Using Topology to Estimate Structural Similarities of Proteins

Andersen, Jørgen Ellegaard; Jensen, Jens Ledet; Koyanagi, Yuki; Nielsen, Jakob Toudahl; Villemoes, Rasmus

Publication date:
2021

Document version:
Submitted manuscript

Using Topology to Estimate Structural Similarities of Proteins

Jørgen Ellegaard Andersen[1, 2], Jens Ledet Jensen[3], Yuki Koyanagi[2], Jakob Toudahl Nielsen[4, 5], and Rasmus Villemoes[6]

[1] *Danish Institute for Advanced Study, University of Southern Denmark*
[2] *Centre for Quantum Mathematics, Department of Mathematics and Computer Science, University of Southern Denmark*
[3] *Department of Mathematics, Aarhus University*
[4] *Interdisciplinary Nanoscience Center (iNANO), Aarhus University*
[5] *Department of Chemistry, Aarhus University*
[6] *Prevas A/S*

Abstract

An effective model for protein structures is important for the study of protein geometry, which, to a large extent, determine the functions of proteins. There are a number of approaches for modelling; one might focus on the conformation of the backbone or H-bonds, and the model may be based on the geometry or the topology of the structure in focus. We focus on the topology of H-bonds in proteins, and explore the link between the topology and the geometry of protein structures. More specifically, we take inspiration from CASP Evaluation of Model Accuracy and investigate the extent to which structural similarities, via GDT_TS, can be estimated from the topology of H-bonds. We report on two experiments; one where we attempt to mimic the computation of GDT_TS based solely on the topology of H-bonds, and the other where we perform linear regression where the independent variables are various scores computed from the topology of H-bonds. We achieved an average ΔGDT of 6.45 with 54.5% of predictions inside 2 ΔGDT for the first method, and an average ΔGDT of 4.41 with 72.7% of predictions inside 2 ΔGDT for the second method.

1 Introduction

It is widely recognised that the diverse functions of proteins are highly dependent on the three-dimensional structures of their native conformations. An effective model for describing the geometric structures of proteins is therefore important for the study of protein structures. One of the earliest models for describing the backbone conformation of proteins is the Ramachandran plots, which plots the dihedral angles (φ, ψ) before and after each C^α atoms in two-dimensional distributions [30]. The method has since been updated and extended to be used in structural validation [19, 32] and a number of other purposes (see, for example, [6] for a review). The extensions to the Ramachandran plots include combining two consecutive pairs of conformation angles [17], and characterising entire proteins by the averages over φ and ψ [7]. Another approach is to use the coordinates of backbone atoms instead of dihedral angles. Examples of this approach include the notion of curvature and torsion taken from differential geometry [27, 28], and projection of nearby atoms to a small sphere centred at each C^α atoms [22]. While the above methods are all based on the geometry of the backbone, an alternative approach is possible by considering its topology. A number of studies have used ideas from knot theory to study the link between topology and geometry of proteins [18, 9, 33]. Yet another approach is to focus on H-bonds, which is one of the main mechanisms determining and stabilising the native structure of the proteins [5, 34, 21]. Studies suggest incorporating H-bond geometry improves the quality of protein structure models [12, 20]. In [23], spatial rotations were introduced as a systematic

three-dimensional descriptor of H-bond geometry, and were found to correspond well to the concrete secondary structures and other local structural motifs. The dataset from [23] has further been used by Penner in [24, 25] to estimate free energy of coronarirus spike proteins, with a view to identifying specific sites of interest for vaccine development. If we concentrate on the topology of H-bonds, we obtain a graph, with backbone atoms as vertices and the covalent and H-bonds as edges. Such H-bond graphs were used to study the dynamics of membrane proteins [35], and for structural comparison [29]. In [26], an extension to this structure was used to study protein structures.

In this paper, we investigate the link between H-bond topology of proteins and their geometric structures. We take inspiration from CASP Evaluation of Model Accuracy (EMA) [10], and investigate how well we can estimate the GDT_TS of the submitted structures using H-bond graphs of the submitted and target structures. GDT_TS has been criticised, among others, for being dependent on the lengths of proteins and for having somewhat arbitrary distance cutoffs [37, 11]. Nonetheless it is a widely accepted measure used to compare protein structures, and we use it here as an indication of structural similarities. We designed two experiments. In the first experiment, we attempt to follow the algorithm for computing GDT_TS, but with only the proteins' topological information (from their H-bond graphs) as the input. The second experiment is a linear regression where independent variables are certain similarity scores computed from the protein H-bond graph, and the dependent variable is GDT_TS. We note here, that our methods are not intended as an attempt for the CASP EMA. Indeed, both methods require the target structure's H-bond graph as part of the input data, which is not available in CASP. They are intended as an investigation into the usefulness of protein topology in comparison of protein structures. However, one could of course imagine combining our methods with an algorithm to predict H-bond graphs from primary sequences to be used in CASP EMA or similar experiments. Indeed, the idea for the investigation originated in a novel approach to the protein folding problem, inspired in part by [23]. It is based on a two-stage process, where in the first stage one or more H-bond graphs are predicted from a primary sequence, then in the second stage the geometric structure is predicted from the H-bond graph(s). We have a method to enumerate possible H-bond graphs, as well as a method to predict local geometric structure of proteins from H-bond graphs [1, 3, 2]. The current study fits in this programme as a "proof of concept" for the idea that H-bond topology of proteins is strongly linked to their geometric structures. Our model is purely based on the topology of protein structures, therefore is less affected by the dynamic nature of the proteins, which is important in their diverse functions [13]. Furthermore, it is independent of alignment, which simplifies its use in potential high throughput applications.

