BIOINFORMATICS ORIGINAL PAPER
Vol. 26 no. 7 2010, pages 889–895
doi:10.1093/bioinformatics/btq066

Structural bioinformatics
Advance Access publication February 17, 2010

How significant is a protein structure similarity with TM-score = 0.5?

Jinrui Xu[1,2] and Yang Zhang[1,2,*]

[1]Department of Medical School, Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109 and [2]Department of Molecular Biosciences, Center for Bioinformatics, University of Kansas, 2030 Becker Drive, Lawrence, KS 66047, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Protein structure similarity is often measured by root mean squared deviation, global distance test score and template modeling score (TM-score). However, the scores themselves cannot provide information on how significant the structural similarity is. Also, it lacks a quantitative relation between the scores and conventional fold classifications. This article aims to answer two questions: (i) what is the statistical significance of TM-score? (ii) What is the probability of two proteins having the same fold given a specific TM-score?

Results: We first made an all-to-all gapless structural match on 6684 non-homologous single-domain proteins in the PDB and found that the TM-scores follow an extreme value distribution. The data allow us to assign each TM-score a P-value that measures the chance of two randomly selected proteins obtaining an equal or higher TM-score. With a TM-score at 0.5, for instance, its P-value is 5.5×10^{-7}, which means we need to consider at least 1.8 million random protein pairs to acquire a TM-score of no less than 0.5. Second, we examine the posterior probability of the same fold proteins from three datasets SCOP, CATH and the consensus of SCOP and CATH. It is found that the posterior probability from different datasets has a similar rapid phase transition around TM-score = 0.5. This finding indicates that TM-score can be used as an approximate but quantitative criterion for protein topology classification, i.e. 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.

Contact: zhng@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 11, 2009; revised on January 20, 2010; accepted on February 12, 2010

1 INTRODUCTION

Protein structure comparison is essential in almost every aspect of modern structural biology, ranging from experimental protein structure determination to computer-based protein folding and structure prediction, from protein topology classification to structure-based protein function annotation and from protein–ligand docking to new compound screening and drug design (Kuntz, 1992; Murzin *et al.*, 1995; Orengo *et al.*, 1997; Zhang, 2009). The most commonly used means to compare protein structures is to calculate the root mean squared deviation (RMSD) of all the equivalent atom pairs after the optimal superposition of the two structures (Kabsch, 1978). However, because all atoms in the structures are equally weighted in the calculation, one of the major drawbacks of RMSD is that it becomes more sensitive to the local structure deviation than to the global topology when the RMSD value is big. For example, the RMSD of two protein structures can be high if the tails or some loops have a different orientation even though the global topology of the core part is the same; this cannot be distinguishable, based on the RMSD value alone, from the case where two structures have completely different topologies.

Aiming at developing protein topology-sensitive measures, Zemla *et al.* proposed a global distance test score (GDT-score), which counts the number of Cα pairs which have a distance <1, 2, 4 and 8 Å after the optimal superposition (Zemla *et al.*, 1999; Zemla, 2003). Similarly, Siew *et al.* proposed MaxSub to identify the maximum substructures which have Cα pairs <3.5 Å (Siew *et al.*, 2000). These measurements were extensively used in the community-wide CASP and CAFASP experiments for assessing the modeling accuracy of protein structure predictions (Fischer *et al.*, 2003; Moult *et al.*, 2007; Zemla *et al.*, 1999). However, the distance cutoffs in both GDT and MaxSub scores are subjective and may need to be manually tuned for different categories of modeling targets (Kopp *et al.*, 2007). Moreover, similar to RMSD, the magnitude of the GDT and MaxSub scores for random structure pairs has a power-law dependence with the protein length (Zhang and Skolnick, 2004), which renders the absolute value of the scores less meaningful. Some structural alignment-based scores, e.g. the MAMMOTH score (Ortiz *et al.*, 2002) and the Dali Z-score (Holm and Sander, 1995), have also been often exploited to access the accuracy of protein structure prediction. However, these measurements neglect the alignment accuracy of the structure modeling. For example, a structure model built on a template protein with a wrong alignment will have the same structural alignment scores as the model on the same template but with a correct alignment. These scores also have the drawback that the similarity of related proteins strongly depends on their length (Pascual-Garcia *et al.*, 2010).

To address these issues, Zhang and Skolnick recently developed a template modeling score (TM-score) (Zhang and Skolnick, 2004), which counts all residue pairs using the Levitt–Gerstein weight (Levitt and Gerstein, 1998) and therefore does not need discrete distance cutoffs. Since the short distance in the Levitt–Gerstein matrix is weighted stronger than the long distance, the TM-score

*To whom correspondence should be addressed.

is more sensitive to the global topology than local variations. Additionally, because it adopts a protein size-dependent scale to normalize the residue distances, the magnitude of the TM-score for random protein pairs is protein size independent (Zhang and Skolnick, 2004).

