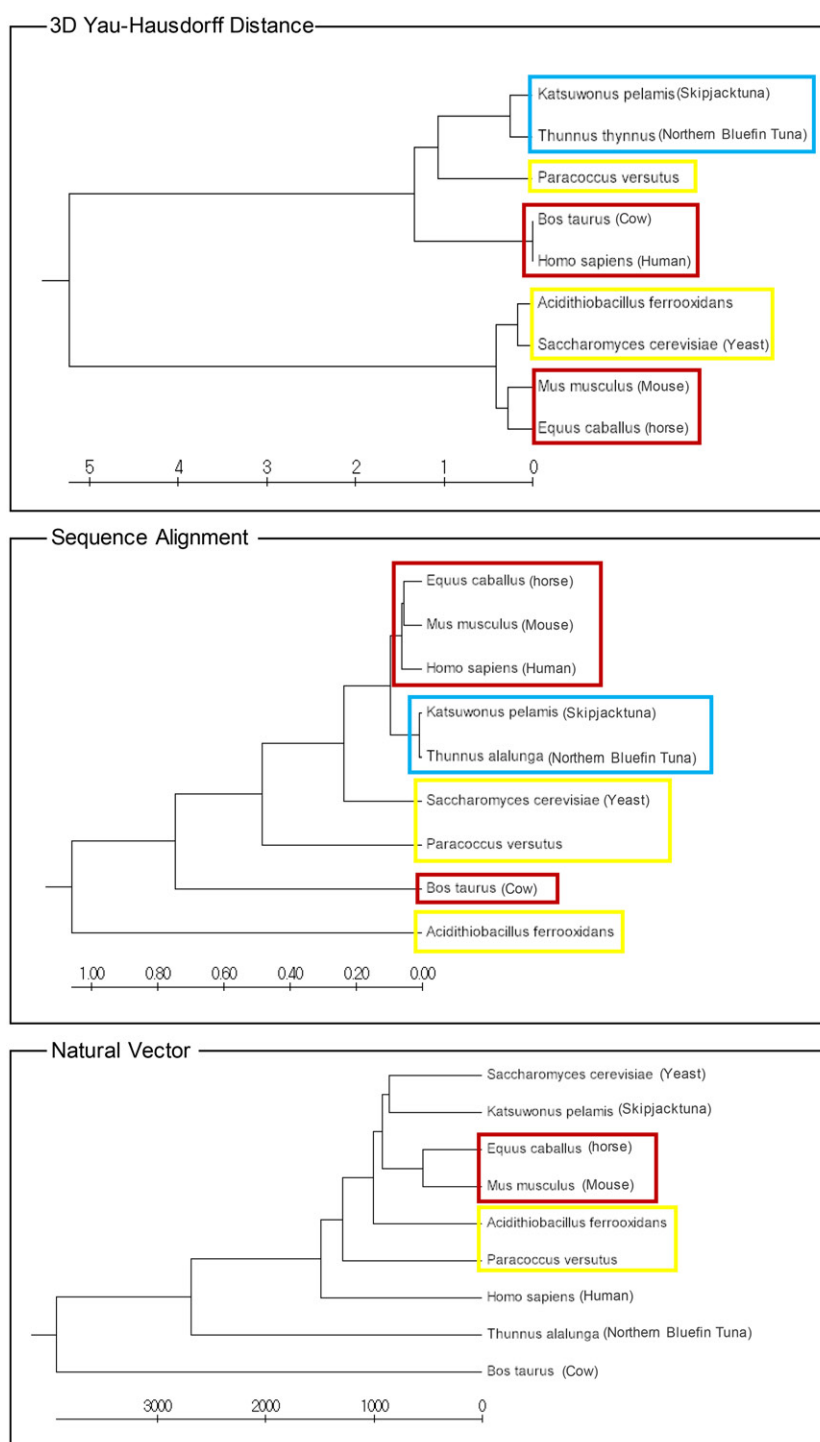
## 4. Discussion and conclusions

### 4.1. Complexity analysis of the (3D) Yau–Hausdorff distance

Given two protein structures with a number of atoms  $m$  and  $n$ , the computational complexity of the (3D) Yau–Hausdorff distance between two structures is  $O(m+n)$  times the minimum 1D Hausdorff distance, which is the same as that of the (2D) Yau–Hausdorff distance (Tian et al., 2015). For two sets of  $m$  and  $n$  points, the complexity of their 1D minimum Hausdorff distance is  $O((m+n)\log(m+n))$  based on Li’s algorithm (Li, Shen, & Li, 2008), and the complexity of our algorithm is  $O((m+n)^2\log(m+n))$ . No other methods are able to calculate the precise value of traditional (3D) Hausdorff distance under rigid motion. The complexity of our method is even lower than  $O((m+n)^5\log^2(mn))$  which is the complexity of (2D) Hausdorff distance under rigid motion (Chew et al., 1997). Thus, the (3D) Yau–Hausdorff distance method significantly decreases