Published in final edited form as:

Proteins. 2009 February 15; 74(3): 701–711. doi:10.1002/prot.22186.

Information theory provides a comprehensive framework for the evaluation of protein structure predictions

Rosemarie Swanson¹, Marina Vannucci², and Jerry W. Tsai^{1,3}

¹Department of Biochemistry and Biophysics, Texas A&M University, 2128 TAMU, College Station, TX 77843, USA.

²Department of Statistics, MS 138, Rice University, 6100 Main Street, Houston, TX 77251-1892, USA.

³Chemistry Department, University of the Pacific, 3601 Pacific Avenue, Stockton, CA 95211, USA.

Abstract

Protein structure prediction has a number of important ad hoc similarity measures for evaluating predictions, but would benefit from a measure that is able to provide a common framework for a broad range of comparisons. Here we show that a mutual information-like measure can provide a comprehensive framework for evaluating protein structure prediction of all types. We discuss the concept of information, its application to secondary structure, and the obstacle to applying it to 3D structure. Based on insights from the secondary structure case, we present an approach to work around the 3D difficulties, and develop a method to measure the mutual information provided by a 3D structure prediction. We integrate the evaluation of all types of protein structure prediction into a single frame work, and compare the amount of information provided by various prediction methods, including secondary structure prediction. Within this broadened framework, the idea that structure is better preserved than sequence during evolution is evaluated quantitatively for the globin family. A nearly perfect sequence match in the globin family corresponds to about 300 bits of information, whereas a nearly perfect structural match for the same two proteins corresponds to about 2500 bits of information, where bits of information describes the probability of obtaining a match of similar closeness by chance. Mutual information provides both a theoretical basis for evaluating structure similarity and an explanatory surround for existing similarity measures.

Keywords

comparison; superposition; probability; volume; prediction sphere; structural variation; sequence homology; GDT_TS; CASP

Introduction

The 7 CASP (Critical Assessment of Methods for Protein Structure Prediction) experiments 1–7 conducted since 1994 have provided an environment that has led to dramatically improved ability to predict 3D protein structure from amino acid sequence. A challenge to

Corresponding author: Rosemarie Swanson, Department of Biochemistry and Biophysics, Texas A&M University, 2128 TAMU, College Station, TX 77843; 979-845-5032 voice, 979-845-9274 fax, rosmar@tamu.edu.

Jerry W Tsai, jtsai@pacific.edu, 209-946-2271 voice

Marina Vannucci, marina@rice.edu, 713-348-6132 voice

This work was performed at Texas A&M University.

the field has been how to measure either the goodness of a prediction or the amount of progress in the field. 8-9 This challenge has been addressed with a strong dependence on an important and widely used empirical measure, the global distance test-total score, GDT_TS 10. However, there still exists no simple way to compare the usefulness of predictions derived by different methods (for example, template-free vs template-based modeling).

In this work we discuss the concept of information as a way to evaluate the goodness of protein structure predictions. Mutual information assesses a proposed answer to a question in terms of how well it reduces the uncertainty surrounding the correct answer. Even though there is a straightforward way to apply information theory to the evaluation of *secondary structure* prediction, it has not been obvious how to approach the *three-dimensional* problem. However, we describe a way around the 3D difficulties, and present a method to measure the information provided by a 3D structure prediction. We apply the method to various sets of protein structures, compare it with other approaches, and present quantitative comparisons of the information provided by different types of structure prediction.

Entropy and Information Concepts

Entropy and information are related concepts and are expressed in the same units (bits) but they are also opposites in a general sense. The following discussion addresses the relation between the two concepts. *Entropy* is a concept associated with the probability distribution derived from a *single* set of outcomes, e.g., a set of observations. It quantifies the *variety* of possible outcomes. *Information* is a concept connected with the probability distribution of the correlation between *two* sets of outcomes, e.g., a set of observations and a set of predictions about them. It quantifies the extent to which a predicted outcome corresponds to an actual outcome, in terms of the *reduction in variety* of possible outcomes, *given the prediction*. The method for calculating the amount of information delivered by a prediction scheme is cleverly illustrated by ET Jaynes 11, ^{esp. p.204}. A more formal but less intuitive treatment is given by others 12-14.

Entropy's numerical value gives the amount of choice, variety, or uncertainty presented by a set of events. For example, the "events" could be the observed secondary structure at each position of a particular 150-residue protein sequence. Two factors can affect the amount of variety associated with a set of events. The *first* factor is the number of *types* of event. The amount of choice/uncertainty associated with a set of observations that classifies secondary structure into the types *coil, beta strand, turn, alpha helix, or 3-10 helix* (5 types) is potentially greater than the amount of choice/uncertainty associated with a set that includes only 3 types of event. The *second* factor affecting the associated amount of choice/uncertainty is the *relative frequency* of each of the types of events. For example, if a protein has 80% of its residues in helix and the rest in coil, the amount of choice/uncertainty associated with its "set of events" (observed secondary structure) is less than it would be for a protein that has 1/3 coil, 1/3 beta strand, 1/3 helix. The mathematical relationship between a probability distribution and its entropy is given by the following expression, where p_{event} is the probability of occurrence of event type *event*, and N different types of event occur.

$$\text{Entropy} = - \sum_{event=1}^N p_{event} \log_2(p_{event})$$

A handy correlate of the numerical value of entropy is its exponential (to base 2, since it is expressed in bits), because that number describes the amount of choice/uncertainty on the same scale as the number of types of events. For example, in the case of 1/3 coil, 1/3 beta strand, 1/3 helix, the value of the exponentiated entropy is 3. In the 80% helix, 20% coil

case, the exponentiated entropy value is 1.65, reflecting the intuitively lower amount of choice/uncertainty in the types of secondary structure present. This value represents the amount of choice/uncertainty in secondary structure associated with one residue, in the particular protein under discussion.

