

MaxSub: an automated measure for the assessment of protein structure prediction quality

Naomi Siew¹, Arne Elofsson², Leszek Rychlewski³ and Daniel Fischer^{4,*}

¹Department of Chemistry, Ben Gurion University, Beer-Sheva 84015, Israel,

²Stockholm Bioinformatics Center, 106 91 Stockholm University, Sweden,

³International Institute of Molecular and Cell Biology, Ks. Trojdena 4, 02-109 Warsaw, Poland and ⁴Department of Bioinformatics Computer Science, Ben Gurion University, Beer-Sheva 84015, Israel

Received on February 17, 2000; revised and accepted on May 5, 2000

Abstract

Motivation: Evaluating the accuracy of predicted models is critical for assessing structure prediction methods. Because this problem is not trivial, a large number of different assessment measures have been proposed by various authors, and it has already become an active subfield of research (Moult et al., 1999). The CASP (Moult et al., 1997, 1999) and CAFASP (Fischer et al., 1999) prediction experiments have demonstrated that it has been difficult to choose one single, 'best' method to be used in the evaluation. Consequently, the CASP3 evaluation was carried out using an extensive set of especially developed numerical measures, coupled with human-expert intervention. As part of our efforts towards a higher level of automation in the structure prediction field, here we investigate the suitability of a fully automated, simple, objective, quantitative and reproducible method that can be used in the automatic assessment of models in the upcoming CAFASP2 experiment. Such a method should (a) produce one single number that measures the quality of a predicted model and (b) perform similarly to human-expert evaluations.

Results: MaxSub is a new and independently developed method that further builds and extends some of the evaluation methods introduced at CASP3. MaxSub aims at identifying the largest subset of C_α atoms of a model that superimpose 'well' over the experimental structure, and produces a single normalized score that represents the quality of the model. Because there exists no evaluation method for assessment measures of predicted models, it is not easy to evaluate how good our new measure is. Even though an exact comparison of MaxSub and the CASP3 assessment is not straightforward, here we use a test-

bed extracted from the CASP3 fold-recognition models. A rough qualitative comparison of the performance of MaxSub vis-a-vis the human-expert assessment carried out at CASP3 shows that there is a good agreement for the more accurate models and for the better predicting groups. As expected, some differences were observed among the medium to poor models and groups. Overall, the top six predicting groups ranked using the fully automated MaxSub are also the top six groups ranked at CASP3. We conclude that MaxSub is a suitable method for the automatic evaluation of models.

Availability: MaxSub is available at: <http://www.cs.bgu.ac.il/~dfischer/MaxSub/MaxSub.html>

Contact: {nomsiew,dfischer}@cs.bgu.ac.il;

arne@biokemi.su.se;

leszek@iimcb.gov.pl

Supplementary Information: Full tables are available at: <http://www.cs.bgu.ac.il/~dfischer/MaxSub/MaxSub.html>
CASP web site: <http://PredictionCenter.llnl.gov/casp3/>
CAFASP web site: <http://www.cs.bgu.ac.il/~dfischer/CAFASP2>

Introduction

Predicting the three-dimensional (3D) structure of a protein of unknown structure from its amino acid sequence is one of the most important current problems of modern biology. A large number of protein models are being built each year, using various methods, which include homology modeling, fold recognition or threading, and *ab initio*. But how accurate are these models? The CASP (Critical Assessment of Structure Prediction; Moult et al., 1997, 1999) and CAFASP (Critical Assessment of Fully Automated Structure Prediction; Fischer et al., 1999) experiments try to address the problem of comparing and

*To whom correspondence should be addressed.

evaluating the models, in order to learn how accurate and reliable the model producing methods are. Predictions are filed by predicting groups on target proteins of unknown structure. Subsequently, when the experimental 3D structures of the targets are determined, the accuracy of the predicted models is evaluated. However, it has become clear that evaluating predicted models is a very difficult task, and this has become an active subfield of research. Ideally, one would like to use an evaluation method that (a) is fully automated, (b) produces one single numerical measure representing the quality of the model, (c) is simple, intuitive and easy to understand, (d) most researchers in the field agree upon, (e) is applicable across the different prediction categories (namely, homology modeling, fold recognition and *ab initio*), (f) is normalized in such a way that it can be added over all targets to produce a total score for each group, and (g) is independent of a pre-computed list of 'correct hits'. Unfortunately, as of today, no such method exists.

The difficulty in assessing models is that simple measures, such as the root mean square deviation (RMSD) computed over all atoms, is a very poor indicator of the quality of a model when only parts of the model are well predicted. The wrongly predicted regions produce such a large RMSD that it is impossible to know if the model contains 'well-predicted' parts at all. Thus, what is needed is a way to identify the 'well-predicted' regions only, and give them a score that represents how close to the experimental structure these regions are. However, defining and finding these regions are difficult problems, conceptually and computationally.

Because it is not clear whether there exists a single, objective, quantitative assessment method based on one single measure, in the CASP experiments a large number of methods and measures have been used, including human visualization and subjective evaluation (see below). For CASP, assessment techniques have been developed by Bryant and colleagues (Moult *et al.*, 1997; Marchler-Bauer and Bryant, 1999), Sippl and colleagues (Lackner *et al.*, 1999; Hubbard, 1999), the CASP Prediction Center (Zemla *et al.*, 1999) and others. Usually, these methods produce a set of numbers that need to be subsequently analyzed by hand. However, for an automated evaluation, it is desirable to have one single scalar representing the quality of a model. The CASP3 assessment was thus based upon the above measures, other measures developed by the assessors themselves (Jones and Kleywegt, 1999; Murzin, 1999; Orengo *et al.*, 1999) and upon human observation. Thus, it is clear that we do not yet have an 'ideal', automated evaluation method. However, it is also clear that efforts towards the adoption of more or less 'standard' methods are needed (Moult *et al.*, 1999).