Despite the advantage and usefulness of RMSD, GDT-, MaxSub- and TM-scores as quantitative measures of protein structure similarities, the scores themselves cannot quantify the statistical significance of the structure superposition/alignment, which is essential in many of the statistical studies of protein structure comparisons and alignment analyses (Levitt and Gerstein, 1998; Sadreyev *et al.*, 2009). Although MAMMOTH and Dali provide P-value or Z-score as a measurement of the significance of sequence-independent structural alignments, they do not appropriately count for the similarities of given alignments that is important in the assessment of protein structure predictions. Another important issue is that these scores do not quantify the probability of two structures sharing same or different folds/topologies. Proteins have been categorized into various structural families based on the structural and/or evolutionary similarities, using either human visual intuition (Murzin *et al.*, 1995) or semi- or fully automated structural comparisons (Holm *et al.*, 1992; Orengo *et al.*, 1997). These hierarchical databases provide important insights to our understanding of protein structure and function, and help gauge the field of structural comparisons and categorization. However, it generally lacks a quantitative correspondence between the structural similarity scores and the various levels of protein structure categorizations. For example, a simple but often-asked question in protein structure prediction and assessment is: does the predicted model have the correct fold (compared with the native structure) given the RMSD, GDT-score and TM-score?

In this work, we try to address these issues by answering two questions: (i) what is the statistical significance of each TM-score value; and (ii) what is the probability of two proteins having the same fold given the TM-score. Here, the reason for focusing on the TM-score is that its magnitude is protein size independent, which facilitates the attainment of length-independent analytical results of the calculations. Although our focus in the second question is on the fold level, the results can be easily extended to other levels of structural similarities.

2 METHODS AND METERIALS

2.1 Definition of TM-score

The TM-score is defined to assess the topological similarity of two protein structures (Zhang and Skolnick, 2004):

$$\text{TM-score} = \frac{1}{L}\left[\sum_{i=1}^{L_{\text{ali}}} \frac{1}{1+d_i^2/d_0^2}\right]_{\max} \qquad (1)$$

where L is the length of the target protein, and L_{ali} is the number of the equivalent residues in two proteins. d_i is the distance of the i-th pair of the equivalent residues between the two structures, which depends on the superposition matrix; the ‘max’ means the procedure to identify the optimal superposition matrix that maximizes the sum in Equation (1). The scale $d_0 = \sqrt[3]{L-15} - 1.8$ is defined to normalize the TM-score in a way that the magnitude of the average TM-score for random protein pairs is independent on the size of the proteins. TM-score stays in (0, 1] with a higher value indicating a stronger similarity.

2.2 Dataset of random protein structure pairs

A total of 6684 single-domain structures were culled from the PDB database (Berman *et al.*, 2002). These proteins share low pairwise sequence similarities (with sequence identity <25%), as filtered by PISCES (Wang and Dunbrack, 2003), with protein lengths between 80 and 200 amino acids. We neglect proteins below 80 residues because they typically have relatively simple topologies. Proteins larger than 200 residues are mainly multiple-domain proteins.

A total of 22 334 586 protein pairs are formed by an all-to-all combination of the 6684 proteins [=6684*(6684 - 1)/2]. The shorter protein in each pair is used as the target protein in TM-score calculations. To increase the size of the statistical sample, for each protein pair, the target protein is first superposed by the TM-score program on the N-terminal structure of the bigger protein structure with the TM-score normalized by the target protein. Then, the target protein slides gaplessly along the sequence of the bigger protein with a window size of 20 residues until less than 20 amino acids remain on the larger protein. A TM-score is obtained from each of the gapless alignments formed in the sliding process; each gapless alignment is counted as an independent structure pair. This procedure on the dataset ends up with a total of 71 583 085 random and protein-like structure pairs.

It should be mentioned that the TM-score superimpositions are obtained from a set of gapless sliding alignments rather than from the optimal structural alignments of the two proteins. The purpose of the gapless alignment is to generate a random structure background, because a structural alignment, produced by tools such as Dali (Holm and Sander, 1995) and TM-align (Zhang and Skolnick, 2005), usually represents an optimal match of a given pair of protein structures that is selected from a huge number of possible combinations of corresponding residues assignments. Thus, a structural alignment (with optimal corresponding residues assigned) does not constitute random structural comparisons even though the non-homologous protein pairs are randomly selected from PDB.

2.3 Dataset of proteins with same/different folds

To estimate the posterior probability for structure pairs at given TM-scores sharing the same topology, a collection of protein pairs in both the same and the different folds is necessary. For this purpose, we borrow the Fold and Topology definition from the standard protein structure classification databases: SCOP (Andreeva *et al.*, 2008) and CATH (Cuff *et al.*, 2009) to generate the same and different fold datasets.

2.3.1 Three sets of same fold structure pairs The first set of protein domains (Set-I) are collected from the SCOP 1.73 database. After filtering out the redundant proteins with a sequence identity >95% and the small proteins with length below 80 amino acids, 11 239 protein domains remain, which cover 551 main Fold families in SCOP. An all-to-all pairing is then carried out for the proteins within the same Fold family and ends up with a total of 746 420 protein pairs which are considered as sharing same folds in SCOP.