2 Methods and Results

2.1 Dataset

The dataset consists of 33 target structures for CASP14 together with the submitted candidate structures, downloaded from CASP data archive [8] (There were 34 target structures available for download, but one, T1044, did not have any corresponding candidates and was dropped.). The size of proteins, measured in the number of residues, ranged from 74 to 922 (Figure 1). Majority of the target structures had length less than 300 residues, with 6 targets having more than 300 residues. The range of the number of candidate structures per protein was from 204 to 599, with the majority of targets receiving more than 500 submissions (Figure 2; Participants are allowed to submit more than one candidate structure). The larger target structures seems to have received as many submissions as the smaller target structures.

We also utilised data from CASP13 to construct our regression model (Section 2.3). There were 20 target structures available for download, with length ranging from 52 to 405 residues, and two structures having more than 300 residues (Figure 3)

The data was processed to obtain information about the H-bonds, following the procedure described in [23]. The H-bonds were determined by the DSSP program [14], with the

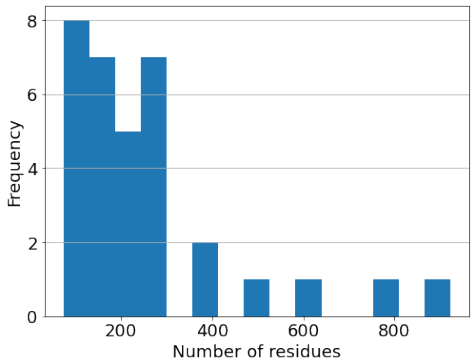

Figure 1: Frequency of target structures in CASP14 by length.

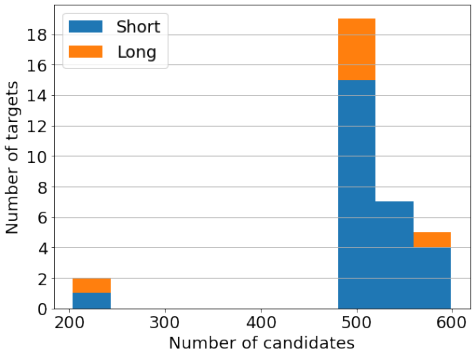

Figure 2: Number of candidates per target structure in CASP14. Short targets are those with fewer than 300 residues, and the long targets are with more than 300 residues.

additional conditions [4];

$$\text{HO-distance} < 2.7\text{Å}$$
$$\text{angle(NHO)}, \text{angle(COH)} > 90°.$$

For the majority of proteins in the resulting data, the number of H-bonds was roughly half of the length measured as the number of residues (Figure 4).

2.2 GDT-like algorithm based on H-bond graphs

We attempt to mimic the GDT algorithm [38], but based only on protein H-bond graphs, i.e. based on information about the protein's hydrogen bonds, but not its geometric structure.

Let T be the graph of the target protein, with vertices $\{v_1, \ldots, v_l\}$ representing the residues, ordered along the backbone, and edges $\{e_1, \ldots, e_m\}$ representing the backbone peptide bonds and H-bonds. Similarly,

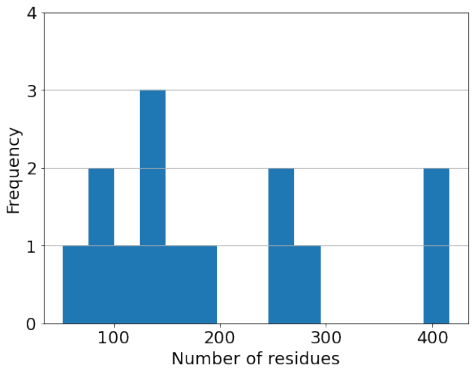

Figure 3: Frequency of target structures in CASP13 by length.

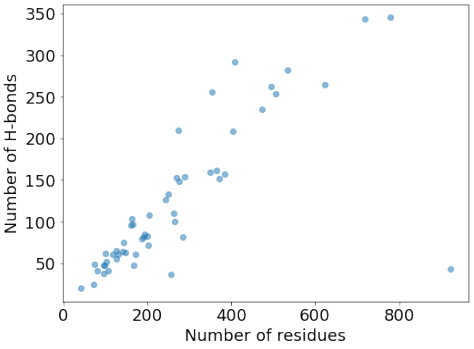

Figure 4: Number of H-bonds versus length for the target structures in CASP13 & 14.

let C be the graph of the candidate protein, with vertices $\{w_1, \ldots, w_l\}$ representing the residues, ordered along the backbone, and edges $\{f_1, \ldots, f_n\}$ representing primary and H-bonds. For a set S, let $\#S$ denote the number of elements in S, and for a graph G, let $\mathcal{V}(G)$ and $\mathcal{E}(G)$ denote the set of vertices and edges in G, respectively. We have $l = \#\mathcal{V}(T)(= \#\mathcal{V}(C))$. The idea is to start with small subgraphs of C and T (corresponding to the same backbone segment), and to "grow" them incrementally, until the difference between the subgraphs is over a predetermined threshold value. We repeat this for different initial subgraphs, and determine the maximum subgraph of C, $\hat{C}_{\text{sub}}$, whose difference from the corresponding subgraph of T (the correspondence is defined below in the detailed description) is below some threshold value r. The score for the candidate graph, which we call g_r, is then given by

$$g_r = 100 \times \frac{\#\mathcal{V}(\hat{C}_{\text{sub}})}{\#\mathcal{V}(C)}. \qquad (1)$$