Entropy measures the amount of choice present in a set of events. In its exponentiated form, it measures uncertainty in terms of the effective number of choices associated with one event.

In contrast, *information* is calculated from the probability distribution derived from the *correlation* between two series of events. Consider, as an example, that *one* set of events is the observed secondary structure mentioned above, and the *second* set of events is a *prediction* of the secondary structure for the same protein.

In order to convert the *correlation* between the two sets into a *numerical value for the information* supplied by the prediction, we must consider the entropy of **two more sets** of events (sets 3 and 4). These two extra sets are derived from the original two sets. The **"third" set** of events describes the correlation between the observed and predicted secondary structure. (The third set is the joint probability distribution). As in the entropy discussion example above, there is one "event" for each residue. For simplicity we consider only 3 types of secondary structure, coil, strand, helix, abbreviated C, E, H. The possible types of event in this third series are observed C, predicted c; observed C, predicted e; observed C, predicted h; etc. Abbreviating the events in an obvious way, there are 9 types of event: Cc, Ce, Ch, Ec, Ee, Eh, Hc, He, Hh. This set has an entropy associated with it, calculated from the probability distribution of the 9 events, as specified by the equation above. The **"fourth" set** of events describes the probability distribution of the same 9 types of events as in the "third" series, but the probability distribution is the distribution that would result from a random pairing of events from the observed series with events from the predicted series. Again, there is one "event" for each residue in the protein. The entropy of this "fourth" set is exactly equal to the sum of the entropies of the original two series, and it is the maximum possible, given the two original series 11·13.

The mutual information provided by the *prediction* set about the *observed* set is given by the difference in entropy between the third and fourth series. The better the prediction is, the higher the correlation between the first and second set will be, and the lower the entropy of the third set will be, because most of the probability will be concentrated on the Cc, Ee, and Hh events; the incorrect prediction events (Ce, Ch, Ec etc) will be rare and have low probabilities, and contribute little to the entropy of the "third" set. A lower entropy for the third set corresponds to greater information provided by the prediction. The exponentiation of the information corresponds to the factor by which the prediction reduced the amount of choice/uncertainty regarding the secondary structure at a residue position. For example, in the 1/3 C, 1/3 E, 1/3 H observed structure example above, the naive uncertainty about the secondary structure at an arbitrary residue could be described with the value 3. A perfect prediction would eliminate the naivety and reduce the uncertainty, by a factor of 3, to 1, to a single choice, that is, to no uncertainty. Less than perfect information would give hints as to the correct answer, without eliminating uncertainty. For example, on the basis of our study of all the secondary structure predictions from CASP 1 through CASP 5, the top predictions of secondary structure improved from reducing uncertainty by a factor of 1.3 in CASP1 to reducing it by 1.7 in CASP5. In terms of the maximum possible information that could have been provided, the top predictions provided 27% of the possible information in CASP1, and had reached 51% by CASP5 15.

Given a set of observations and a scheme for predicting them, the numerical value of mutual information measures how good a hint about an upcoming event is given by the prediction scheme. In its exponentiated form, mutual information measures the reduction, provided by the prediction scheme, in the effective number of choices associated with one event, compared to chance expectation. Mutual information measures how much detail a prediction supplies.

However, applying an identical approach to 3D structure prediction is difficult; in contrast to the case with secondary structure prediction, there is no predetermined set of correct answers, no discrete set of right-or-wrong atom positions. No direct entropy estimate can be made for the observed structure ("first" series) nor for the predicted structure ("second" series).

Instead, a different approach provided a solution. The idea, "mutual information measures the ability of one set of numbers to predict another set compared to the ability expected from chance", gives a way to estimate the information provided by a 3D prediction scheme, despite the lack of discrete, right-or-wrong atom positions. The ratio of the *volume* within which a predicted Ca position lies relative to the *volume* within which it would fall at random is a measure of the information provided by a 3D prediction scheme. In terms of the discussion of secondary structure, this ratio corresponds to the exponentiated information for one residue position. In the case of the perfect secondary structure prediction scheme discussed, the equivalent ratio was 3. In the 3D cases we examined, the chance:predicted volume ratio typically is 1000:1 or higher. The numbers 3 and 1000 suggest the relative odds against guessing correctly in secondary structure and 3D structure problems, and therefore the relative difficulty of the two types of prediction problem.

To pursue the mutual information analogy an estimate is needed for the probability that two atoms will fall close to each other by chance. Choosing a *maximum* volume within which an atom may fall by chance is important, because the probability model breaks down if this volume is infinite. The details are described under Materials and Methods.

Materials and Methods

Superpositions

In order to evaluate the mutual information supplied by two 3D structures about each other, the two must be put into a common coordinate system and, in the case where the sequences are not identical, corresponding Ca atoms must be identified. The LGA protein structure superposition program 10, a program heavily used in the context of CASP structure comparisons, is an accepted solution to these problems 5–7. We used the LGA program for all superpositions in this work. Details of settings used are included in Appendix A in Supplemental Material.