The second set of protein domains (Set-II) are from CATH 3.2.0. The structure pairs are generated from the proteins in the same ‘Topology’, a structural level equivalent to the ‘Fold’ in SCOP (Hadley and Jones, 1999). After the same redundancy and length filtering, 14 830 domains covering 700 main Topologies in CATH are obtained. An all-to-all pairing among proteins of the same Topology families results in 2 769 868 domain pairs. The reason for Set-II being much bigger than Set-I is due to the fact that some CATH families have a dominantly large size.

The third protein pair set (Set-III) is a consensus of the SCOP and CATH databases where the proteins are of the same fold in both SCOP and CATH. Due to the different domain splitting system, SCOP and CATH may have protein domains with the same ID (the same PDB names and chains) but having different sequence segments. To ensure that SCOP and CATH deal with the same structures, we filter out those inconsistent domains and collect only the structures which have the same IDs in the SCOP and CATH and meanwhile have the identical regions covering >90% of both the SCOP and

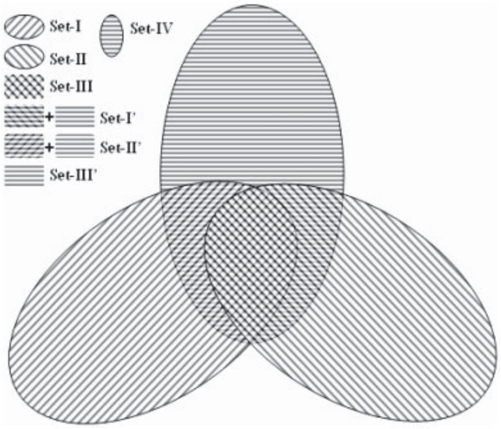

Fig. 1. Venn diagram of datasets of the same/different folds. Set-I contains 746 420 same Fold domain pairs generated from 11 239 protein domains in SCOP. Set-II consists of 2 769 868 same Topology domain pairs generated from 14 830 protein domains in CATH. Set-III is the overlap part of Set-I and Set-II, which includes 186 359 pairs from 5105 consensus domains. Set-IV contains 13 027 960 all-to-all pairs from the 5105 consensus domains. Set-I′ is the different fold set for SCOP, generated by subtracting a subset of Set-I from Set-IV. Set-II′ is the different fold set for CATH, generated by subtracting a subset of Set-II from Set-IV. Set-III′ is the different fold set for Set-III and obtained by subtracting subsets of Set-I and Set-II from Set-IV.

CATH domains. By these criteria, 5105 domain structures are culled from SCOP with a counterpart in CATH, which cover 328 different fold families. An all-to-all pairing is carried out among the proteins which are consistently defined by SCOP and CATH as being of the same fold, resulting in 186 359 protein pairs.

2.3.2 Three sets of different fold structure pairs There are three sets of different fold protein pairs corresponding to the same fold pairs in Set-I, II and III. Due to the big size of the protein sets, we found that the TM-score distributions for different fold proteins are very similar between these sets. Therefore, we generated all the different fold protein pairs from the well-defined consensus set of the 5105 protein domains from Set-III.

The first different fold protein set is named Set-I′. It contains all-to-all pairings of the 5105 protein domains (named as Set-IV, 13 027 960 pairs) but excludes all the pairs that are in the same SCOP Fold family (a subset of Set-I), which results in 12 815 737 protein pairs. The Set-II′ is similar to Set-I′ but excludes from Set-IV the domain pairs that are in the same CATH Topology family (a subset of Set-II), which results in 11 508 804 protein pairs. To generate Set-III′ from Set-IV, any pairs which are either in the same SCOP Fold family or in the same CATH Topology family are excluded. This results in 11 506 777 protein pairs. Figure 1 is a Venn diagram to illustrate the generation of all the datasets in this section.

3 RESULTS

3.1 Statistical significance of TM-score

Extreme value distribution (EVD) is often used to model the smallest or largest value among a large set of independent, identically distributed random values (Embrechts *et al.*, 1997). It has been shown that both sequence and structure comparison scores of proteins follow the EVD (Levitt and Gerstein, 1998). The general function of EVD is described as

$$y=f\left(x|\mu,\sigma\right)=\sigma^{-1}\exp\left(\frac{\mu-x}{\sigma}\right)\exp\left(-\exp\left(\frac{\mu-x}{\sigma}\right)\right) \qquad (2)$$

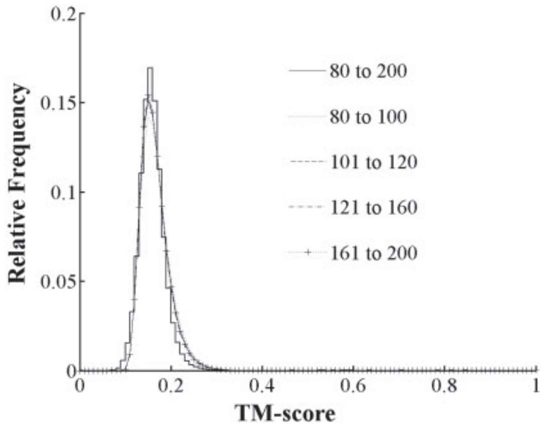

Fig. 2. TM-score distribution of 71 583 085 gapless comparisons among 6684 non-homologous protein structures. The continuous curve represents an EVD with the location parameter and the scale parameter being 0.1512 and 0.0242, respectively; the reduced χ^2 of fitting is 0.001 obtained by the *Evfit* module of MATLAB7 software. The TM-score distributions of four subdivisions are from proteins with length in [80, 100], [101, 120], [121, 160] and [161, 200], respectively.

where μ is the so-called location parameter and σ is the scale parameter.