We give a detailed description of the algorithm below, together with the pseudocode in Algorithm 1. Let $d(A, B)$ be a distance function between graphs A and B and let $r \in (0, \infty)$. We compute the score of similarity between the graphs T and C as follows;

1. Set $i = 1$.
2. We select a subgraph $C_{\text{sub}}(i)$ of C, consisting of three vertices $\{w_i, w_{i+1}, w_{i+2}\}$ starting from the ith position and the edges connecting them. Select a subgraph $T_{\text{sub}}(i)$ of T in the same way.
3. Compute the distance measure $d_i = d(T_{\text{sub}}(i), C_{\text{sub}}(i))$.
4. If $d_i \geq r$, where r is the pre-determined limit, the initial segment is already over the limit value. Increment i by 1, go to 2.
5. If $d_i < r$, "grow" the subgraph $C_{\text{sub}}(i)$ by adding all edges that are connected to $C_{\text{sub}}(i)$, together with the vertices connected to these edges. Call the selected edges and vertices, together with $C_{\text{sub}}(i)$, $C_{\text{sub}}2(i)$. Select $T_{\text{sub}}2(i)$ from T in the same way.
6. Compute the distance measure $d_i = d(T_{\text{sub}}2(i), C_{\text{sub}}2(i))$.
7. Repeat the above two steps ("growing" subgraphs and computing the distance measure), until $d_i \geq r$. If $d_i \geq r$, move to the next starting segments by incrementing i by 1, go to 2.
8. If $C_{\text{sub}}(i) = C$, we have the entire graph under the limit value.
9. After going through all starting segments, we have a set $S = \left\{C_{\text{sub}}(i) \middle| i \in \{1, \ldots, l-2\}\right\}$ of maximal $C_{\text{sub}}(i)$'s. Select the longest $C_{\text{sub}}(i)$ in S, which we call $\hat{C}_{\text{sub}}$.
10. $g_r = 100 \times \frac{\#\mathcal{V}(\hat{C}_{\text{sub}})}{\#\mathcal{V}(C)}$.

For the current analysis we define the distance function d by

$$d(A, B) = \#(\mathcal{E}(A) \ominus \mathcal{E}(B)), \qquad (2)$$

where $\ominus$ denotes the symmetric difference of two sets;

$$A \ominus B = (A \setminus B) \cup (B \setminus A).$$

Algorithm 1 Pseudocode for GDT-like algorithm

for i in $\{1, \ldots, l-2\}$ **do**
  Let $C_{\text{sub}}(i)$ be the subgraph of C obtained by taking three vertices $\{w_i, w_{i+1}, w_{i+2}\}$ and the edges connecting them in C
  Let $T_{\text{sub}}(i)$ be the subgraph of T obtained by taking three vertices $\{v_i, v_{i+1}, v_{i+2}\}$ and the edges connecting them in T
  Compute the distance measure $d_i = d(T_{\text{sub}}(i), C_{\text{sub}}(i))$
  if $d_i \geq r$ **then**
    Continue to next i
  end if
  while True **do**
    Let $T_{\text{sub}}2(i)$ be the subgraph of T obtained by taking $T_{\text{sub}}(i)$ together with all edges connected to the vertices in $T_{\text{sub}}(i)$, and the end-vertices of these edges (i.e. "grow" the subgraph by 1 edge+vertex pair)
    Let $C_{\text{sub}}2(i)$ be the subgraph of C obtained in the same manner
    Compute the distance measure $d_i = d(T_{\text{sub}}2(i), C_{\text{sub}}2(i))$
    if $d(i) \geq r$ **then**
      Break out of while loop
    end if
    Set $T_{\text{sub}}(i) = T_{\text{sub}}2(i)$
    Set $C_{\text{sub}}(i) = C_{\text{sub}}2(i)$
    if $T_{\text{sub}}(i) == T$ **then**
      Break out of while loop
    end if
  end while
end for
$\hat{C}_{\text{sub}} = \max\left\{C_{\text{sub}}(i) \middle| i \in \{1, \ldots, l-2\}\right\}$
Score $= 100 \cdot \#\mathcal{V}(\hat{C}_{\text{sub}})/\#\mathcal{V}(C)$