Informative volume

In principle, an atom position placed by chance could fall anywhere in the universe; but as a matter of both common sense and mathematical utility, the maximum volume defining a chance atom placement should be limited. (Specifying a maximum volume captures the assumption that the atoms are connected and bond lengths are limited.) We required that, to provide any information at all, a predicted atom position had to be close enough to its target atom's position to fall within a sphere around the target atom with volume equal to the protein's volume (the total volume for the number of residues that get matched, that is). Thus, the "informative" volume was taken as a standard residue volume multiplied by the number of Ca positions that were matched by the superposition algorithm. Predicted atom

positions that are farther from their target than the radius of the informative volume contribute nothing to the information score but do not reduce it.

The standard residue volume was chosen empirically, as the average volume associated with a residue in the unit cell of an approximately hundred-amino acid protein, arbitrarily chosen as tuna cytochrome c, 16 3cyt, a per-residue volume of about 250 Å³, corresponding to an average of 9 non-hydrogen atoms per residue. We used this “natural averaging”, rather than creating our own average of individual residue volumes, for simplicity. This “protein-crystallographic” volume per residue includes the contribution of hydrogen atoms, water of hydration, and molecular packing effects. Our standard residue volume implies a corresponding atom volume of about 28 Å³.

Calculating mutual information

As a predicted position falls closer to its target’s position, it falls within a sphere of smaller volume. We took the ratio of the total volume of the protein to the volume of the “prediction sphere” as an indication of the informativeness of the predicted atom position. We used the logarithm of this ratio to define a measure, which we call “3Dbits”, of the informativeness of the prediction for this atom. Using logarithms to base 2 expresses the mutual information in bits, the standard form. The information provided by an entire prediction is the sum of the bit values for the predicted atoms. In other words, we gain one bit of information for a Ca position with each halving of its prediction volume relative to the whole volume of the residues matched by the superposition program.

Using a cube with half edge equal to the distance between the predicted and actual atom position rather than a sphere with that same radius simplifies the calculation. The effect of this simplification is to subtract a constant from the 3Dbit score for each atom, $(\log_2(8/((4/3)\pi))) = \sim 1$. This is acceptable since it is equivalent to modifying the choice for the informative volume, and that choice was arbitrary (though intuitively appealing), and every calculation is done the same way. An extensively commented version of the PERL script used for the calculation is provided as supplementary material, in a file, tsCommentsPerl.txt.

3Dbits vs 3DbitSum

In the secondary structure example discussed, the information was calculated with respect to a prediction for one residue, and also for one residue, in the discussion of 3D information. However, in the 3D case, the quality of a prediction depends not only on the closeness of superposition but also on the number of residues that match. On the other hand, a characteristic of 3D superpositions of molecules is that generally the more atoms that are superimposed the less precise the match becomes. There is a trade-off between quantity and quality of match. This trade-off is explicitly addressed by summing the 3Dbit values for all the matching residues. This approach causes predictions that include more atoms to supply more information, assuming equal rmsd (root mean square distance) for predictions being compared.

Mathematical Formulation for 3Dbits and 3DbitsSum

N = the number of Ca’s matched by the LGA superposition

V = the volume of a standard residue (250 Å³)

D_j = the distance between matched Ca pair j , where j runs from 1 to N

$$3Dbits_j = \log_2(N * V / (2D_j)^3)$$

for j such that $D_j < \text{half-edge of cube of volume } (N \cdot V)$; otherwise[†] $3\text{Dbits}_j = 0$

$$3\text{DbitSum} = \sum_{j=1}^N 3\text{Dbits}_j$$

Application to structures and comparisons with other similarity measures

The behavior of 3DbitSum was examined by applying it to several sets of structure comparisons: a set of globins, 33 single-domain template-based targets from CASP 4 and 6, 13 template-free targets from CASP 7, molecules related to these targets, and some sets of nmr structures. Direct comparisons with GDT_TS were made in selected cases. Along with 3DbitSum, a variety of other measures were calculated. Measures examined were rmsd, the root mean square deviation between C α atoms in superimposed structures; the % sequence identity for the superimposed structures; lgaS and lgaQ, two similarity measures from the LGA superposition program; the total number of C α s in each of the superimposed molecules, as well as the number of C α s superimposed; “bits/C α ”, which is 3DbitSum divided by the number of C α s superimposed; and “effective distance,” which is the side of a cube which represents the geometric-average uncertainty volume associated with the match between superimposed C α s. Effective distance is the mutual information analog of rmsd. Correlation coefficients between several different similarity measures were also calculated.

Additional details

Further description of the sets of molecules chosen for comparisons and details of structure superpositions is in Appendix A in Supplemental Material.

Results and Discussion

Measures of similarity

3DbitSum is an *extensive* quantity (in the thermodynamic sense) and increases with the number of atoms matched and is suitable for measuring the total information provided by a model or template. An *intensive* measure of the quality of match, which doesn't depend (at least not in the same way) on the number of residues in the match, is the “bits/C α ” score. Bits/C α is a way to compare different-length predictions on a per-residue basis.

How effective in increasing our knowledge of unknown structures are different methods of structure prediction, and how do they compare with experimental structural similarities?

Figure 1 plots the average and standard deviation of values of 3DbitSum for several kinds of comparison, arranged from most to least mutually informative. Figure 2 plots bits/C α for the same set of comparisons, and Table S-I lists values for these and several other comparison measures.

The data show that the structures that provide the most information – the most detail -- about each other are independent copies of identical protein chains in the asymmetric unit of a crystal structure. The sample of 52 such pairs used here averaged nearly 4000 bits of mutual information. The average bits/C α was 19.5, corresponding to an average effective distance of 0.56 Å, an rmsd of 0.46 Å, and an average number of residues matched of 149.