In Figure 2, we show the distribution of TM-score values calculated from 71 583 085 random protein pairs which are collected from 6684 non-homologous proteins in the PDB library by gapless threading. The distribution matches well to the Equation (2) with the best fitting parameter $\mu=0.1512$ and $\sigma=0.0242$ estimated by the Maximum Likelihood method, which is implemented by the *Evfit* module in MATLAB7; the error tolerance of fitting is 1.0e-6.

We also split the protein samples into four groups according to protein size, i.e. [80, 100], [101, 120], [121, 160], [161, 200]; all of them follow well the same EVD. This data demonstrate the robustness of the EVD for the TM-score distribution of random protein pairs within a gapless alignment match (or random protein matches). Also, the data confirm the previous conclusion that the TM-score magnitude and distribution of random protein pairs are independent of protein size (Zhang and Skolnick, 2004).

We are interested in the probability of having a TM-score equal to or greater than a certain value (x) among random protein pairs, i.e. P-value of a TM-score. The P-value can be obtained by integrating Equation (2) from x to 1, i.e.

$$P\text{-value}(x)=\int_x^1 f(x|\mu,\sigma)\mathrm{d}x=1-\exp\left[-\exp\left(\frac{\mu-x}{\sigma}\right)\right] \qquad (3)$$

Figure 3 shows the curve of the P-value versus TM-score with μ and σ taken from the data in Figure 2. In general, the probability of finding a TM-score ≤ 0.17 from random structural pairs is close to 1. The P-value then decreases rapidly as the TM-score becomes >0.17; it is significantly <1 when TM-score >0.3. In the inset of Figure 3, we plot the P-value for the TM-score range in [0.3, 1], which follows approximately an exponential regression. When TM-score = 0.5, it corresponds to a P-value of 5.5×10^{-7}.

Many authors have demonstrated that the magnitude of RMSD, GDT-score and several other matrices are all protein length dependent (Betancourt and Skolnick, 2001; Levitt and Gerstein,

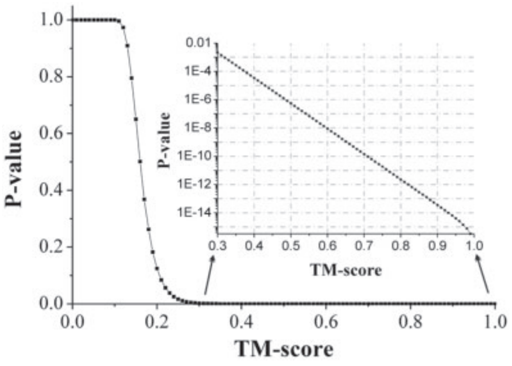

Fig. 3. The P-value versus TM-score. The curve is a sigmoid like Boltzmann function with reduced χ^2 equal to 0.0001. Inset: P-value (in logarithm scale) versus TM-score in [0.3, 1].

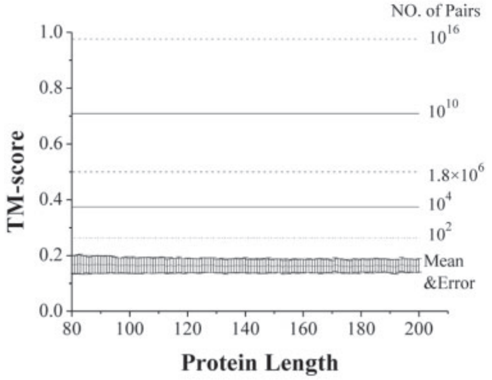

Fig. 4. The average TM-scores (with error bars) of gapless alignment matches on random structural pairs with protein length from 80 to 200 amino acids. The straight and dash lines above TM-scores = 0.2 indicate the number of random protein pairs (values on the right-hand side) needed to achieve or surpass a certain TM-score level. By doing random structure comparisons in 10^2, 10^4, 10^{10} and 10^{16} times, one can hit a match with a TM-score $\geq$0.263, 0.374, 0.709 and 0.977, respectively. 1.8×10^6 random matches are needed to achieve a TM-score $\geq$0.5.

1998; Ortiz *et al.*, 2002; Zhang and Skolnick, 2004). A basic assumption of our work is that the magnitude of TM-score is protein length independent, which enables us to express the P-value as a sole function of TM-score. Figure 4 (the bottom) shows explicitly the average TM-score value and the deviation with proteins of different sizes (from 80 to 200 amino acids). The data again confirm the size independence of the TM-score values in random protein pairs.

As an intuitive explanation of the P-value, we also present in Figure 4 the number of random protein pairs which are needed to achieve or surpass certain TM-score values; this is converted from the P-value data shown in Figure 3. For a TM-score = 0.5, for instance, we need at least 1.8 million random structural matches so that one structure match can hit a TM-score $\geq$0.5. When a TM-score = 0.72, this number increases to 10 billion.