The algorithm is also dependent on the limit value r for the distance between two subgraphs. We tested for the effect of different r values by computing the score g_r for $r = 5, 10, 20, 40, 80$ for all candidate structures and looking at their distributions, together with their correlation with GDT_TS (Figure 5). As a result $r = 80$ was excluded as being too high (resulting in more than 20% of all structures having score of 100). GDT_TS is computed as an average of scores for four different cutoff values (1,2,4, and 8 Å) [16]. We imitate this by computing an average over different sets of r-values. The distributions of (average) scores for different sets of r-values are shown in Figure 6. Based on these, we chose the average over r-values 10, 20 and 40 as our final score, since the combination had the widest spread of values. We call the final composite score Γ-GDT;

$$\Gamma\text{-GDT} = (g_{10} + g_{20} + g_{40})/3. \qquad (3)$$

We then predicted the best candidate structure for each target by selecting the structure with the highest Γ-GDT, and look at the difference between GDT_TS of our prediction and GDT_TS of the best candidate for each target structure, which we call ΔGDT. The distribution of ΔGDT is shown in Figure 7. The average ΔGDT for all targets was 6.45, with the highest value of 36.38. We were able to identify a candidate with ΔGDT < 2 in 18 targets, and with ΔGDT < 10 in 24 targets (Figure 7). The distribution of GDT_TS for the best candidate structure against GDT_TS for the selected structure is shown in Figure 8. It turned out the particular set of r-values we chose for the computation of Γ-GDT gives the best prediction result (Table 1).

r-values	ΔGDT < 2	ΔGDT < 10
10,20,40	18	24
5,10,20	14	22
5,10,20,40	17	23

Table 1: Number of predictions (out of 33) with specified ΔGDT range for different combinations of r-values.

We also tested for two different distance function to (2). In the first, we reduced the contribution made by an edge in the set $\mathcal{E}(C) \ominus \mathcal{E}(T)$, if there is an edge that lies close to it. For $x \in \mathbb{R}$, define a function f_1 by

$$f_1(x) = \begin{cases} 1 & \text{if } |x| > 4 \\ |x|/4 & \text{otherwise.} \end{cases} \qquad (4)$$

We define a new distance function d_1 by

$$\begin{aligned} d_1(A,B) = & \\ & \sum_{(p,q)\in A\setminus B} \min\{\frac{f_1(p-p') + f_1(q-q')}{2} \\ & \qquad |(p',q') \in B \setminus A\} \\ & + \sum_{(p,q)\in B\setminus A} \min\{\frac{f_1(p-p') + f_1(q-q')}{2} \\ & \qquad |(p',q') \in A \setminus B\}. \end{aligned} \qquad (5)$$

In the second, we tried to reduce the contribution by a "close" edge further by setting

$$f_2(x) = \begin{cases} 1 & \text{if } |x| > 4 \\ \frac{\exp(|x|)-1}{\exp(4)-1} & \text{otherwise,} \end{cases} \qquad (6)$$

and

$$\begin{aligned} d_2(A,B) = & \\ & \sum_{(p,q)\in A\setminus B} \min\{\frac{f_2(p-p') + f_2(q-q')}{2} \\ & \qquad |(p',q') \in B \setminus A\} \\ & + \sum_{(p,q)\in B\setminus A} \min\{\frac{f_2(p-p') + f_2(q-q')}{2} \\ & \qquad |(p',q') \in A \setminus B\}. \end{aligned} \qquad (7)$$

The prediction results for different distance functions are shown in Figure 9. We see that the performance of the original distance function d (2), which is a simple count of the elements in the symmetric difference of the sets of edges, is significantly better than the two modified distance functions.

2.3 Linear regression based on the protein fatgraph model

The second method is a linear regression on the similarity scores which we compute based on the hydrogen bonds in the candidate and target structures. Each hydrogen bond is identified by the position of its donor- and acceptor atoms, so each bond can be expressed as a 2-tuple of integers (p, q), where the donor is the p'th atom along the backbone and the acceptor the q'th.

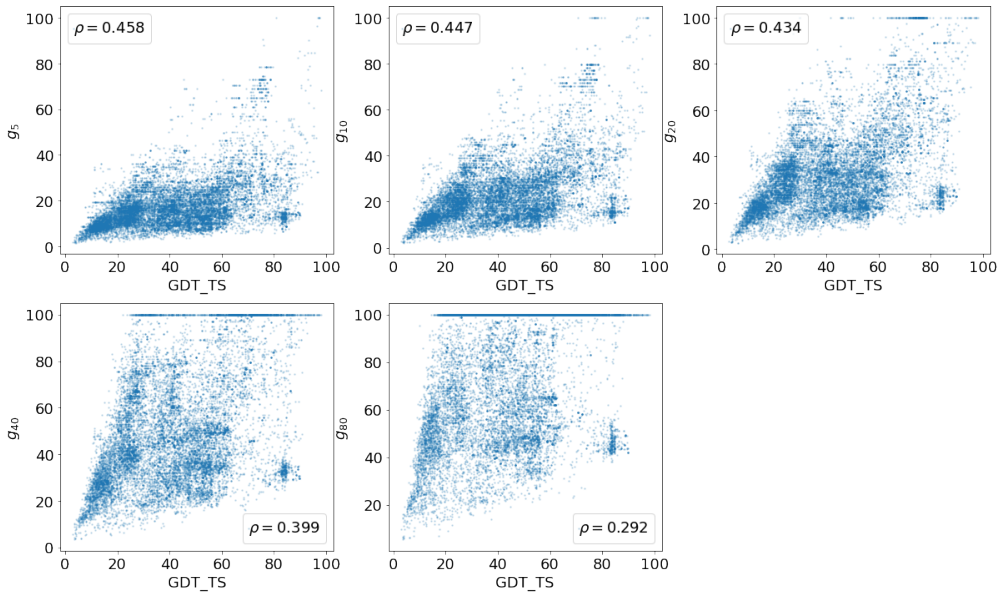

Figure 5: Distribution of g_r scores for different values of r, against GDT_TS. ρ is the Spearman's correlation coefficient between GDT_TS and g_r scores.