[†]With 4 Å as the LGA superposition cutoff distance (“d:4”), in mode 4, sequence-independent mode, no predicted atom positions have any mismatches great enough to invoke this rule. However, in mode 3, sequence-dependent mode, using a 5 Å cutoff, this rule did come into play.

For less similar molecules -- 21 pairs of homologs with an average sequence identity of about 50% -- the mutual information dropped to about 2400 bits, the average bits/C α to 13.1, average effective distance 2.14 Å, rmsd 1.45 Å, for an average number of residues matched of 176. For the least accurate models -- 13 top GDT_TS-score single-domain free models from CASP7 compared to native -- the mutual information fell to about 500 bits, average bits/C α was 8.4, effective distance 7.26 Å, rmsd 2.65 Å, average number of residues matched was 64.

For the well-studied globin family, the two most similar structures examined here (99% sequence identity) provided about 2500 bits of information about each other (Bits/C α = 17.0, effective distance 0.67 Å, rmsd 0.54 Å, 152 residues superimposed) while the *least* similar globin structures, soybean and sea-hare, provided about 965 bits of mutual information (Bits/C α = 8.2, effective distance 4.58 Å, rmsd 2.54 Å, 117 residues superimposed). Soybean (1bin) and sea-hare (1mba) sequences are so different (15% sequence identity) that a PSI-blast search with one sequence cannot find the other.

As might be expected, the least amount of information is provided by secondary structure predictions, about 120 bits on average, for CASP1–5 15.

The most similar structures provide an estimate of the noise level in structures

The variation shown by pairs of identical molecules in the same asymmetric unit indicates the “noise level” in protein conformation, and implies the highest accuracy to be expected from predictions. From the 52 pairs of identical molecules discussed above, the average upper limit of predictability appears to be about 20 bits/C α $\pm$ 5, effective distance of 0.6 Å $\pm$ 0.3, with rmsd of 0.5 Å $\pm$ 0.2 (see Table S-I).

Similarity of NMR models

In Figure 2., the downward trend of bits/C α follows 3DbitSum for the chosen ordering of the collections, except for nmr models, where it is relatively higher. What this means is that the nmr models examined were “intrinsically” more similar (14.5 bits/C α) than our selected CASP4 & 6 templates, targets, and predictions (~11.3 bits/C α), but were shorter on average (103 residues compared to 180 and 145), because nmr has more stringent length restrictions than crystallography.

Template-based predictions

The single-domain template-based predictions provided about 2000 and 1500 bits of information, respectively, about their targets in CASP4 and CASP6. (The average target length was greater in the CASP4 sample, so that the mutual information from unmodified templates was greater than in CASP6.) However, in CASP4, the best *predictions* provided slightly less information on average than the best *templates*, whereas in CASP6, the best predictions provided about 90 bits more information about their targets on average than the best templates did. This amounted to about 0.3 bits/C α more for the predictions in CASP6 than for the best templates.

From a different point of view -- in CASP4, only 3 of 7 targets had predictions with a 3DbitSum score higher than the best pre-existing PDB structure, whereas in CASP6, 20 of 26 targets had predictions this good. Using a binomial test for proportions, with

$$p_1 - p_2 = (20/26 - 3/7),$$

and the standard deviation, σ , of this difference approximated by the square root of

$$(p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2),$$

where n_1 is 26 and n_2 is 7, we found

$$(p_1 - p_2)/\sigma = 1.666$$

The value 1.666 corresponds to a significance level of 4.8%. That is, the degree of improvement seen in CASP6 would be expected by chance in slightly fewer than 1 in 20 analyses. Although this is not highly significant, it does suggest measurable progress. If progress were significant at the 1% level, it might not be subject to question in the first place. Additional detail about the CASP4 and CASP6 comparisons is provided in Supplemental Table S-II and Supplemental Table S-III.

Sequence and structural similarity can be measured on the same scale

Because the E-values and bit scores that are reported for sequence matches in BLAST searches are also based on estimates of the probability that such matches would occur by chance, the bit scores from BLAST searches can be directly compared with 3DbitSum scores. Maximum BLAST bit scores found for perfectly matching hits for protein sequence searches done in connection with this work were around 300 bits for sequences of about 150 residues (globin sequences provided one example - data not shown). Comparing 300 bits for sequence matching with 2500 bits for structure matching (closest globin example) sheds quantitative light on the widely understood idea that structure is better preserved during evolution than sequence.

It must be true that the entirety of the information carried by an amino acid sequence is not captured by letter-to-letter comparisons of sequences, thus limiting the recognition reach of sequence comparisons. In contrast, structure expresses spatial correlations between residues, and that factor significantly extends the recognition reach of structure comparison. 3DbitSum provides some quantitative indication of the amount of additional information that evidently is “hidden” in correlations between residues in sequences. An interesting possibility is that template-free modeling might be a more mature method to recognize spatial correlations between different parts of amino acid sequences than any algorithm that depends only on sequence pattern recognition. A different approach to the question of spatial correlations between residues involves contact potentials. In the context of discriminating among correct and incorrect decoys by threading, a recent comprehensive examination of contact potentials in terms of information theory found the equivalent of between 0.03 and 0.35 bits/contact, for the range of contact criteria examined.¹⁷

Comparison of 3DbitSum with other measures of similarity

Twenty-four diverse globins were compared in nine ways with sperm whale myoglobin. Results are shown in Supplemental Table S-IV and Supplemental Table S-V and discussed in the Appendix (Supplemental material). The overall average correlation between pairs of measures was 0.76.