computational complexity by a descending dimension without losing information of the structure and makes a great improvement than traditional (3D) Hausdorff distance.

### 4.2. The choice of rotation number and analysis of protein length for calculating the (3D) Yau–Hausdorff distance

We discuss how many rotations are appropriate and enough to obtain a stable result for computing the (3D) Yau–Hausdorff distance of two structures. We set up a



**Figure 4.** UPGMA phylogenetic tree constructed using cytochrome c as the molecular barcode. The results are based on distance matrix calculated by (A) the (3D) Yau–Hausdorff distance, (B) sequence alignment using BLOSUM substitution matrix, and (C) natural vector method. Species in the yellow frame are fungi. Species in the red frame are mammals. Species in the blue frame are fish.

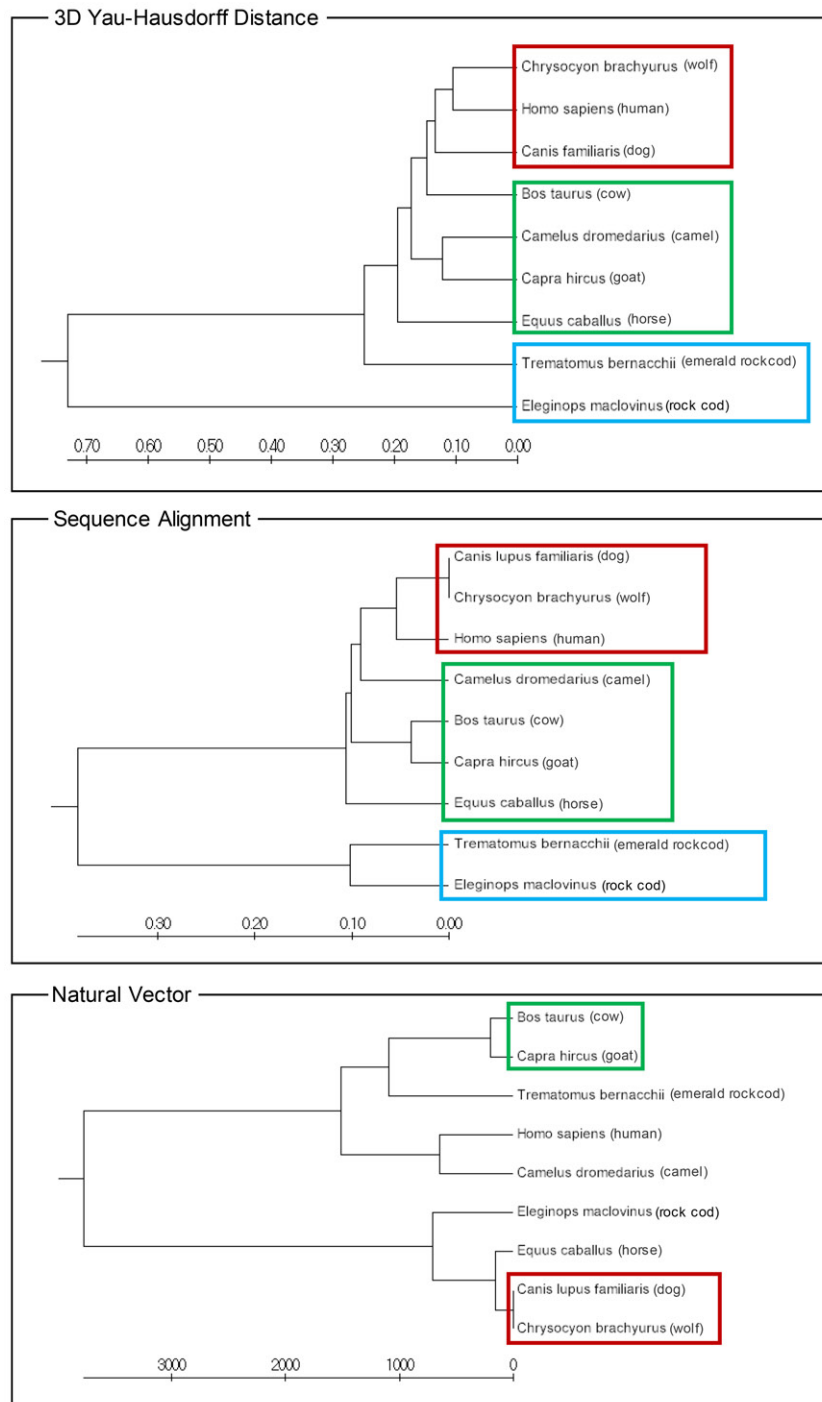
control experiment here to show that 50 is enough for the rotation number of the set  $M$  and  $N$  of two point sets  $A$  and  $B$ .