3.2 TM-score of proteins with the same fold

Although the P-value can give a quantitative measure of the statistical significance of each TM-score value, researchers often want to know what TM-score approximately corresponds to the protein pairs sharing the same fold. For example, an often-asked question in *ab initio* and template-based protein structure prediction is how to judge whether a predicted model has the same fold or topology as the experimental structure (Jauch *et al.*, 2007; Kopp *et al.*, 2007; Zhang, 2009). Here, we address this issue by calculating the posterior probability for proteins at certain TM-score having the same or different folds. We will examine the results of the posterior probabilities using three different fold/topology standards.

3.2.1 TM-score corresponding to the SCOP Fold level According to the Bayesian rules, for a given TM-score, the posterior probabilities of proteins having the same or different fold can be expressed as:

$$\begin{cases} P(F|\mathrm{TM}) = \dfrac{P(\mathrm{TM}|F)P(F)}{P(\mathrm{TM}|F)P(F)+P(\mathrm{TM}|\overline{F})P(\overline{F})} \\ P(\overline{F}|\mathrm{TM}) = \dfrac{P(\mathrm{TM}|\overline{F})P(\overline{F})}{P(\mathrm{TM}|F)P(F)+P(\mathrm{TM}|\overline{F})P(\overline{F})} \end{cases} \qquad (4)$$

Here, TM stands for the TM-score of the compared proteins as calculated by the structural alignment program TM-align (Zhang and Skolnick, 2005); F and $\overline{F}$ represent the events that the proteins belong to the same and different Fold in SCOP, respectively; $P(F)$ and $P\left(\overline{F}\right)$ are the prior probabilities of proteins in same and different folds; $P(\mathrm{TM}|F)$ and $P\left(\mathrm{TM}|\overline{F}\right)$ are the conditional probabilities of TM-score when the two proteins are in the same or different Fold families, respectively.

In the Set-I and Set-I′, 746 420 pairs of proteins are considered by SCOP1.73 as the same Fold and 12 815 737 are as not in the same Fold. Thus, the conditional probabilities can be calculated by

$$\begin{cases} P(\mathrm{TM}|F) = \dfrac{N(\mathrm{TM})}{\sum N(\mathrm{TM})} \\ P(\mathrm{TM}|\overline{F}) = \dfrac{\overline{N}(\mathrm{TM})}{\sum \overline{N}(\mathrm{TM})} \end{cases} \qquad (5)$$

where N(TM) is the number of protein pairs with a certain TM-score (TM) in the Set-I, and $\overline{N}(\mathrm{TM})$ is the number of protein pairs with the TM-score in the different fold protein Set-I′. The denominators are the summation of the same and different fold protein pairs for all TM-scores in (0, 1], which equals to the total number of protein pairs in Set-I and Set-I′, respectively.

In Figure 5 (‘filled squares’ and ‘open squares’), we divide the TM-score space into 20 bins and present the conditional probability for both the same and different fold proteins. As expected, the same fold and the different fold proteins are well grouped in two different TM-score ranges, i.e. the same fold proteins have a higher TM-score and the different fold proteins have a lower one. However, since TM-score and SCOP fold do not have a one-to-one correspondence, there is a small overlap of TM-score between the two protein datasets.

To calculate the prior probabilities: $P(F)$ and $P(\overline{F})$, for the purpose of minimizing statistical bias, we collect all the 85 685 protein domains in the SCOP database. An all-to-all pairing is then carried out on these proteins. The prior probabilities can be calculated by

$$\begin{cases} P(F) = \dfrac{N(F)}{N(F)+N(\overline{F})} \\ P(\overline{F}) = 1-P(F) \end{cases} \qquad (6)$$

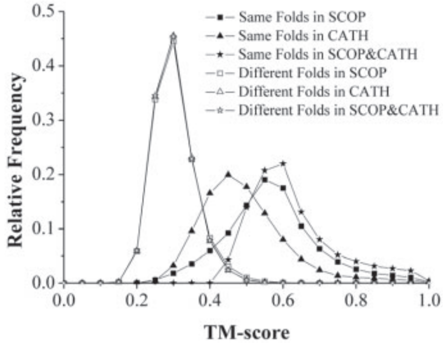

Fig. 5. The conditional probabilities of TM-score for proteins in the same fold and different fold families as defined by SCOP (Set-I; Set-II′), CATH (Set-II; Set-II′) and SCOP and CATH (Set-III; Set-III′).

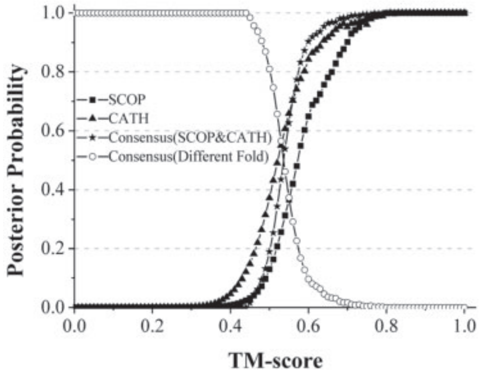

Fig. 6. The posterior probability of proteins with a given TM-score being in the same Fold (squares, triangles and stars points) or different Fold family (circle points). The Fold family is defined based on either the SCOP Fold level (SCOP, Set-I) or the CATH Topology level (CATH, Set-II) or a consensus of SCOP Fold and CATH Topology families (Consensus, Set-III).

where $N(F)$ and $N(\overline{F})$ are, respectively, the numbers of the same fold and the different fold pairs according to the SCOP definition. Overall, $P(F) = 0.0142$ and $P(\overline{F}) = 0.9858$ in our counting.