Two heavily used measures of progress in protein structure prediction are the root mean square deviation (rmsd) in distance between superimposed Ca atoms, and the GDT_TS score.¹⁰ These measures, especially GDT_TS, are widely used in judging the quality of predictions.⁸⁻⁹ GDT_TS has an important feature that makes it a significant step forward from rmsd for evaluating structural similarity. It is constructed to give greater influence to closer matches, in contrast to rmsd, which is more sensitive to larger mismatches. However,

we believe both GDT_TS and rmsd scores are subtly biased in favor of comparisons of smaller structures; that is, comparisons of smaller structures tend to have better scores than comparisons of larger structures. In contrast, we believe that it is more difficult to correctly predict more atom positions than fewer, and that similarity scores should be higher for cases where more residues are superimposable.

In addition to rmsd and GDT_TS, the LGA program reports less often discussed measures, lgaS3, a more complex but more sensitive measure of structural similarity than GDT_TS, which has an extremely high (>0.9) correlation with GDT_TS, and lgaQ, a measure meant to indicate the quality of a superposition. We found a high correlation (>0.8) between lgaS3 and bits/C α , and an even higher correlation between lgaQ and 3DbitSum. These high correlations suggest to us that GDT_TS is measuring an intensive quantity (like bits/C α does), and that lgaQ is measuring an extensive quantity (like 3DbitSum does.)

Superposition length dependence of similarity measures

The main reason for our belief that similarity scores should be higher for larger superpositions is the observation that superimposing more atoms means that there are more conditions to satisfy with the same number of rotation and translation variables. A possible counter argument is that there are more choices for locating superimposable stretches of chain in longer structures. This is an interesting question to raise. We found simple probabilistic models were inadequate to answer the question, so we examined the question empirically.

If comparing longer chains makes it easier to find superpositions of fragments, the absolute number of residues superimposed ought to increase with chain length. For the 13 CASP7 single-domain free modeling targets, we calculated the average number of residues superimposed for the top 200 GDT_TS predictions for each target, and calculated the correlation between this average "N" and the number of residues in the "fixed" chain (recommended to be chosen as the shorter chain). The correlation was only 0.30. Empirically this provides little support for its being easier to predict a given number of C α atoms in a large structure than in a small structure.

On the other hand, we found for our set of globin comparisons (a more uniform set of molecules) that the correlation between rmsd and number of residues superimposed was -0.69 (Table S-V). Since GDT_TS depends on the *percent* of residues that are superimposable within defined distance constraints, GDT_TS trends downward for larger structures.

Advantage in behavior for 3DbitSum vs GDT_TS and rmsd in comparing predictions for different targets in the same class

As mentioned, the measures rmsd and GDT_TS do not vary with number of residues successfully superimposed in a way that allows for unbiased comparisons between the quality of predictions for different targets. Table I shows the results of sequence-dependent (GDT_TS-calculating) superpositions for 13 single-domain CASP 7 template-free targets and their top GDT_TS score predictions, as determined through the CASP 7 website Table Browser (<http://www.predictioncenter.org/casp/casp7/public/cgi-bin/tablebrowser.cgi>). Examination of the target-length dependence of GDT_TS and the information measure shows opposite trends. GDT_TS has a correlation coefficient of -0.66 with number of residues superimposed, whereas 3DbitSum has a correlation coefficient of $+0.77$ with number of residues superimposed, which is a more desirable behavior, given the greater difficulty of predicting longer targets. This difference has little bearing on comparisons of

different predictions for the same target, of course, which is the context in which GDT_TS is primarily used. But it still represents a limitation, as shown by the following examples.

Behavior for low-end GDT_TS values

The results in Table I are sorted by GDT_TS score, highest first. Table lines 8–10 each show almost identical rmsd values of about 2.93 Å, so that the rms deviations are the same. However, the lengths of the targets and the number of C α superimposed differ. Line 8 (T0309) reports only **29** GDT_TS-acceptable superimposed Cas (and 39 Cas superimposed inside the informative volume) from a target of length 62, and a GDT_TS score of **36.7%**. Line 9 (T0361) has **79** GDT_TS-acceptable superimposed Cas (and 128 Cas superimposed inside the informative volume) from a target of length 158, and a slightly lower GDT_TS score of **35.8%**. But clearly the longer prediction provides more information in an absolute sense, and it would be good for this to be reflected in the prediction scores for the two models. The information score, 3DbitSum, does reflect this, yielding **211 bits** for the first case and **810 bits** in the second case. The prediction for T0350 (line 1) has the highest GDT_TS score (**63.7%**) in the set, but although it is certainly good, it predicts only **85** Cas (91 inside the informative volume) and provides less information in an absolute sense than the prediction for T0361 (line 9), which is a more demanding prediction target. The 3DbitSum for T0350 (line 1) is **647 bits**, which is high but not as high as for T0361 (line 9).