Here we present an automated method that attempts to fulfill this need. Rather than simply adding yet another

evaluation method to the already large number of methods, we (a) concentrate on a method that produces a single number measuring the quality of a model and (b) show that a fully automatic evaluation based on this number identifies the same top groups and models as those identified by human-expert evaluations. Although an exact, direct comparison of the results of automated methods vis-a-vis the human evaluation is not straightforward, the human-expert evaluation can be a useful standard against which methods are compared. A method lacking the biological insights, expertise and intuition of a human, but whose performance is similar to a human-expert evaluation is thus an achievement, and can safely be used in the automated evaluation of prediction experiments such as CAFASP.

The variety of existing assessment methods reflects the different aspects that can be evaluated in a prediction. For example, in the homology modeling category, one may wish to evaluate the accuracy of the loops or the side chain packing (Zemla *et al.*, 1999; Jones and Kleywegt, 1999). Or in the fold-recognition category, one may be interested in evaluating whether a compatible fold was recognized, regardless of the quality of the alignment obtained (Fischer *et al.*, 1999). Here we concentrate on evaluation methods that focus on the alignment quality of the models only.

Mainly two types of assessment methods have been used, which, in CASP jargon, are called 'sequence-independent' and 'sequence-dependent' (Moult *et al.*, 1999). In the former, the structural similarity of the predicted model (herein referred to as 'the model') and the experimental structure is measured, without requiring that each model residue be structurally matched to its corresponding residue in the experimental structure. The displacement of a model residue from its corresponding residue in the experimental structure, as measured from the best structural alignment, is referred to as a 'shift' (Marchler-Bauer and Bryant, 1999; Lackner *et al.*, 1999). In the latter (sequence-dependent assessment), only corresponding residues are compared (Hubbard, 1999; Zemla *et al.*, 1999). Thus, this is a stricter assessment criterion.

The method of Sippl and colleagues (Lackner *et al.*, 1999) is a sequence-independent method and is based on the structural superposition of the model over the experimental structure. From this structural superposition, a set of numbers is generated, which include the number of equivalent residues of the optimum match and the number of residues aligned at shifts zero, one, five, and above five, plus the average over the shifts. A similar method based on structural superposition that measures both the shift error and 'contact specificity' was developed by Bryant and colleagues (Moult *et al.*, 1997; Marchler-Bauer and Bryant, 1999). These sequence-independent methods award credit to fold-recognition predictions that

resemble the correct fold, but in which, due to errors in the alignment methods, some fragments may have been displaced (i.e. inaccurate alignments). Being based on structural superposition, these methods suffer from some of the limitations inherent in structural superposition programs (Lackner *et al.*, 1999), such as the need of a similarity score definition or the need of predefined thresholds, among others.

Another set of approaches is based on the sequence-dependent alignment, where each predicted residue is compared to its corresponding residue in the experimental structure. Sequence-dependent approaches are more strict in their evaluation criteria, and in particular, fold-recognition models generated with a correct fold but with a wrong alignment can score poorly. In addition, the sequence-dependent approach is a simpler and more straightforward measure of similarity between a model and an experimental structure. Hubbard's RMS/Coverage graphs (Hubbard, 1999) is a sequence-dependent method that samples the best RMSD from a large number of structural superpositions, each having a different number of equivalent residues. The graphs plot the best RMSD values against the number of equivalent residues. The graphs can then be judged by manual inspection. A related method named GDT was developed by Adam Zemla as part of the evaluation tools devised for CASP3 (Zemla *et al.*, 1999). GDT is aimed at identifying any accurately, not necessarily contiguous, predicted substructures. GDT attempts to find the maximum number of predicted residues that can be superimposed over the experimental structure within a given threshold. Because each model residue lies at a distance below the given threshold, the resulting RMSD of the superimposed residues is always smaller than the given threshold. GDT's approach corresponds in our minds to the notion of identifying the largest 'well-predicted' subset in the model based on given constraints. However, because finding the largest 'well-predicted' subset is a hard problem, approximations and/or heuristics need to be used.

For other evaluation methods used in CASP3, we refer the reader to the reports of the homology modeling (Jones and Kleywegt, 1999) and the *ab initio* (Orengo *et al.*, 1999) assessors, and to the papers of Sippl and colleagues (Sippl *et al.*, 1999) and Venclovas and colleagues (Venclovas *et al.*, 1999).

Here we propose an independently developed and new measure called MaxSub, which is based on similar principles as GDT. Briefly, MaxSub computes a single scalar in the range of 0 to 1, which measures the similarity of a model to its corresponding experimental structure (0 for a completely wrong model, 1 for a perfect model). The scalar is a normalization of the size of the largest 'well-predicted' subset and is computed using a variation of a formula suggested by Levitt and

Gerstein (1999). In the Methods section we describe the algorithm and the normalization procedure, and in the Results section we present the results of applying MaxSub to the CASP3 fold-recognition models, and show that our automated evaluation is in good agreement with the evaluations reported by the CASP3 assessors (Murzin, 1999; Marchler-Bauer and Bryant, 1999). We conclude that MaxSub is a good method for the fully automated evaluation of models, and thus is a suitable measure to be used in the upcoming CAFASP2 event, to be held in 2000.