Figure 6 (‘filled squares’) shows the posterior probability for two proteins with a certain TM-score to be in the same SCOP Fold, which is calculated by integrating the data of Equations (5) and (6) into Equation (4). When TM-score <0.4, there are almost no protein pairs which are in the same SCOP Fold family. On the other hand, when TM-score >0.6, the probability of the two proteins in the same SCOP Fold rapidly increases to >65%. There is a clear phase transition occurring around the half scale of TM-score.

3.2.2 TM-score corresponding to the CATH Topology level Since the fold definition can be dependent on subjective choices, to examine the robustness of the TM-score distribution, we calculate the posterior probability using another widely-used database, CATH (Cuff *et al.*, 2009). A total of 2 769 868 protein pairs are considered by CATH as having the same Topology in Set-II and 11 508 804 pairs as having different Topology in Set-II′.

In Figure 5 (‘filled triangles’ and ‘open triangles’), we show the conditional probabilities of the same and different fold protein pairs in the CATH database. Compared with the SCOP data, there is a clear shift of the distribution of the same fold pairs toward a smaller TM-score value, which indicates that the fold definition in CATH Topology is on average broader than that in SCOP Fold, although the different fold distribution of CATH is similar to that of SCOP. This is consistent with the fact that CATH is dominated by big families and therefore the number of protein pairs in Set-II is much larger than that in Set-I. Correspondingly, the prior probability $P(F)$ of CATH Topology calculated from all the 114 125 CATH domains based on Equation (6) is 0.0299, which is higher than that of SCOP (0.0142), because more protein pairs are categorized into the same Topology families due to the broader structural cutoff in CATH. The prior probability of the different Topology proteins $P(\overline{F}) = 0.97$.

Figure 6 (‘filled triangles’) shows the posterior probabilities of protein pairs being in the same CATH Topology families with given TM-scores. There is a slight shift of CATH compared with SCOP toward smaller TM-score as well; but a similar rapid phase transition is observed in TM-score between 0.4 and 0.6.

3.2.3 TM-score corresponding to the consensus SCOP and CATH fold families Due to the slight inconsistence between SCOP and CATH databases, we next considered a consensus set of protein pairs, Set-III, where the proteins are considered as the same Fold/Topology by both SCOP and CATH, which covers 328 consensus structural families. The different fold protein pairs (Set-III′) are those which are categorized into different structural families by both databases.

As shown in Figure 5 (‘filled stars’), the conditional probabilities of the TM-score for proteins in the same families in the consensus set are slightly shifted toward a larger TM-scores compared with SCOP, because of the even tighter definition of the fold family. Similarly, the prior probability for the same fold $P(F) = 0.0149$ while $P(\bar{F}) = 0.985$ for the different fold proteins.

In Figure 6 (‘filled stars’), we present a posterior probability of proteins at the same Fold and different Fold families based on the consensus definition from both SCOP and CATH. There is again a rapid phase transition around the TM-score = 0.5. Compared with SCOP and CATH, however, this transition is more rapid.

3.2.4 Robustness of posterior probability to structural variations The ‘same Fold’ in the protein structure classifications usually refers to the similar topological arrangements of the secondary structure elements in the core region of the structures (Andreeva *et al.*, 2008; Cuff *et al.*, 2009). However, the calculations in the above sections were made on all protein pairs in the SCOP and CATH families, where some of the same fold domains have long tails and outlier supersecondary/loop structures that have different orientation, despite the similar topology in the core region. In this section, we examine the influence of these outlier structural variations on the posterior probability of TM-score.

We consider two structure filters to refine our structure datasets. First, we exclude from Sets-I, II, III the protein pairs which have a difference in the radius of gyrations (RGs) >10%. This reduces the number of protein pairs in Sets-I, II, III from 746 420, 2 769 868, 186 359 to 449 281, 1 360 782 and 117 446, respectively. In Figure S1A (see Supplementary Materials), we show the conditional probability of the filtered datasets, which has a slight shift toward higher TM-score values, compared with the raw datasets from SCOP, CATH and the consensus, due to the removal of the structure outliers. Supplementary Figure S1B shows the posterior probability of the

TM-score for the filtered datasets, which is almost the same as that calculated from the raw data. The deviation in the posterior probability between the raw and filtered datasets is much smaller than that in the conditional probability. For example, based on Figure 5 and Supplementary Figure S1A, the correlation coefficients of conditional probability curves between filtered and raw ones (Set-I, Set-II, Set-III) are 0.947, 0.969 and 0.926, respectively, while based on Figure 6 and Supplementary Figure S1B, the corresponding correlation coefficients between posterior probabilities increase to 0.999, 1.0 and 0.999. This indicates that posterior probability is indeed less influenced by the changes in the input datasets. This robustness is due to the posterior equation $P(F|\mathrm{TM}) = P(\mathrm{TM}|F)P(F)\big/P(\mathrm{TM})$, where the prior probability from a large base of datasets keeps constant and the change in conditional probability is partially canceled out by the normalization of P(TM) which changes accordingly. Thus, the overall posterior probabilities are less influenced by the filter of the structural outliers.