High-end GDT_TS

The template-free models just discussed tend toward the lower end of GDT_TS scoring. A second comparison of GDT_TS and 3DbitSum is at the very high end of GDT_TS scoring. It is based on the 25 nmr structures examined in the original LGA paper¹⁰. The LGA superpositions were repeated with the same parameters as the original work (Table 2 in that reference), and 3DbitSum was also calculated. Results are shown in Table II. The nmr structures are very similar according to GDT_TS, with scores ranging between 97% and 93%. Rmsd ranges between 0.70 and 1.14 Å. 3DbitSum has a much larger range than GDT_TS, 2247 to 1439, and no (0.07) correlation with GDT_TS, probably because the structures are barely distinguishable by GDT_TS. (IgaQ, however, has a broader range and a 0.81 correlation with 3DbitSum.) Another interesting comparison is the correlations of 3DbitSum and GDT_TS with rmsd and the number of residues superimposed. GDT_TS has a correlation of 0.67 with the number of residues superimposed, but a correlation of only -0.34 with rmsd. In contrast, 3DbitSum has little correlation (-0.15) with the number of residues superimposed, but a correlation of -0.78 with rmsd. The range of residues superimposed is 135 to 129 out of 135. It could be argued that the range of 3DbitSum does better reflect the differences among the nmr structures, because the estimated mis-match volume associated with an rmsd of 1.14 Å (Table II entry 20), is 4.3 times the mis-match volume associated with an rmsd of 0.70 Å (Table II entry 3). If the different nmr structures were template-based models, a model with rmsd of 0.7 Å would clearly be better than one with rmsd of 1.14 Å. Thus, 3DbitSum provides discrimination beyond the power of GDT_TS at both ends of the spectrum of similarity.

Visual assessment of template-free models

We have presented evidence that 3DbitSum has useful properties for comparing model quality for different targets. We were also interested in whether 3DbitSum might be an effective evaluation score *within* a set of models for a single target. Because template-free models are still of substantially lower accuracy than template-based models, assessment of their model quality still involves visual assessment by expert humans. 18 The best template-free models as determined by human assessment in CASP7 generally have GDT_TS ranks in the top 5 to 10% of models, but often do not have the very highest rank. We calculated 3DbitSums for all of the template-free models for the same 13 single-domain targets we had

examined above, and examined the ranks on various scales for the 33 assessor-designated best models for them. (Details in Appendix A, supplemental material.) Results are shown in Supplemental Table S-VI.

The average GDT_TS rank of the 33 assessor-designated best models was 12 $\pm$ 14; their average 3DbitSum rank was 17 $\pm$ 18, their average rank for the number of residues superimposed under 4 Å was 21 $\pm$ 22, and for the number of residues superimposed within the informative distance, 24 $\pm$ 44. However, the average rank of their rmsd value was 255 $\pm$ 116. Since each target had around 500 models submitted, the average rank of the visual assessors-best models was at the 95th percentile or better, except for rmsd, for which the average rank was near the 50th percentile. Models with low rmsd values tended to have shorter sections superimposed. When the assessors-best models were compared with the top GDT_TS score models, in about 2/3 of cases, the assessors-best models ranked higher on at least one of the length rankings. This observation suggests that assessors often valued models that fit moderately well over large lengths more than they valued models with closer fits over shorter lengths, although not invariably (T0307TS178_2).

Three targets (300, 304, and 361) had at least one model with a 3DbitSum score that was substantially higher than any score for the assessors-best models. We visually examined the high 3DbitSum models and the assessors preferred models. For target 304, assessor-preferred model T0304TS564_4 was clearly better: it had a stretch of residues in a good conformation that was somewhat mis-oriented relative to the native structure, reducing the superposition-dependent scores. The reasons for the assessors' preferences were less clear-cut in other cases, but still we felt convinced that human evaluation is necessary to choose among the top-scoring template-free models, even though GDT_TS (or 3DbitSum, or, potentially, lgaQ) is an effective screening tool.

Conclusions

3DbitSum expands the range of possible comparisons to new territory

In light of the acknowledged difficulty 8·9 of evaluating the quality of predictions, there are three useful results from the concepts presented here. The first is the ability to compare different kinds of predictions and different targets in a common framework, on a common scale. An information-based measure can not only usefully compare predictions for different length targets, it can usefully compare predictions made using radically different methods, ranging from template-free modeling to secondary structure prediction. The second is the ability to distinguish the performance of predictions at the very high end of success, which is important in template-based predictions. The third is the ability of 3DbitSum to provide a more nuanced evaluation of high-difficulty predictions such as with template free modeling.

The “common framework” capability is a result of the nature of information – something that restricts the number of acceptable choices that can be made among possible answers. Unlike rms differences in coordinates, a concept that applies only to coordinate comparisons, or SOV and GDT_TS 10·19, important and ingenious concepts that apply respectively only to secondary structure prediction and 3D protein structure comparisons, information is a general concept that can be applied to any sort of prediction or comparison system.

Nevertheless, it is noteworthy that the relative content of the theoretically based measures 3DbitSum and bits/C α could be captured in significant degree by empirical measures like lgaQ and GDT_TS (and lgaS and lgaS3).

An understanding of the *conceptual meaning* of information, as opposed to a purely mathematical understanding, made it possible to expand the use of information to a problem area where it had not been applied before, and to address existing difficulties in the evaluation of protein structure predictions. The information approach brings all methods of structure prediction into a common evaluation framework, and this framework is furthermore compatible with existing methods of amino acid and nucleic acid sequence comparison.

The common evaluation framework shows quantitatively that in an absolute sense *template-free* modeling provides the greatest increase in information from prediction methods (500 bits per structure, on average, in CASP7). Even though *template-based* models are substantially superior to *template-free* models, the information *added* by *template-based* modeling was only about 90 bits on average in CASP6, and was none at all in CASP4. For comparison, without any modeling at all, the information provided by one structure about another with clear sequence homology (average 50% sequence identity) was about 2400 bits for the examples in this study.