We would like to emphasize that we do not claim that MaxSub is the best evaluation method; to make such a claim, a systematic comparison of the dozens of previously developed individual measures is needed. This is part of ongoing research. In this paper we only show that the application of one particular single measure (i.e. MaxSub) results in an evaluation comparable to that obtained by the human-expert CASP3 assessors. Future comparisons may demonstrate that other measures perform better. Before such thorough comparisons are carried out, we feel that it is important to make our findings available to other methods' developers, especially because MaxSub may be used as the evaluation method in the upcoming CAFASP2 event. This will allow predictors to know in advance how their predictions will be assessed. In addition, because this paper is the first step towards a systematic comparison of evaluation measures, it may serve as a benchmark for future comparisons.

System and methods

Here we assume that a model is given by a set of 3D coordinates corresponding to the C_α atoms of the predicted residues and ignore the details of the other atoms. To assess the accuracy of the model we aim to find the largest subset of model residues that superimpose well upon their corresponding residues in the experimental structure. Finding the largest subset that satisfies a given condition is apparently a computationally difficult problem, because enumeration is not feasible. However, as we show below, with the aid of heuristics, this problem can approximately be solved in $O(n^2)$ time, where n is the number of residues in the model.

Definitions

Given are two ordered sets of points in 3D, $A = \{a_1, a_2, \dots, a_n\}$ and $B = \{b_1, b_2, \dots, b_n\}$. A corresponds to the experimental structure (or a subset of it; see below) and B corresponds to the model, obtained by homology modeling, fold-recognition, *ab initio* or any other method. For each i , a_i and b_i refer to the same residue; a_i and b_i represent the 3D coordinates of residue i in the experimental structure and in the model, respectively. A match is an ordered set $M = \{(a_i, b_i) \mid a_i \in A, b_i \in B\}$, $|M| \leq n$. A match M defines an optimal transformation T

(rotation and translation) that best superimposes the points of B over A , that is, T minimizes

$$RMS(M) = \sqrt{\frac{\sum_{(a_i, b_i) \in M} \|a_i - T(b_i)\|^2}{|M|}}, \quad (1)$$

where $\|\cdot\|^2$ is the cartesian distance.

Problem statement

Given two sets of points in 3D, A and B , we want to find the largest subset M such that for all $(a_i, b_i) \in M$, $\|a_i - T(b_i)\|^2$ is below some distance threshold value d . Notice that T here is the transformation that minimizes $RMS(M)$, and that M can be composed of non-contiguous fragments. Because finding the largest subset M that satisfies the above condition is computationally expensive (NP-complete), we apply a heuristic, approximate, algorithm. The algorithm given below requires as input the value of the distance threshold d . In the Results section we demonstrate that MaxSub's dependency on the particular value of d is slight, and that the current choice of $d = 3.5 \text{ \AA}$ appears to be as good as other similar choices.

The algorithm

The heuristic algorithm for finding the largest 'well-predicted' subset. The assumption that allows us to tackle this problem efficiently is that the largest subset M sought contains at least one matching segment of length $L \geq 4$ of consecutive pairs from M : i.e. $(a_i, b_i), (a_{i+1}, b_{i+1}) \dots (a_{i+L-1}, b_{i+L-1}) \in M$, for some i . Although this is a strong assumption in general, it does not restrict the applicability of our problem when dealing with polypeptide chains, because the protein models we are interested in either contain L consecutive matching residues for small values of L , or are completely wrong.

The heuristic algorithm used here is similar to that used by (Hubbard, 1999; Zemla *et al.*, 1999) and its basic idea is to generate $O(n)$ different initial 'seed' matches of size L (determined by each consecutive matching segment of length L). Each such 'seed' match is subsequently extended to include additional pairs. At the end, the extended match (M_{\max}) of largest size (s_{\max}) is returned.

Input: The model B and the experimental structure A , and the constants L (the length of the 'seed' segment) and d (the maximum distance between residue pairs that is allowed after superposition).

Output: The largest subset M_{\max} found.

1. $s_{\max} = 0$; /* s_{\max} holds the size of the largest subset found so far */

2. **for** $i = 1$ **to** $n - L + 1$

Let $M = \{(a_i, b_i), (a_{i+1}, b_{i+1}), \dots, (a_{i+L-1}, b_{i+L-1})\}$

$M = \text{Extend}(M, A, B, d)$ /* Increases the size of the current match */

If $|M| > s_{\max}$ **then** $\{s_{\max} = |M|; M_{\max} = M\}$

3. **return** M_{\max}

Extend(M, A, B, d) /* Extends the current subset iteratively */

1. **for** $j = 1$ **to** k /*Extends M in $k = 4$ iterations.*/
 - 1.1 Compute the transformation T that optimally superimposes the residues in M .
 - 1.2 $N = \emptyset$
 - 1.3 **for** $i = 1$ **to** n **do**
 - i. If the distance between a_i and $T(b_i)$ is below the threshold $\frac{j \times d}{k}$, **then** $N = N \cup \{(a_i, b_i)\}$
 - 1.4 $M = N$
2. Using the last M , re-compute the transformation T that optimally superimposes B onto A . If for some $(a_i, b_i) \in M$, the distance of a_i to $T(b_i)$ is above d , then remove (a_i, b_i) from M .
3. **return** M

The algorithm considers $n - L + 1$ segments (n is the number of residues in the model). M is extended over $k = 4$ iterations. At each iteration j , pairs at distance below $\frac{j \times d}{k}$ are added to M . Steps 1 and 3 of **Extend** each require $O(n)$ steps; the superpositions are carried out using the quaternion method (Besl and McKay, 1992); each superposition requires $O(n)$ steps. Thus, the total running time is $O(n^2)$; typical running times are less than one second. In the **Implementation** section we attempt to evaluate how accurate this heuristic algorithm is.