Second, we detect and remove the long tails at the N- and C-terminals by STRIDE (Frishman and Argos, 1995) from all protein domains in SCOP and CATH families. Thus, the TM-score of each pair may be increased because of the decreasing of the target protein length while the core region alignment in TM-align is unchanged. Supplementary Figures S2A and S2B show the conditional and posterior probabilities for the refined datasets. Cutting the long tails has a relatively larger influence on the conditional probabilities than the first filter from the RG cutoffs. However, both filters have no obvious influence on the posterior probabilities because of the robustness of the prior probability and the normalization effect of the probabilities in the posterior probability calculation.

3.2.5 Summary Combining the results from the three different datasets, as well as bearing in mind the robustness of the posterior probability to the structural variations, it seems quite safe to assign TM-score = 0.5 as a rough but quantitative cutoff for protein Fold/Topology definition, i.e. most of proteins with TM-score >0.5 can be considered as of the same topology whereas most proteins with a TM-score <0.5 should be of different topology. Surely, this cutoff may vary slightly with the different definitions of ‘Same Fold’. For CATH, a TM-score at 0.5 indicates that the structures have a 37% probability being in the same Topology family; when TM-score increases to 0.6, this probability increases to 80%. As for SCOP, a TM-score of 0.5 only corresponds to a 13% probability for the structures to be in the same Fold family; but a TM-score = 0.7 has the posterior probability rapidly increased to 90%. These results are consistent with the conditional probability data in Figure 5, i.e. the CATH Topology level has relatively looser criteria to define the same fold proteins than the SCOP Fold level. When the TM-score is further away from the cutoff value of 0.5, the conclusion becomes gradually safer in all the criterions. When TM-score = 0.4, for example, >99.9% of proteins are not in the same fold according to the consensus definition of SCOP and CATH; when TM-score = 0.6, >90% of proteins are in the same fold based on the consensus criterion.

4 DISCUSSION AND CONCLUSION

We first examined the TM-score distribution of randomly selected non-homologous protein pairs using gapless threading, and found that it follows a simple EVD independent of protein length. This allows us to calculate the P-value to estimate the statistical significance of each TM-score value. When the TM-score <0.17, the P-value is close to 1, which means that any protein structures or computer models at this level of similarity is indistinguishable from random structure pairs. The P-value decreases rapidly <0.001 when TM-score >0.3, a region of structural similarity which is significantly different from random structures. When TM-score = 0.5, the P-value is reduced to 5.5×10^{-7}, meaning that at least 1.8 millions of random protein pairs are needed to achieve on average one of this similarity.

It should be noted that this data does not contradict a previous study (Zhang *et al.*, 2006) where the average TM-score of the structural alignment by TM-align on random structure pairs is ~0.3. For a given pair of proteins, the structure alignment program, TM-align (Zhang and Skolnick, 2005), needs to scan a huge number of possible alignments ($\sim L_1^{L_2}$ in principle, where L_1 and L_2 are the size of proteins) and selects and returns an optimal alignment that corresponds to the highest TM-score. Thus, the average TM-score reported by TM-align for random structure pairs is much higher than the average TM-score from a gapless alignment for the same random structure pairs because the former represents an optimal alignment that is far from a random alignment selection although the structure pair itself is randomly selected. Interestingly, in the recent CASP7 and CASP8 blind protein structural predictions, the average TM-score of the worst three models for each target are 0.161 ± 0.041 and 0.168 ± 0.042, respectively (data taken from http://zhanglab.ccmb.med.umich.edu/casp7 and http://zhanglab.ccmb.med.umich.edu/casp8); both are below and near 0.17. This means that the predicted models from these bottom groups are not more than a random pickup of structures from the PDB library.

Second, we developed an approach for estimating the posterior probability of proteins with given TM-scores to be in the same or different fold family. Using three different datasets which has Fold/Topology defined from the standard SCOP and CATH databases, we consistently observed a similar rapid phase transition of the probability around TM-score = 0.5. This indicates that TM-score may be used as a quantitative criterion for assessing whether protein structures or model predictions are of the same topology, i.e. for TM-score <0.5, proteins are mostly not in the same fold while for TM-score >0.5, proteins are generally in the same fold. This criterion becomes gradually safer when the actual TM-score reaches a value further away from the cutoff.

We also examined the influence of the structural outliers on the TM-score by cutting the long tails or excluding protein pairs of different RGs. Although the conditional probabilities of the TM-score values have marginal changes, the posterior probability is not influenced by the structural outliers mainly because of the canceling-out effect of probabilities in the posterior calculation. A robust phase transition is observed at TM-score = 0.5 in both datasets whether including or excluding the structure outliers.