Basing the evaluation of predictions on the degree to which they reduce uncertainty about the details of their targets provides a meaningful standard of comparison over many different forms of prediction and comparison, and such an approach overcomes some limitations of existing methods of comparison.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Adam Zemla for access to and help with the LGA program and for valuable discussion of the role of chain length in structure similarity measures, and of the similarity measures lgaS and lgaQ, and we thank Ioannis Kagiampakis for help in setting up the automated downloads from the CASP website. MV is supported by NIH/NHGRI grant R01HG003319 and by NSF/DMS grant number DMS-0605001. JWT is supported by NIH grant R01 GM81631-01.

References

1. Moulton J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. *Proteins*. 1995; 23(3):ii–v. [PubMed: 8710822]
2. Moulton J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins*. 1997 Suppl 1:2–6. [PubMed: 9485489]
3. Moulton J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins*. 1999 Suppl 3:2–6. [PubMed: 10526346]
4. Moulton J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins*. 2001 Suppl 5:2–7. [PubMed: 11835476]
5. Moulton J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*. 2003 Suppl 6:334–339. [PubMed: 14579322]
6. Moulton J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)--round 6. *Proteins*. 2005; 61 Suppl 7:3–7. [PubMed: 16187341]
7. Moulton J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction-Round VII. *Proteins*. 2007; 69(S8):3–9. [PubMed: 17918729]
8. Kryshtafovych A, Venclovas C, Fidelis K, Moulton J. Progress over the first decade of CASP experiments. *Proteins*. 2005; 61 Suppl 7:225–236. [PubMed: 16187365]

9. Kryshchuk A, Fidelis K, Moult J. Progress from CASP6 to CASP7. *Proteins*. 2007; 69(S8):194–207. [PubMed: 17918728]
10. Zemla A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003; 31(13):3370–3374. [PubMed: 12824330]
11. Jaynes, ET. Confidence Intervals vs Bayesian Intervals. In: Rosenkrantz, RD., editor. *ET Jaynes: Papers on Probability, Statistics and Statistical Physics*. DReidel Publishing Co: Dordrecht, Holland; 1983. p. 149-209.
12. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*. 1993; 232(2):584–599. [PubMed: 8345525]
13. Baldi, P.; Brunak, S. *Bioinformatics: The Machine Learning approach*. Cambridge, MA: MIT Press; 1998.
14. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000; 16(5):412–424. [PubMed: 10871264]
15. Swanson R, Kagiampakis G, Tsai J. An information measure of the quality of protein secondary structure prediction. *J Comput Biol*. 2008; 15(1):1–15. [PubMed: 18257674]
16. Swanson R, Trus BL, Mandel N, Mandel G, Kallai OB, Dickerson RE. Tuna cytochrome c at 2.0 Å resolution. I. Ferricytochrome structure analysis. *J Biol Chem*. 1977; 252(2):759–775. [PubMed: 13079]
17. Solis AD, Rackovsky S. Information and discrimination in pairwise contact potentials. *Proteins: Structure, Function, and Bioinformatics*. 2007; 9999(9999) NA.
18. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. *Proteins*. 2007; 69 Suppl 8:57–67. [PubMed: 17894330]
19. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*. 1999; 34(2):220–223. [PubMed: 10022357]
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res*. 2000; 28(1):235–242. [PubMed: 10592235]
21. Holmes JB, Tsai J. Characterizing conserved structural contacts by pair-wise relative contacts and relative packing groups. *J Mol Biol*. 2005; 354(3):706–721. [PubMed: 16269154]
22. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. *Proteins*. 2001 Suppl 5:22–38. [PubMed: 11835479]
23. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins*. 2005; 61 Suppl 7:27–45. [PubMed: 16187345]

Figure 1. 3DbitSum: averages and standard deviations of 3DbitSum are shown for several different groups of comparisons, arranged from most informative to least informative. All comparisons used in making this figure are between molecules related to one of the CASP targets studied in this paper or else in the set of globin structures. For a detailed description of the sets of comparisons that underlie the averages plotted, see the end of Appendix A (Supplementary material).

Figure 2. Bits/Cα: averages and standard deviations of Bits/Cα are shown for the same sets of comparisons as shown in Figure 1. Bits/Cα for CASP6 predictions is slightly higher than for CASP4 predictions, reflecting slightly better prediction quality, even though the total information supplied on average by templates was higher for CASP4 than for CASP6, as explained in the detailed description of comparison sets at the end of Appendix A (Supplementary material).

Table 1
3DbitSum scores for the best CASP7 single-domain free predictions compared with native, sorted by GDT_TS score.

3Dbit Sum	bits/C α	# C α sup'd (1)	# C α informative (2)	N1	N2	rmsd Å	eff dist Å	gdt_ts	lgaS3 (3)	lgaQ (4)	Comparison ID (5)
647	7.1	85	91	91	91	2.93	5.48	63.74	62.74	2.80	t350-113_1.2hc5_a
382	6.3	48	55	61	61	2.47	5.84	59.84	57.37	1.87	t348-026_5.2hfl_a
867	7.3	105	119	119	119	2.88	5.76	59.66	59.24	3.53	t382-024_3.2t9c_a
550	6.2	61	88	89	89	2.84	6.75	55.90	53.45	2.08	t300-033_2.2h3r_a
475	5.7	55	71	83	83	2.70	7.32	51.51	47.63	1.97	t353-024_1.2hfg_a
521	5.2	65	88	101	101	3.16	8.92	46.04	42.19	2.00	t304-015_1u_2h_28_a
565	4.6	62	90	123	123	2.73	10.84	39.02	37.46	2.19	t307-020_4.2h5n_a
211	3.4	29	39	62	62	2.93	11.37	36.69	34.78	0.96	t309-034_1.2h4o_a
810	5.1	79	128	158	158	2.92	10.42	35.76	32.60	2.61	t361-010_1.2hkt_a
485	3.6	54	85	135	135	2.93	14.10	31.11	27.71	1.78	t319-020_3.2j6a_a
365	3.5	43	78	103	103	3.39	13.02	30.58	26.47	1.23	t314-025_5.2hg6_a
486	3.2	51	128	152	152	3.95	16.07	28.29	23.82	1.26	t287-025_2.2g3v_a
1080	2.6	98	151	413	413	3.14	25.63	16.28	14.54	3.03	t296-020_3.2ha9_a