Firstly, we explain that the specific matrix of  $50 \times 50$  theta and phi values results in 2500 rotation positions. Let  $M = \{\theta_1^3, \theta_2^3, \dots, \theta_u^3\}$ ,  $N = \{\varphi_1^3, \varphi_2^3, \dots, \varphi_v^3\}$  be two sets containing  $u$  and  $v$  3D rotations, respectively. Fixing each rotation  $\theta^3$  in  $M$  of the point set  $A$ ,  $v$  rotations of the point set  $B$  are considered and the minimum value of  $H^1(P_x(A^{\theta^3}), P_x(B^{\varphi^3}))$  is

computed. Thus, for the grid of  $u$  theta and  $v$  phi values,  $D_1$  is calculated according to  $uv$  relative rotation positions of the original 3D point sets  $A$  and  $B$ . Thus, 50 theta and 50 phi values result in 2500 rotation positions.  $D_2$  is analogous to the above steps.

We constructed 10 point sets in a ball with radius 50, and each set contained 200 points randomly generated in this ball. The (3D) Yau–Hausdorff distances between each pair of the 10 sets were calculated by choosing the number of

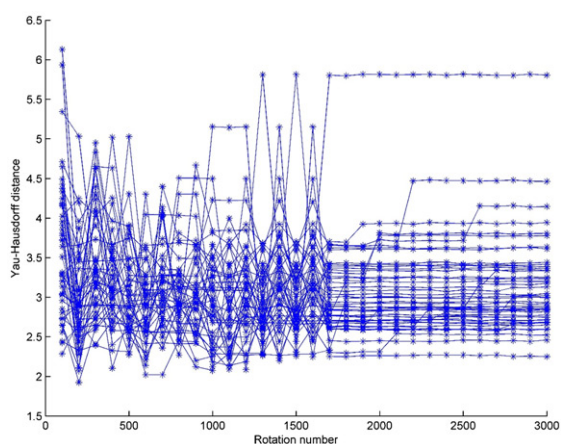




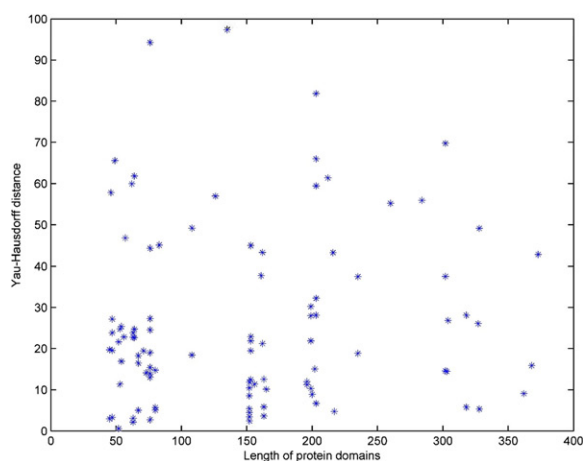
**Figure 5.** UPGMA phylogenetic tree constructed using  $\beta$  globin as the molecular barcode. The results are based on distance matrix calculated by (A) the (3D) Yau–Hausdorff distance, (B) sequence alignment using BLOSUM substitution matrix, and (C) natural vector method. Species in green frame are herbivores. Species in blue frame are fish. Species in red are omnivorous.

rotation positions from 100 to 3000. We drew the curves of the relation between distance value and rotation number for each pair of point sets. Since there are 45 pairs of distances among the 10 point sets, a total of 45 curves are shown in Figure 6. The rotation number increases 100 each time, and the corresponding distances are marked with an asterisk. In this figure, we can see that most pairs of distances achieve a stability when rotation number reaches about 2000. Therefore, 2500 rotations are enough to get a stable distance.

On the other hand, we analyze the (3D) Yau–Hausdorff distance distribution by randomly choosing 100 distance values of the 260 protein domains dataset used in this study shown in Figure 7. The  $x$  coordinate represents the minimum length of two protein domains and the  $y$  coordinate shows their (3D) Yau–Hausdorff distance. In this figure, we can see that the 100 values distribute randomly which have no direct relation with respect to the protein length. It indicates that protein length will not influence the (3D) Yau–Hausdorff distance.