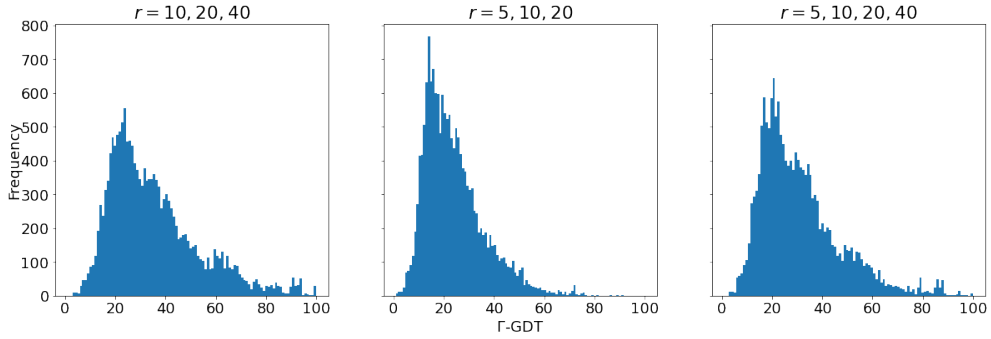

Figure 6: Distribution of Γ-GDT for different combinations of r-values.

The first of our similarity scores is the proportion of the bonds, which are correctly identified in the candidate structure. In other words, if H_T, H_C are the sets of hydrogen bonds respectively in the target structure and in the candidate structure, then the first score P is defined as;

$$P = \frac{\#(H_T \cap H_C)}{\#H_T},$$

where we use the fact that for two bonds (p, q) and (p', q'), $(p, q) = (p', q')$ iff $p = p'$ and $q = q'$.

The second similarity score S_n depends on a parameter $n \in \mathbb{N}$. For a non-negative integer $x \in \mathbb{Z}$, define

$$f_n(x) = \begin{cases} 1 - x/n & \text{if } x \leq 2n \\ -1 & \text{otherwise} \end{cases}. \tag{8}$$

For a bond $(p, q) \in H_C$, set

$$s_C((p,q)) = \max\left\{ f_n(|p - p'|) + f_n(|q - q'|) \mid (p', q') \in H_T \setminus H_C \right\}.$$

Similarly for $(p, q) \in H_T$, set

$$s_T((p,q)) = \max\left\{ f_n(|p - p'|) + f_n(|q - q'|) \mid (p', q') \in H_C \setminus H_T \right\}.$$

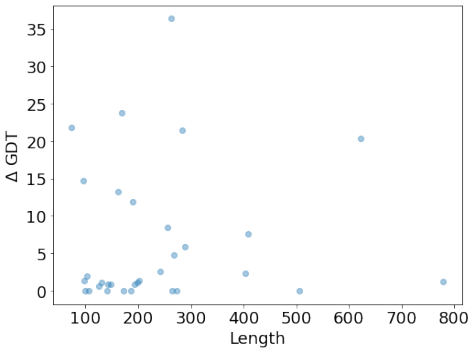

Figure 7: Distribution of ΔGDT against length (measured in number of residues) for 33 target structures, predicted using Γ-GDT.

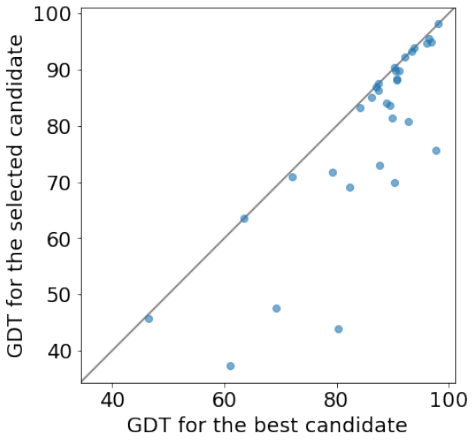

Figure 8: GDT_TS for the best candidate structure against GDT_TS for the selected structure.

S_n is then given by

$$S_n = \frac{1}{\#((H_T \setminus H_C) \cup (H_C \setminus H_T))} \times \left(\sum_{(p,q) \in H_C \setminus H_T} s_C((p,q)) + \sum_{(p',q') \in H_T \setminus H_C} s_T((p',q')) \right).$$

So for a given candidate structure, we can compute S_n for different n's.

Having calculated P and S_n, $n \in I$, where I is a subset of $\{1, 2, \ldots, 10\}$, for all candidate structures, we perform a linear regression with P, S_n as independent variables and GDT_TS as the dependent variable. We estimated the regression model using data from CASP13,

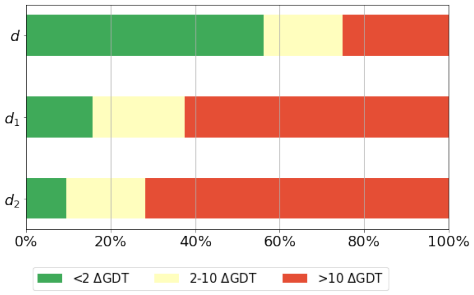

Figure 9: Percentages of predictions with ΔGDT < 2, $2 \leq$ ΔGDT < 10, and ΔGDT ≥ 10 for different distance functions.