The normalized score

From the resulting M_{\max} we aim to produce a single normalized score, S . S allows us to differentiate the more accurate models with the same (or similar) sizes of M_{\max} ; S awards more points to models whose largest 'well-predicted' subset superimposes better on the experimental structure than models that have a subset of a similar size, but that superimpose at a larger distance. S is a variation of Levitt-Gerstein's score (Levitt and Gerstein, 1999), which is an alternative way, other than the RMSD, to measure structural similarity. S is computed as:

$$S = \frac{\sum \frac{1}{1 + \left(\frac{d_i}{d}\right)^2}}{q}$$

where the sum is over each residue pair i , in M_{\max} , q is the number of C_{α} atoms in the experimental structure, d_i is the distance of the corresponding i th C_{α} atoms (after the optimal superposition of the residue pairs in M_{\max}), and d is the distance threshold. Notice that $q \geq n$ because the experimental structure may contain a larger number of residues than the model. Notice also that residue pairs in M_{\max} contribute to S between one half, if they are at distance d , and 1, if they lie at distance 0. Therefore, in a perfect model, d_i would be 0 for every residue pair, q would equal n , and thus S would be equal to 1. In a completely wrong model the size of M_{\max} is zero, and thus $S = 0$.

One disadvantage of the above S score, is that it does not consider the fragmentation of the models. However, this disadvantage is only minor, because the amount of fragmentation becomes an important factor only for the less accurate models. Devising an alternative S score that penalizes the number of gaps and chain breaks (Levitt and Gerstein, 1999) is part of our ongoing research.

Implementation

MaxSub has been implemented as a web interface accessible at:

<http://www.cs.bgu.ac.il/~dfischer/MaxSub/MaxSub.html>.

The user inputs a model and an experimental structure and MaxSub returns the largest ‘well-predicted’ subset along with its corresponding S score.

Evaluating MaxSub’s heuristic algorithm. To test whether our heuristic algorithm is able to find the largest (or close to the largest) ‘well-predicted’ subset of residues we have carried out an extensive search on one CASP3 model, T0063AL066.1_1. Of the predicted 46 residues, MaxSub identifies a ‘well-predicted’ subset of size 38[†]. To verify whether this subset is the (or close to the) largest ‘well-predicted’ subset, we need to carry out an exhaustive search on all subsets of size greater or equal to 38. Because this is unfeasible computationally, we randomly chose 70,101 subsets of residues in the size range of 38 to 46. Using each such subset, we superimposed the model over the experimental structure, and counted the number of pairs of residues in this superposition that lie within a distance less than or equal to d . The histogram in Figure 1 shows the distribution of the number of residue pairs found within a distance less than or equal to d in each of the 70,101 subsets. 0.2% out of the 70,101 subsets had only 29 pairs within d ; 3.0% of the subsets had 30 pairs, and 0.9% of the subsets had 38 pairs, the same number of pairs found by MaxSub (shown in Figure 1 with an arrow). The largest number of pairs found was 39; this

[†] For the purposes of testing our heuristic algorithm, the particular value of d is not important and an arbitrary value of $d = 4 \text{ \AA}$ was used here; all other computations in this paper use the default $d = 3.5 \text{ \AA}$.

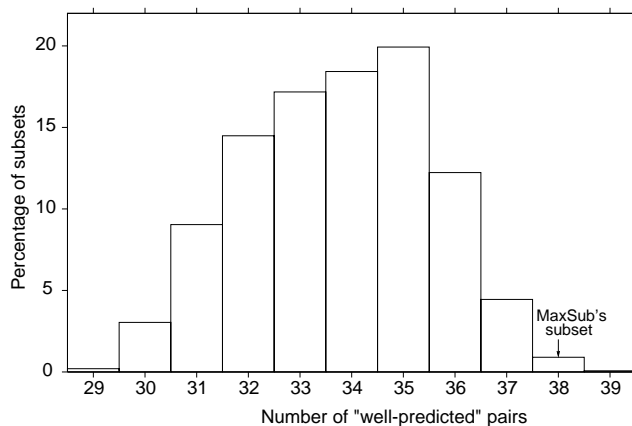


Fig. 1. MaxSub’s heuristics performs well. The histogram shows the distribution of the number of residue pairs found within a distance less than or equal to d in each of 70,101 randomly generated subsets. MaxSub identifies 38 superimposable pairs in the model, one less than the maximum number of superimposable pairs found in the random subsets. Only 0.06% of the random subsets contained 39 superimposable pairs.

occurred in only 0.06% of the subsets. This means that the difference between the largest subset that was found among 70,101 random subsets and the subset that was found by MaxSub is only one residue pair. Thus, it is possible that further improvements in the heuristics may lead to a better approximation. Nevertheless, from this experiment, we feel safe to conclude that although MaxSub is a heuristic, approximate algorithm, it is able to find, if not the largest, a close to the largest ‘well-predicted’ subset in a model.