**Figure 6.** The 45 (3D) Yau-Hausdorff distance curves of 10 point sets randomly generated in a ball with radius 50. Each curve represents the relation between distance and rotation number for one pair of point sets. The rotation number increases 100 each time, and the corresponding distances are marked with an asterisk. The distances achieve a stability when the rotation number reaches about 2000.



**Figure 7.** The protein lengths and (3D) Yau-Hausdorff distances of 100 randomly choosing pairs from the 260 protein domains dataset. All the 100 values perform randomly with no direct relation regarding the protein length.

### 4.3. The possible applications of structure comparison

In this study, we presented the possibility of using structure comparison to discover proteins with similar functions based on structural similarity. Given a protein with a specific function, we could use structure comparison to find more protein candidates with the same or similar function. This could be applied on a broad range of fields, and one of them might be the field of gene editing. The enzyme with the powerful function of gene splicing in the famous CRISPR-Cas9 system, Cas9, was selected by screening proteins with a sequence similarity to other endonucleases (Mali et al., 2013). A newly published protein with an even more powerful function, Argonaute, was found by a similar approach (Gao, Shen, Jiang, Wu, & Han, 2016). As structure is directly related to protein function, structure comparison may work better than sequence comparison in this circumstance. Structure comparison may provide more protein candidates with less sequence similarity but more functional similarity. Conversely, given a protein with unknown property, we could predict its possible functions by referring to known

proteins with similar structures. In this way, we would be able to predict protein functions in a more accurate high-throughput way.

### 4.4. Room for improvement of the accuracy of phylogenetic analysis based on structure comparison

When using homologous protein structures to measure the evolutionary distance between species, the structures should be the natural state of proteins in theory. However, proteins do not maintain one static shape in biological processes. Most proteins function in a dynamic manner. Under this circumstance, the structure that could represent one protein best is a series of dynamic conformations. Structure comparison between two proteins should be the Hausdorff distance between two sets of dynamic structures, where pairwise structure comparison is calculated by the (3D) Yau-Hausdorff distance. Cryo-EM method could solve multiple dynamic structures of a protein at the same time but only around 1% of the PDB data was solved by EM. Most proteins only have one crystal structure solved. In our analysis, we dealt with the multiple structures of  $\beta$  globin in human as different dynamic conformations since it was solved multiple times. If the dynamic conformations of other proteins were also available, the resolution of structure comparison might be improved.