and applied the model to data from CASP14. After testing for all subsets $I \subset \{1, 2, \ldots, 10\}$ by running multiple regression with P and $S_n, n \in I$ as independent variables, we found that setting $I = \{2\}$ gave the best results with CASP13 data. The regression equation, based on all candidate structures in CASP13, was determined to be

$$\text{GDT\_TS} = 10.70 + 0.63P + 1.26S_2.$$

Using this equation, we estimated the GDT_TS for CASP14 candidate structures, and selected the structure with the highest estimated GDT_TS for each target. We were able to identify a candidate structure with ΔGDT < 2 for 23 out of 33 targets, with the average ΔGDT of 4.57 (Table 2). The frequency distribution of ΔGDT is shown in Figure 10. The large ΔGDT values were observed in shorter proteins, although it must be noted that most of targets have lengths less than 300 residues.

We also investigated the effect of the score function (8) by scaling it with the exponential function;

$$\tilde{f}_n(x) = \begin{cases} 1 - \frac{2(\exp(x)-1)}{\exp(2n)-1} & \text{if } x \leq 2n \\ -1 & \text{otherwise} \end{cases}. \quad (9)$$

Compared to (8), the new score function (9) gives smaller penalties to difference in bond positions, especially when the difference is small. Using the data from CASP13 and (9), we found that setting $I = \{2, 6, 8, 10\}$ gave the best result with regards to identifying the most candidates with ΔGDT < 2. However, the S_n scores are strongly correlated, and we decided to use $I = \{2\}$ again. This gave a result close to that obtained with

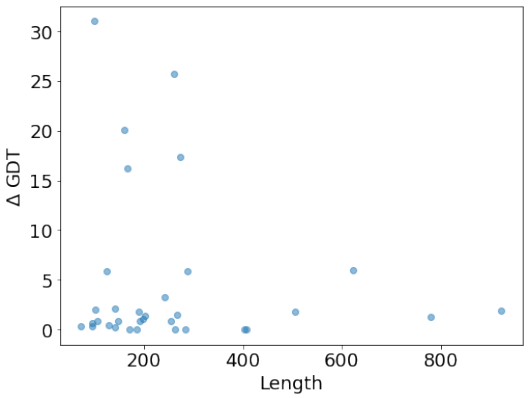

Figure 10: Distribution of ΔGDT against length (measured in number of residues) for 33 target structures, by linear regression method.

$I = \{2, 6, 8, 10\}$ (11 targets with ΔGDT < 2, compared to 12 targets with ΔGDT < 2). The regression equation was

$$\text{GDT\_TS} = 9.68 + 0.63P + 0.20S_2.$$

The new score function (9) resulted in a small improvement for prediction with CASP14 data, where we were able to identify a candidate structure with ΔGDT < 2 for 24 out of 33 targets, and the average ΔGDT of 4.41 (Table 2).

We then removed the S_n scores from the independent variables, and ran the regression with only P scores as the independent variable. The regression equation now read

$$\text{GDT\_TS} = 9.56 + 0.63P.$$

This only resulted in a small drop in our ability to identify the best candidate structure, with ΔGDT < 2 for 22 out of 33 targets and the average ΔGDT of 5.53 (Table 2).

	% of candidates with	
	ΔGDT < 2	ΔGDT < 10
f_n	69.70	84.85
$\tilde{f}_n$	72.73	84.85
No S_n	66.67	84.85

Table 2: Prediction results for different score functions, showing the percentages of targets (out of 33), where the selected candidate structure had ΔGDT less than 2 and 10, respectively.

3 Discussion

We have shown that the information on H-bonds alone can, to a large extent, correctly assess similarities in geometric structures of proteins. Even though a direct comparison between our results and CASP EMA is not possible, the performance of our methods, measured as the percentages of predictions with ΔGDT < 2 and ΔGDT < 10, are clearly numerically superior to the best performance in CASP 14 EMA (Figure 11). It could be argued that both our methods essentially rely on simply counting the matched (or unmatched, in the case of Γ-GDT) H-bonds in two structures. The modified distance functions in the Γ-GDT and the S scores in the linear regression, which measures the differences between unmatched H-bonds, have negative or relatively small positive effect on the overall accuracy of predictions. The fact that these relatively simple methods can nonetheless assess similarities in protein structures correctly, demonstrates the strong link between the topology and the geometry of proteins.

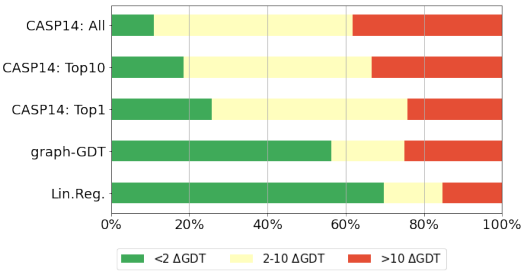

Figure 11: Percentages of the predictions with ΔGDT < 2, $2 \leq$ ΔGDT < 10 and ΔGDT ≥ 10. The figures for CASP14 are averages of all models, top 10 models, and the best-performing model, ordered by the number of predictions with ΔGDT < 2. The data for CASP14 was obtained from CASP Data Archive (`https://www.predictioncenter.org/download_area/CASP14/`) and processed by the authors.