Results

To illustrate MaxSub, we show the detailed assessment of one prediction. Subsequently, we present a full evaluation of all models in the fold-recognition section of CASP3.

Detailed results for a prediction of target T0063

The CASP3 target T0063 is the translation initiation factor 5A from *P. aerophilum*. The experimental structure consists of two domains, but here we consider only the C-terminal domain (residues 70–138), containing $q = 69$ residues (see also Table 1 footnote). The model evaluated here is the T0063TS009_1 model submitted to CASP3, which includes all the 69 residues of the C-terminal domain ($n = 69$). A structural superposition of the full 3D model (69 predicted residues) with the experimental structure gives an RMS of 10.7 Å. In this superposition only 2 residues superimpose at a distance less than $d = 3.5 \text{ \AA}$; furthermore, only 14 residues are found at a distance less than 6 Å. However, MaxSub

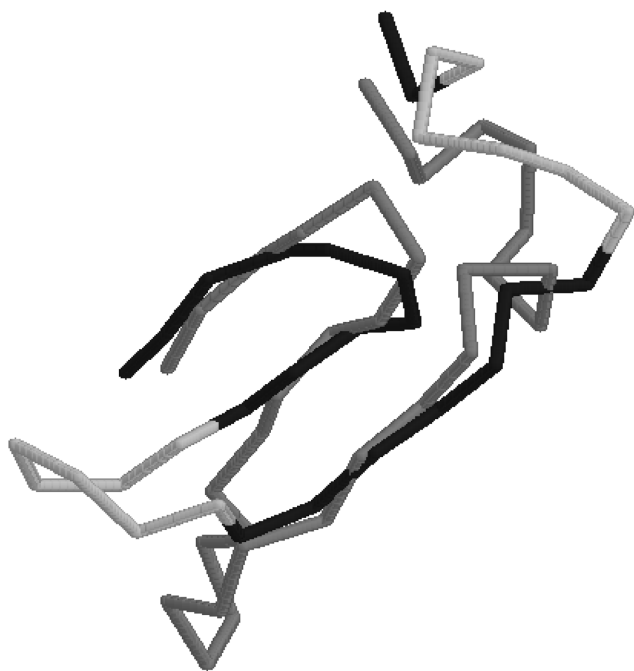


Fig. 2. Superposition of residues 84–108 of the predicted model T0063TS009_1 (light gray) with the experimental structure of T0063 (dark gray), the translation initiation factor 5A from *P. aerophilum*. The rotation and translation used to produce the figure were obtained by optimally superimposing the well-predicted subset identified by MaxSub (black). This subset includes three β -strands corresponding to residues 84–88, 90–93 and 101–107, plus residues 89, 114 and 115. Each of the well-predicted residues superimposes with the experimental structure within $d = 3.5$ Å. The loop at residues 94–100 is wrongly predicted. The termini were also wrongly predicted and are excluded from the figure for clarity. The superposition of the well-predicted subset identified by MaxSub (19 residues) has an RMSD of 2.2 Å, whereas a superposition of the full model (69 residues) gives an RMSD of 10.7 Å.

identifies a subset M of 19 well-predicted residues that include the three β -strands (residues 84–88, 90–93 and 101–107) of one of the protein's β -sheets, plus residues 89, 114 and 115. This means that each of these 19 residues superimposes with the experimental structure within $d = 3.5$ Å (see Figure 2), with an RMSD of 2.2 Å. Four fragments were considered by MaxSub to be wrongly predicted: the two termini (residues 70–83 and 116–138), one loop (residues 94–100), plus residues 108–113 (see Figure 2).

The normalized score S obtained by MaxSub for this model is 0.20, a score indicating that this prediction is only correct in part (see below). Nevertheless, of all models submitted at CASP3 for this target, only four other models had larger S scores.

CASP3 targets

We applied MaxSub to the fold-recognition models submitted at CASP3. Only models labeled as number one were considered here[‡]. Table 1 shows for each predicting group and for each target the score computed by MaxSub using the submitted models. Only the 20 top groups are listed and the scores shown are multiplied by 10 and rounded to the nearest integer. Thus the highest possible score for each prediction is 10 (for a perfect prediction) and the lowest score is 0 (zero is also given when a group did not file a prediction).

The MaxSub numerical score enables us the addition of scores over all models of each group, to get a better overall evaluation of the performance of the predicting groups. Table 1 lists the groups in decreasing order according to the groups' total score (accumulated over all targets, column TOTAL; the totals were computed before the rounding and thus are shown as real numbers). The table shows that the top groups are roughly clustered in four different score ranges. The top five groups correspond to the total score range >7.8 ; the four next groups fall into the score range 5.6–6.1, and the groups ranked 10 to 14 fall into the score range of 3.1–4.1. The last six groups shown have scores <2.4 . Because within each score range, the score variations among the groups vary slightly, it is best to consider the score ranges rather than a straightforward ranking of the individual groups. A ranking based on the small score differences would not be very meaningful (see below).