### Acknowledgements

The authors wish to thank the Department of Mathematical Science at Tsinghua University for providing the work space and library facilities.

### Disclosure statement

No potential conflict of interest was reported by the authors.

### Funding

This study is supported by the National Natural Sciences Foundation of China [91746119], Tsinghua University startup fund. The funders did not take part in study design; in collection and analysis of data; in the writing of the manuscript; and in the decision to publish this manuscript.

### References

- Bilder, P. W., Ding, H., & Newcomer, M. E. (2004). Crystal structure of the ancient, Fe-S scaffold IscA reveals a novel protein fold. *Biochemistry*, 43(1), 133–139. DOI: 10.1021/bi035440s.
- Chew, L., Goodrich, M., Huttenlocher, D., Kedem, K., Kleinberg, J., & Kravets, D. (1997). Geometric pattern matching under Euclidean motion. *Computational Geometry*, 7(1–2), 113–124.
- Deng, M., Yu, C., Liang, Q., He, R., & Yau, S. S.-T. (2011). A novel method of characterizing genetic sequences: Genome space with biological distance and applications. *PLoS One*, 6, 1–9. DOI: 10.1371/journal.pone.0017293.
- DePristo, M. A., Weinreich, D. M., & Hartl, D. L. (2005). Missense meanderings in sequence space: A biophysical view of protein evolution. *Nature Reviews Genetics*, 6(9), 678–687. DOI: 10.1038/nrg1672.