One of the immediate uses of the quantitative TM-score thresholds is for automated protein structure classifications (Murzin *et al.*, 1995; Orengo *et al.*, 1997). Because of the rapid increase of protein structures accelerated by various proteomic projects, it becomes increasingly infeasible for the human visualization to conduct large-scale protein structure classification, where development of the quantitative scoring functions that corresponds to specific structural similarity levels is the key. Besides, the proposed TM-score cutoffs can also be of important use for the automated assessment of protein structure predictions, especially for that in the template free modeling category where the majority of the predictions are incorrect and even the human visual assessments have difficulties in judging whether the predicted models have the same or different topology to the native (Ben-David *et al.*, 2009; Jauch *et al.*, 2007). The reported quantitative TM-score thresholds is promising to provide an automated and yet reliable reference to the human-based assessments in this category. Moreover, in protein structure prediction, for example, the multitemplate-based method ModFOLD that is designed for identifying the template models by training multiple MQAP scores with TM-score (McGuffin, 2008), the quantitative TM-score cutoff system may help in designing the specific training TM-score cutoffs for selecting models sharing specific levels of structural similarities. Finally, the fold/topology assessment of protein domains is of critical importance in elucidating the functional and evolutionary relations among protein molecules (Chothia *et al.*, 2003; Zhang, 2009), where the quantitative correspondences of TM-score and topology classifications have potential use in constructing various structure-based networks from large-scale structure databases for functional and evolutional annotation purposes (Dokholyan *et al.*, 2002; Qian *et al.*, 2001). To facilitate these uses, we have built an online server called TM-fold, which helps calculate the fold classifications for any given structure pairs. The server also allows users to upload large-scale structure datasets and generate TM-score analyses based on the refined core-region structures with the long tails/loops outliers truncated. TM-fold is freely available at http://zhanglab.ccmb.med.umich.edu/TM-fold.

The second part of the studies in this article has been focused on the fold level of protein structures, which is mainly because this concept of topology has been most generally used in protein folding and protein structure prediction; also, this category of structure similarity is clearly defined and has equivalency in both SCOP and CATH databases (Hadley and Jones, 1999). Nevertheless, the extension of our approach to other levels of homologous family and superfamily should be straightforward.

Funding: Alfred P. Sloan Foundation, NSF Career Award (DBI 0746198); National Institute of General Medical Sciences (GM083107, GM084222).

Conflict of Interest: none declared.

REFERENCES

Andreeva,A. *et al.* (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.

Ben-David,M. *et al.* (2009) Assess ment of CASP8 structure predictions for template free targets. *Proteins*, **77**(Suppl. 9), 50–65.

Berman,H.M. *et al.* (2002) The protein data bank. *Acta Crystallogr., Sect D: Biol. Crystallogr.*, **58**, 899–907.

Betancourt,M.R. and Skolnick,J. (2001) Universal similarity measure for comparing protein structures. *Biopolymers*, **59**, 305–309.

Chothia,C. *et al.* (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.

Cuff,A.L. *et al.* (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.

Dokholyan,N.V. *et al.* (2002) Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl Acad. Sci. USA*, **99**, 14132–14136.

Embrechts,P. *et al.* (1997) *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin

Fischer,D. *et al.* (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins*, **53**(Suppl. 6), 503–516.

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

Hadley,C. and Jones,D.T. (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, **7**, 1099–1112.

Holm,L. *et al.* (1992) A database of protein structure families with common folding motifs. *Protein Sci.*, **1**, 1691–1698.

Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison, *Trends Biochem. Sci.*, **20**, 478–480.

Jauch,R. *et al.* (2007) Assessment of CASP7 structure predictions for template free targets. *Proteins*, **69**, 57–67.

Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta. Cryst.*, **A34**, 827–828.

Kopp,J. *et al.* (2007) Assessment of CASP7 predictions for template-based modeling targets. *Proteins*, **69**, 38–56.

Kuntz,I.D. (1992) Structure-based strategies for drug design and discovery. *Science*, **257**, 1078–1082.

Levitt,M. and Gerstein,M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.

McGuffin,L.J. (2008) The ModFOLD server for the quality assessment of protein structural models. *Bioinformatics*, **24**, 586–587.

Moult,J. *et al.* (2007) Critical assessment of methods of protein structure prediction-Round VII. *Proteins*, **69**(Suppl. 8), 3–9.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Orengo,C.A. *et al.* (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Ortiz,A.R. *et al.* (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Pascual-Garcia,A. *et al.* (2010) Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. *Proteins*, **78**, 181–196.

Qian,J. *et al.* (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.*, **313**, 673–681.

Sadreyev,R.I. *et al.* (2009) Discrete-continuous duality of protein structure space. *Curr. Opin. Struct. Biol.*, **19**, 321–328.

Siew,N. *et al.* (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

Wang,G. and Dunbrack,R.L. Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.

Zemla,A. *et al.* (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, **37**(Suppl. 3), 22–29.

Zhang,Y. (2009) Protein structure prediction: when is it useful?. *Curr. Opin. Struct. Biol.*, **19**, 145–155.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.

Zhang,Y. *et al.* (2006) On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl Acad. Sci. USA*, **103**, 2605–2610.