A direct and exact comparison of MaxSub's assessment vis-à-vis the human-expert CASP3 assessment (Murzin, 1999; Marchler-Bauer and Bryant, 1999) is out of the scope of this paper. Such comparison would in any case be impossible because the criteria and the targets used are different (see Table 1 footnote). One of the main differences is that in the CASP3 evaluation, the only models that received grades were those built on a parent structure considered by the CASP3 assessor to be of the 'correct' fold-type, whereas here we apply our evaluation to all models, regardless of how they were produced. Another minor difference is that for some cases the CASP3 assessor used also the models labeled as number two, whereas here, we use only those labeled as number one. However, because it seems that such a comparison is inevitable, a rough, qualitative, comparison nevertheless sheds some light on the performance of MaxSub as an automatic assessment measure.

In general, we observed that the better models were scored high by both MaxSub and the CASP3 assessors. For example, the first column shows the MaxSub scores obtained for target T0081. Two predictions, of groups 166

[‡] In CASP3 groups were allowed to submit more than one model for each target; these models were labeled by the predictors, so that the best model was labeled as number one.

Table 1. MaxSub results for CASP3 fold-recognition models

Group number	Target number														MaxSub total
	81	44	85	83	54	53	63	79	46	71	67	43	61		
Score range ≥ 7.8															
005	2	0	0	3	0	2	3	0	0	0	0	1	—	—	11.8
212	4	—	—	—	2	2	0	0	2	0	—	—	—	—	10.9
166	5	—	0	—	3	0	—	2	0	0	0	0	—	—	10.1
217	3	1	—	—	2	1	—	2	0	0	0	0	0	0	9.0
176	1	0	0	2	—	2	0	2	0	—	—	0	0	0	7.8
Score range 5.6–6.1															
019	2	1	0	3	—	0	0	0	0	0	0	—	0	—	6.1
061	1	0	0	0	0	0	0	2	2	—	0	0	0	—	6.1
017	—	1	0	3	—	0	—	2	0	—	0	—	—	—	6.0
028	0	1	0	3	0	0	0	2	0	—	0	—	0	—	5.6
Score range 3.1–4.1															
066	—	—	0	—	0	—	4	0	0	0	0	0	—	—	4.1
074	0	0	—	—	2	0	0	—	2	—	0	0	—	—	3.9
185	—	—	—	—	—	—	—	0	—	—	—	—	—	4	3.6
009	1	0	0	0	0	2	0	0	0	0	0	0	0	0	3.5
090	1	—	—	—	—	0	0	2	0	0	—	0	0	0	3.1
Score range ≤ 2.4 (Only first six groups shown)															
076	0	0	0	—	0	0	—	2	0	—	—	0	0	—	2.4
190	—	—	—	2	—	—	—	0	—	—	—	—	0	—	2.3
003	—	—	—	—	—	0	—	—	2	0	—	—	0	—	2.2
201	2	0	—	0	—	0	0	0	0	—	—	0	0	—	1.8
035	0	—	—	0	0	—	0	2	0	0	—	0	0	—	1.7
163	—	—	—	—	—	—	0	2	—	—	—	—	—	—	1.6

Group and target numbers correspond to those used at CASP3 (see <http://Predictioncenter.Inl.gov/casp3>). Only the models labeled as prediction number one were considered here. Scores were multiplied by 10 and rounded to the nearest integer. A ‘—’ indicates that no prediction was filed. To avoid accumulation of low scores, we assigned zeroes to (a) models with an S score ≤ 0.10 , (b) models for which MaxSub identified less than 25 superimposable pairs, and (c) models with scores between 0.10 and 0.125 but with less than 40 superimposable pairs. Targets 45, 51, 72, 77, and 78 were unavailable at the time of preparation of this manuscript, and therefore they could not be evaluated; however, this probably makes only a small difference because for targets 45, 72, 77 and 78, no model received any points by the CASP3 assessors, and target 51 was assessed by only one of the CASP3 assessors. Targets 52, 56, 59, 75 and 80 are not shown, because no group shown in the table received any credit by MaxSub or by the CASP3 assessors (for a complete table see MaxSub’s url). Because MaxSub is independent of structural classification schemes, and because it evaluates fold-recognition, homology modeling and *ab initio* models in the same way, we don’t need to determine whether a target corresponds to a ‘new’ fold or to a ‘known’ one; all targets were evaluated, including those identified by the CASP3 assessors as ‘new’ folds (target numbers 45, 52 and 56). Automatic evaluation using MaxSub for the homology modeling and the *ab initio* CASP3 models will be presented elsewhere.

For the illustration purposes of this paper, we opted to consider all multidomain targets as single prediction targets, with the exception of targets 63 and 71. Although other targets may be better evaluated at the domain level (e.g. target 79), we only considered the domains of targets 63 and 71 in order to be consistent with the CASP3 assessors; the domain considered for target 63 corresponds to residues 70–138 and that of target 71, to residues 1–121. In a real experiment like the upcoming CAFASP2, it may be more appropriate to consider each domain of all multi-domain targets as an individual prediction target; this would result in higher scores for some of the models. However, identifying the domain boundaries is one aspect that requires considerable human intervention, and is out of the scope of this paper.

and 212, received the highest scores (5 and 4). These two groups also received the highest scores by both the CASP3 assessors, ‘A’ and ‘B’, respectively (in CASP3 a ‘formula-1’ (Marchler-Bauer and Bryant, 1999) type of scoring was used, with grades ‘A’ to ‘F’). The next scoring group found by MaxSub is group 217 with a score of 3. Group 217 is also third in both the CASP3 assessors’ ranking (‘C’). Similarly, the two MaxSub highest scoring models for target T0063, from groups 005 and 066, received scores of 3 and 4, respectively. These same models were also identified by the CASP3 assessors as

the top two predictions for this target (with grades ‘A’ and ‘B’). Similar agreements were observed among the better models of the other targets.