In the linear regression analysis, we observe that the larger values of n in S_n scores, which in effect enlarges the search window for "similar" H-bonds, do not improve the prediction accuracy. An explanation could be that it is simply a consequence of using GDT_TS as the measure of structural similarity, as a small

local difference (e.g. an extra turn where there should be none) can result in a significantly lower GDT_TS. Further investigation is needed to ascertain the cause of this behaviour.

There are broadly two types of methods used in CASP model accuracy estimation. Consensus, or clustering methods take multiple candidate structures as input and tries to identify a structure, that is the "best match" for the input structures according to some criteria. Single-model methods, on the other hand, takes a single candidate structure as an input and tries to estimate its accuracy, independent of other candidate structures. The consensus methods have generally outperformed the single-model methods, and this resulted in the development effort being concentrated on the consensus methods in the past [31]. More recently the single-model methods have received more attention and development effort [15], as the potential issues with the consensus methods are recognised. One issue, for example, is that the consensus methods may not be very useful in the environment outside the CASP-setup, where a large number of candidate structures may not be available for the input. Another potential issue, related to the first, is that the consensus methods may simply be taking advantage of the fact that many CASP models are now able to produce high-quality candidate structures, which are, naturally, similar to each other [36]. Our method is, by construction, unlikely to be improved to outperform the best accuracy estimation methods, as it ignores the geometric data in the candidate structures and only utilises the topological data. However, the relative simplicity of our method means it should be relatively easy to combine it with an existing method to improve its performance. We chose not to attempt it in this paper, as our focus here has been to investigate the link between the topology and the geometry of proteins, rather than to participate in CASP EMA. Nonetheless, as we mentioned in Introduction, one could easily imagine combining our method with an algorithm for predicting hydrogen bonds from a primary sequence. When a high-accuracy prediction of hydrogen bonds becomes possible, our method has the advantage that it could be combined with both a consensus method and a single-model method.

Acknowledgement

This paper is partly a result of the ERC-SyG project, Recursive and Exact New Quantum Theory (ReNewQuantum) which received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810573.

References

[1] Andersen, Jørgen Ellegaard, Fuji, Hiroyuki, and Koyanagi, Yuki. "Enumeration of protein structures by matrix model techniques". In preparation.

[2] Andersen, Jørgen Ellegaard, Fuji, Hiroyuki, and Koyanagi, Yuki. "Topology of protein metastructure and β-sheet topology". Preprint.

[3] Andersen, Jørgen Ellegaard, Koyanagi, Yuki, and Nielsen, Jakob Toudahl. "Prediction of H-bond rotations from protein H-bond topology". Preprint.

[4] Baker, E.N. and Hubbard, R.E. "Hydrogen bonding in globular proteins". *Progress in Biophysics and Molecular Biology* 44.2 (1984), pp. 97–179. ISSN: 0079-6107.

[5] Bordo, Domenico and Argos, Patrick. "The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins". *Journal of molecular biology* 243.3 (1994), pp. 504–519.

[6] Carugo, Oliviero and Djinović-Carugo, K. "Half a century of Ramachandran plots". *Acta Crystallographica Section D: Biological Crystallography* 69.8 (2013), pp. 1333–1341.

[7] Carugo, Oliviero and Djinović-Carugo, Kristina. "A proteomic Ramachandran plot (PRplot)". *Amino acids* 44.2 (2013), pp. 781–790.

[8] *CASP Data Archive*. `https://predictioncenter.org/download_area/CASP14/`.

[9] Chen, Shi-Jie and Dill, Ken A. "Symmetries in proteins: A knot theory approach". *The Journal of chemical physics* 104.15 (1996), pp. 5964–5973.

[10] Cozzetto, Domenico, Kryshtafovych, Andriy, Ceriani, Michele, and Tramontano, Anna. "Assessment of predictions in the model quality assessment category". *Proteins: Structure, Function, and Bioinformatics* 69.S8 (2007), pp. 175–183.

[11] Garg, Shikhin, Kakkar, Smarth, and Runthala, Ashish. "Improved protein model ranking through topological assessment". *Computational Biology and Bioinformatics* (2016), pp. 410–428.

[12] Grishaev, Alexander and Bax, Ad. "An empirical backbone- backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation". *Journal of the American Chemical Society* 126.23 (2004), pp. 7281–7292.

[13] Henzler-Wildman, Katherine and Kern, Dorothee. "Dynamic personalities of proteins". *Nature* 450.7172 (2007), pp. 964–972.

[14] Kabsch, Wolfgang and Sander, Christian. "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features". *Biopolymers* 22.12 (1983), pp. 2577–2637.

[15] Kryshtafovych, Andriy, Monastyrskyy, Bohdan, Fidelis, Krzysztof, Schwede, Torsten, and Tramontano, Anna. "Assessment of model accuracy estimations in CASP12". *Proteins: Structure, Function, and Bioinformatics* 86 (2018), pp. 345–360.

[16] Kryshtafovych, Andriy, Prlic, Andreas, Dmytriv, Zinoviy, Daniluk, Pawel, Milostan, Maciej, Eyrich, Volker, Hubbard, Tim, and Fidelis, Krzysztof. "New tools and expanded data analysis capabilities at the Protein Structure Prediction Center". *Proteins: Structure, Function, and Bioinformatics* 69.S8 (2007), pp. 19–26.