(1) # C α sup'd is the number of C α superimposed as reported by the LGA program, mode=3, d=5 Å.
(2) # C α informative is the number of C α superimposed within the "informative distance" ($< 1/2$ max edge)
(3) lgaS3 is a measure produced by the LGA superposition program, which correlates strongly with GDT_TS and is a more sophisticated and sensitive measure of structural similarity than GDT_TS.
(4) lgaQ is a "quality of superposition" measure produced by the LGA superposition program, using the formula $0.1 * N / (0.1 + \text{RMSD})$. It correlates strongly with 3DbitSum.
(5) ComparisonID identifies target#_group#_model#_targetpdbID_chain for the prediction with the highest gdt_ts score for each single-domain free model target on the CASP7 website.

Table II
Comparison of 3DbitSum and gdt_ts for 25 nmr structures, examined by Zemla 10

3Dbit Sum	bits/C α	# C α sup'd (1)	# C α Inform ative (2)	N1	N2	rmsd Å	eff dist Å	gdt_ts	lgaS3 (3)	lgaQ (4)	Comparison ID (5)
2067	15.3	135	135	135	135	0.79	0.94	96.85	98.58	15.18	1m2f-m08_1m2e
1771	13.1	133	135	135	135	0.93	1.56	96.30	97.99	12.90	1m2f-m01_1m2e
2247	16.6	133	135	135	135	0.70	0.69	96.30	98.11	16.63	1m2f-m16_1m2e
1578	11.7	135	135	135	135	0.91	2.17	96.30	98.27	13.31	1m2f-m02_1m2e
1821	13.5	134	135	135	135	0.84	1.43	96.11	98.16	14.27	1m2f-m11_1m2e
1635	12.1	134	135	135	135	0.95	1.97	96.11	97.98	12.70	1m2f-m19_1m2e
1941	14.4	133	135	135	135	0.80	1.17	95.93	97.99	14.78	1m2f-m17_1m2e
1499	11.1	133	135	135	135	0.94	2.48	95.93	97.98	12.83	1m2f-m20_1m2e
1696	12.6	132	135	135	135	0.85	1.77	95.74	97.75	13.87	1m2f-m14_1m2e
1512	11.2	131	135	135	135	0.95	2.43	95.56	97.34	12.48	1m2f-m13_1m2e
1570	11.6	134	135	135	135	1.04	2.20	95.56	97.80	11.74	1m2f-m05_1m2e
1543	11.4	135	135	135	135	1.09	2.30	95.37	97.73	11.34	1m2f-m10_1m2e
1699	12.6	133	135	135	135	0.92	1.77	95.37	97.60	13.00	1m2f-m12_1m2e
1652	12.2	134	135	135	135	0.89	1.91	95.37	97.79	13.51	1m2f-m18_1m2e
1538	11.4	130	135	135	135	0.85	2.33	95.37	97.38	13.74	1m2f-m21_1m2e
1592	11.8	133	135	135	135	0.96	2.12	95.37	97.55	12.58	1m2f-m07_1m2e
1928	14.3	131	135	135	135	0.85	1.19	95.19	97.20	13.82	1m2f-m22_1m2e
1717	12.7	133	135	135	135	0.89	1.71	95.19	97.66	13.50	1m2f-m24_1m2e
1518	11.2	134	135	135	135	0.94	2.41	95.00	97.44	12.90	1m2f-m25_1m2e
1439	10.7	132	135	135	135	1.14	2.75	95.00	97.30	10.64	1m2f-m09_1m2e
1968	14.6	130	135	135	135	0.80	1.11	94.81	97.08	14.46	1m2f-m15_1m2e
1616	12.0	130	135	135	135	1.01	2.03	94.26	96.56	11.71	1m2f-m04_1m2e
2124	15.7	129	135	135	135	0.74	0.85	93.89	96.37	15.29	1m2f-m03_1m2e
1871	13.9	132	135	135	135	1.00	1.31	93.70	96.64	11.98	1m2f-m23_1m2e

(1) # C α sup'd is the number of C α superimposed as reported by the LGA program

(2) # C α informative is the number of C α superimposed within the "informative distance" ($< 1/2$ max edge)

(3) lgaS3 is a measure produced by the LGA superposition program, which correlates strongly with GDT_TS and is a more sophisticated and sensitive measure of structural similarity than GDT_TS.

(4) IgaQ is a "quality of superposition" measure produced by the LGA superposition program, using the formula $0.1 * N / (0.1 + \text{RMSD})$. It correlates strongly with 3DbitSum.

(5) ComparisonID identifies the pdbID and nmr model number and the standard of comparison, pdbID 1m2e.