However, as can be expected, some differences appeared in the medium to poor models. Out of the 50 non-zero MaxSub scores, 33 received grades by both the CASP3 assessors, and 13 received either a ‘plus’ or a ‘dot’ by one of the assessors (‘plusses’ and ‘dots’ were given by the CASP3 assessor to partially correct models). Conversely, a large number of the models with the lower CASP3 grades (‘D’ to ‘F’) had MaxSub scores below 1. These

largely correspond to models with significant alignment errors that contain only a small subset of superimposable residues, but that received grades by the CASP3 assessors because their parent structures had the ‘correct fold-type’ (see below). For a detailed list of the CASP3 assessors’ ranking we refer the reader to the published tables (Murzin, 1999; Marchler-Bauer and Bryant, 1999).

Overall, the top six groups according to the CASP3 assessors are ranked as the top six groups also by MaxSub. Furthermore, MaxSub ranks 1 to 10 include the ranks 1 to 13 of both the CASP3 assessors. Some of the differences between MaxSub’s rankings and the CASP3 assessors’ can partially be attributed to the different methodologies used. For example, the CASP3 assessors did not award any credit to predictions built on parent structures of different fold-types than that of the experimental structure, even if a superimposable subset of atoms having a significant size can be found, and conversely, as noted above, grades were given in CASP3 to models based on a ‘correct’ parent-structure even if the alignment was majorly incorrect. Because 3D structures do not evolve in nature to accommodate discrete classifications, two proteins sharing partial structural similarity can fall into two different fold-types in a structural classification scheme. Thus, it is not uncommon for fold-recognition methods to identify a parent structure that is partially similar to the experimental structure, but has a different fold-type than the one assigned to the experimental structure by a given structural classification method. Whether such predictions need to be awarded some credit or not, is an open question, and the answer is probably dependent on what the biologist who is attempting to model a protein is looking for. In any case, MaxSub’s evaluation is based exclusively on the 3D coordinates of the models and is independent of any fold classification scheme. Thus, *ab initio* models can be evaluated in the exact same way as homology modeling or fold-recognition models.

The agreement between MaxSub and the CASP3 assessors is not as good for MaxSub’s ranks 11 to 20; some groups are ranked similarly by MaxSub and by the CASP3 assessors, but for other groups, significant differences are observed. However, it is important to notice that the number of correct predictions of the groups in these ranks is very small, and thus, the differences may not be very meaningful. Indeed, for a number of the groups, discrepancies also occur between the two CASP3 assessors.

A more straightforward comparison of MaxSub with the CASP3 assessment is to take into account only those models considered by the CASP3 assessors (i.e. those models built on parent structures considered by the CASP3 assessors to be of the ‘correct’ fold-type). To this end, only the MaxSub scores of models also considered by the CASP3 assessors are used. Not surprisingly, such a comparison (results not shown), shows an even better

agreement between the three methods in the ranking of the top 15 groups.

We conclude that overall, there is good agreement between the three methods in ranking the best eight or nine groups. A complete listing of all the remaining groups is available at MaxSub’s url.

MaxSub’s dependency on the threshold d

The advantage of having a single measure to assess the models was one of the main motivations to develop MaxSub. However, to achieve this, MaxSub sets the distance threshold at which a pair of corresponding residues is considered to be ‘well predicted’ (the d parameter; see **Methods**) to a value of 3.5 Å. When it is not clear what this threshold should be, or when one does not wish to use any arbitrary setting, several thresholds need to be selected (see Hubbard, 1999), with the disadvantage that this may impede automatic assessment. The approach used by GDT (Zemla *et al.*, 1999) is to average the values obtained at various thresholds. Here, we prefer to use a fixed value of d , set at the vicinity of the distance at which most researches would consider a residue to be ‘well predicted’.

To test the sensitivity of MaxSub to the d threshold, we have tried values of d in the range 2–7 Å, and compared the overall ranking of the groups to the ranking shown in Table 1 (computed with $d = 3.5$ Å). The comparison demonstrates that most groups remain within the same score-ranges shown in Table 1 (results not shown); the five top groups found at $d = 3.5$ Å remain within the top six positions, regardless of the value of d used. This is expected because as the score differences among the groups within the same score range are slight, some parameter variations are likely to produce small variations in the ranking of the groups. We also noticed that the rank variation observed among the different thresholds is similar to the rank variation observed among the two CASP3 assessors (not shown).

In summary, the ranking of the groups using different values for the d threshold varies slightly, especially among those groups with similar performance. We conclude that MaxSub’s ranking is robust and relatively independent of the particular value chosen. We also conclude that for the evaluation of fold-recognition models, setting the value of $d = 3.5$ Å is as good a choice as other close values. To verify that MaxSub is also robust on homology modeling predictions, we have applied it to the CASP3 homology modeling models, and we have compared the results with those reported by the CASP3 homology modeling assessor (Jones and Kleywegt, 1999). The results show a high degree of agreement between the two assessments, providing further validation of the applicability of MaxSub. A table showing MaxSub’s scores and rankings of the homology modeling groups is available at MaxSub’s url.