[17] Levitt, Michael. "A simplified representation of protein conformations for rapid simulation of protein folding". *Journal of molecular biology* 104.1 (1976), pp. 59–107.

[18] Levitt, Michael. "Protein folding by restrained energy minimization and molecular dynamics". *Journal of molecular biology* 170.3 (1983), pp. 723–764.

[19] Lovell, Simon C, Davis, Ian W, Arendall III, W Bryan, De Bakker, Paul IW, Word, J Michael, Prisant, Michael G, Richardson, Jane S, and Richardson, David C. "Structure validation by Cα geometry: ϕ, ψ and Cβ deviation". *Proteins: Structure, Function, and Bioinformatics* 50.3 (2003), pp. 437–450.

[20] Morozov, Alexandre V, Kortemme, Tanja, Tsemekhman, Kiril, and Baker, David. "Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations". *Proceedings of the National Academy of Sciences* 101.18 (2004), pp. 6946–6951.

[21] Pace, C Nick, Scholtz, J Martin, and Grimsley, Gerald R. "Forces stabilizing proteins". *FEBS letters* 588.14 (2014), pp. 2177–2184.

[22] Peng, Xubiao, Chenani, Alireza, Hu, Shuangwei, Zhou, Yifan, and Niemi, Antti J. "A three dimensional visualisation approach to protein heavy-atom structure reconstruction". *BMC structural biology* 14.1 (2014), pp. 1–16.

[23] Penner, Robert, Andersen, Ebbe Sloth, Jensen, Jens Ledet, Kantcheva, Adriana Krassimirova, Bublitz, Maike, Nissen, Poul, Rasmussen, Anton Michael Havelund, Svane, Katrine Louise, Hammer, Bjørk, Rezazadegan, Reza, Nielsen, Niels Christian, Nielsen, Jakob Toudahl, and Andersen, Jørgen Ellegaard. "Hydrogen bond rotations as a uniform structural tool for analyzing protein architecture". *Nature Communications* 5 (2014).

[24] Penner, Robert C. "Conserved High Free Energy Sites in Human Coronavirus Spike Glycoprotein Backbones". *Journal of Computational Biology* 27.11 (2020), pp. 1622–1630.

[25] Penner, Robert Clark. "Antiviral Resistance against Viral Mutation: Praxis and Policy for SARS CoV-2". *BioRxiv* (2021).

[26] Penner, Robert, C., Knudsen, Micheal, Wiuf, Carsten, and Andersen, Jørgen Ellegaard. "Fatgraph models of proteins". *Communications on Pure and Applied Mathematics* 63.10 (2010), pp. 1249–1297.

[27] Rackovsky, S and Scheraga, HA. "Differential geometry and polymer conformation. 1. Comparison of protein conformations1a, b". *Macromolecules* 11.6 (1978), pp. 1168–1174.

[28] Rackovsky, S and Scheraga, HA. "Differential geometry and protein folding". *Accounts of Chemical Research* 17.6 (1984), pp. 209–214.

[29] Rahat, Ofer, Alon, Uri, Levy, Yaakov, and Schreiber, Gideon. "Understanding hydrogen-bond patterns in proteins using network motifs". *Bioinformatics* 25.22 (2009), pp. 2921–2928.

[30] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. "Stereochemistry of polypeptide chain configurations". *J. Mol. Biol.* 7 (1963), pp. 95–99.

[31] Ray, Arjun, Lindahl, Erik, and Wallner, Björn. "Improved model quality assessment using ProQ2". *BMC bioinformatics* 13.1 (2012), p. 224.

[32] Read, Randy J, Adams, Paul D, Arendall III, W Bryan, Brunger, Axel T, Emsley, Paul, Joosten, Robbie P, Kleywegt, Gerard J, Krissinel, Eugene B, Lütteke, Thomas, Otwinowski, Zbyszek, et al. "A new generation of crystallographic validation tools for the protein data bank". *Structure* 19.10 (2011), pp. 1395–1412.

[33] Røgen, Peter and Fain, Boris. "Automatic classification of protein structure by using Gauss integrals". *Proceedings of the National Academy of Sciences* 100.1 (2003), pp. 119–124.

[34] Rose, George D and Wolfenden, Richard. "Hydrogen bonding, hydrophobicity, packing, and protein folding". *Annual review of biophysics and biomolecular structure* 22.1 (1993), pp. 381–415.

[35] Siemers, Malte, Lazaratos, Michalis, Karathanou, Konstantina, Guerra, Federico, Brown, Leonid S, and Bondar, Ana-Nicoleta. "Bridge: A graph-based algorithm to analyze dynamic H-bond networks in membrane proteins". *Journal of chemical theory and computation* 15.12 (2019), pp. 6781–6798.

[36] Won, Jonghun, Baek, Minkyung, Monastyrskyy, Bohdan, Kryshtafovych, Andriy, and Seok, Chaok. "Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning". *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019), pp. 1351–1360.

[37] Xu, Jinrui and Zhang, Yang. "How significant is a protein structure similarity with TM-score= 0.5?" *Bioinformatics* 26.7 (2010), pp. 889–895.

[38] Zemla, Adam. "LGA: a method for finding 3D similarities in protein structures". *Nucleic acids research* 31.13 (2003), pp. 3370–3374.