- Gao, F., Shen, X. Z., Jiang, F., Wu, Y., & Han, C. (2016). DNA-guided genome editing using the *Natronobacterium gregoryi* Argonaute. *Nature Biotechnology*, *34*(7), 768–773. DOI: 10.1038/nbt.3547.
- Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(Suppl\_1), S96–S99. DOI: 10.1098/rsbl.2003.0025.
- Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*, 10915–10919.
- Holm, L., & Sander, C. (1996). Mapping the protein universe. *Science*, *273*, 595–603.
- Huang, H. H., Yu, C., Zheng, H., Hernandez, T., Yau, S. C., He, R., ... Yau, S. S.-T. (2014). Global comparison of multiple-segmented viruses in 12-dimensional genome space. *Molecular Phylogenetics and Evolution*, *81*, 29–36. DOI: 10.1016/j.ympev.2014.08.003.
- Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *34*(5), 827–828. DOI: 10.1107/S0567739478001680.
- Li, B., Shen, Y., & Li, B. (2008). A new algorithm for computing the minimum Hausdorff distance between two point sets on a line under translation. *Information Processing Letters*, *106*(2), 52–58. DOI: 10.1016/j.ipl.2007.10.003.
- Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., ... Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science*, *339*(6121), 823–826. DOI: 10.1126/science.1232033.
- Mishra, A., Rana, P. S., Mittal, A., & Jayaram, B. (2014). D2N: Distance to the native. *Biochimica et Biophysica Acta*, *1844*(10), 1798–1807. DOI: 10.1016/j.bbapap.2014.07.010.
- Mishra, A., Rao, S., Mittal, A., & Jayaram, B. (2013). Capturing native/native like structures with a physico-chemical metric (pcSM) in protein folding. *Biochimica et Biophysica Acta*, *1834*(8), 1520–1531. DOI: 10.1016/j.bbapap.2013.04.023.
- Mittal, A., & Acharya, C. (2012). Extracting signatures of spatial organization for biomolecular nanostructures. *Journal of Nanoscience and Nanotechnology*, *12*(11), 8249–8257. DOI: 10.1166/jnn.2012.6732.
- Mittal, A., & Jayaram, B. (2011). Backbones of folded proteins reveal novel invariant amino acid neighborhoods. *Journal of Biomolecular Structure and Dynamics*, *28*(4), 443–454. DOI: 10.1080/073911011010524954.
- Mittal, A., Jayaram, B., Shenoy, S., & Bawa, T. S. (2010). A stoichiometry driven universal spatial organization of backbones of folded proteins: Are there Chargaffs rules for protein folding?. *Journal of Biomolecular Structure and Dynamics*, *28*(2), 133–142. DOI: 10.1080/07391102.2010.10507349.
- Noel, M. D., & Natasa, P. (2014). GR-Align: Fast and flexible alignment of protein 3D structures using Graphlet degree similarity. *Bioinformatics*, *30*, 1259–1265. DOI: 10.1093/bioinformatics/btu020.
- Ollagnier-de-Choudens, S., Sanakis, Y., & Fontecave, M. (2004). SufA/IscA: Reactivity studies of a class of scaffold proteins involved in [Fe-S] cluster assembly. *JBIC Journal of Biological Inorganic Chemistry*, *9*(7), 828–838. DOI: 10.1007/s00775-004-0581-9.
- Prlc, A., Bliven, S., Rose, P. W., Bluhm, W. F., Bizon, C., Godzik, A., & Bourne, P. E. (2010). Pre-calculated protein structure alignments at the RCSB PDB website. *Bioinformatics*, *26*(23), 2983–2985. DOI: 10.1093/bioinformatics/btq572.
- Saio, T., Kumeta, H., Ogura, K., Yokochi, M., Asayama, M., Katoh, S., ... Inagaki, F. (2007). The cooperative role of OsCnfU-1A Domain I and Domain II in the iron-sulphur cluster transfer process as revealed by NMR. *Journal of Biochemistry*, *142*(1), 113–121. DOI: 10.1093/jb/mvm120.
- Smith, J. M. (1970). Natural selection and the concept of a protein space. *Nature*, *225*, 563–564. DOI: 10.1038/225563a0.
- Tian, K., Yang, X., Kong, Q., Yin, C., He, R., & Yau, S. S.-T. (2015). Two dimensional Yau-Hausdorff distance with applications on comparison of DNA and protein sequences. *PLoS One*, *10*, 1–19. DOI: 10.1371/journal.pone.0136577.
- Tian, K., Zhao, X., & Yau, S. S.-T. (2018). Convex hull analysis of evolutionary and phylogenetic relationships between biological groups. *Journal of Theoretical Biology*, *456*, 34–40. DOI: 10.1016/j.jtbi.2018.07.035.
- Wilson, A. C., & Sarich, V. W. (1969). A molecular time scale for human evolution. *Proceedings of the National Academy of Sciences*, *63*, 1088–1093.
- Yu, C., Deng, M., Cheng, S. Y., Yau, S. C., He, R., & Yau, S. S.-T. (2013). Protein space: A natural method for realizing the nature of protein universe. *Journal of Theoretical Biology*, *318*, 197–204. DOI: 10.1016/j.jtbi.2012.11.005.
- Yu, C., Hernandez, T., Zheng, H., Yau, S. C., Huang, H. H., He, R., ... Yau, S. S.-T. (2013b). Real time classification of viruses in 12 dimensions. *PLoS One*, *8*, 1–10. DOI: 10.1371/journal.pone.0064328.
- Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*(4), 702–710. DOI: 10.1002/prot.20264.
- Zhao, X., Tian, K., He, R., & Yau, S. S.-T. (2017). Establishing the phylogeny of *Prochlorococcus* with a new alignment-free method. *Ecology and Evolution*, *7*(24), 11057–11065. DOI: 10.1002/ece3.3535.
- Zhao, X., Wan, X., He, R., & Yau, S. S.-T. (2016). A new method for studying the evolutionary origin of the SAR11 clade marine bacteria. *Molecular Phylogenetics and Evolution*, *98*, 271–279. DOI: 10.1016/j.ympev.2016.02.015.