Discussion and conclusion

As protein modeling becomes a widely-used tool, the importance of evaluating the produced models becomes greater. This is particularly true for events like CASP and CAFASP, where a large number of models has to be assessed on a fairly simple, objective, quantitative and reproducible scale.

To date a number of evaluation methods and measures have been devised, and no single method producing a single measure has been agreed upon. In the CASP experiments, human expertise has played a large part both in the prediction and in the evaluation processes. In the latter, human intervention has been applied to interpret the large set of measures generated for each model. Here we have presented a fully automated measure called MaxSub that is a further development of some of the principles developed for the assessment of the CASP3 models. MaxSub searches for the maximum superimposable subset of residues in a given model and an experimental structure and produces a normalized score in the range of 0 to 1 that represents the quality of the model. MaxSub is a simple and intuitive measure that allows for the objective, quantitative, reproducible, and fully automated comparison of various models. MaxSub considers all models as a set of 3D coordinates and it does not require a discrete fold classification scheme to generate a list of 'correct' structural matches. MaxSub does not distinguish between models produced by fold recognition, *ab initio*, or homology modeling methods and evaluates all models in the exact same way.

Although a direct comparison of MaxSub's results and the CASP3 results is not possible (partly because of the different approaches used in the evaluation, and partly because not all the experimental structures evaluated by the CASP3 assessors were available to us) we still felt that the CASP3 targets are an excellent test set for evaluating the performance of MaxSub. We found that there is good agreement between MaxSub's evaluation and the human-based evaluation of the better models submitted to CASP3. When comparing the overall ranking of the groups we also observed that there is good agreement in the selection of the top performing groups. Nevertheless, as expected, the agreement was not as good for the medium to poor models and groups.

It may be impossible to obtain a single method to assess predicted models that will be agreeable to all researchers in the field, because any one single method is likely to suffer from one limitation to another. MaxSub is not an exception. In this paper we do not claim that MaxSub is the best method, we only show that this automated method, which produces a single normalized score, is able to reproduce human-expert evaluations reasonably well. To determine what single measure is the best one, large-scale, systematic comparisons need to be carried out. Here

we have presented the first step towards such a large-scale comparison, and we are looking forward to seeing further comparisons of existing and new measures.

Advocates of the sequence-independent type of assessment, may raise conceptual objections to MaxSub. Because MaxSub's approach is sequence-dependent, a predicted model based on a 'correct' fold but with a wrong alignment may not receive a high score by MaxSub, while a sequence-independent assessment may award some credit to it. However, it is evident that a good model must be based on a good alignment, and such a model will be scored highly by MaxSub. Nevertheless, if one seeks for a sequence-independent assessment, then approaches like those proposed by Bryant and colleagues or by Sippl and colleagues may be required.

MaxSub's approach is sequence-dependent as are the RMS/coverage graphs and GDT. MaxSub differs from previous methods in at least one of the following: (a) MaxSub produces a single number that measures the quality of a model, (b) it does not require visualization or human intervention, and (c) the number produced by MaxSub is a normalized score that considers the size of the well-predicted subset and, at the same time, its closeness to the experimental structure.

It is obvious that expert human intervention is likely to add evaluation aspects that are hard or impossible to automate. When the number of models to be evaluated is not too large, and human-expert assessment is available, then MaxSub can at least be used as a companion tool to ease the burden of the human assessment. Nevertheless, when the number of models is large and no human-expert is available, an automated tool is required. We believe that MaxSub's advantages demonstrated above, outweigh its possible limitations. We conclude that an approach like MaxSub can serve as a basis for the automated evaluation of models for large-scale experiments like CAFASP. With an automated method like MaxSub, the CAFASP2 experiment will be free of human intervention in both the prediction and the evaluation processes. This will be a step forward in our efforts towards a higher level of automation in the field.

The results of additional evaluations of CASP3 models using MaxSub as well as an automatic server that implements MaxSub are available through the internet at: <http://www.cs.bgu.ac.il/~dfischer/MaxSub/MaxSub.html>.

Acknowledgements

Thanks to Erez Karpas for his assistance in setting the MaxSub server. We are grateful to Adam Zemla and Krzysztof Fidelis from the Livermore Prediction Center for critically reading this manuscript, for providing the data via the Prediction Center url, and for discussions. We acknowledge Adam Zemla for developing GDT. We thank Manfred Sippl for critically reading the manuscript and

for helpful suggestions. Thanks also to Mark Gerstein, Klara Kedem, Adam Godzik and John Moult for fruitful discussions. Special thanks to John Moult and to the organizers and assessors of CASP3 for their important contribution to the field. Finally, we are also grateful to other CASP and CAFASP organizers and participants for encouragement. N.S. is partially supported by the Israeli High-Tech and by the Kreitman Foundation Fellowships.

References

- Besl,P.J. and McKay,N.D. (1992) A method for registration of 3-D shapes. *Trans. Pattern Analysis and Machine Intelligence*, **14**(2), 239–256.
- Fischer,D., Barret,C., Bryson,K., Elofsson,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MacCallum,R.M., Pawlowski,K., Rost,B., Rychlewski,L. and Sternberg,M. (1999) CAFASP-1:critical assessment of fully automated structure prediction-methods. *Proteins*, Suppl 3, 209–217.
- Hubbard,T.J.P. (1999) RMS/Coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. *Proteins*, Suppl 3, 15–21.
- Jones,A.W. and Kleywegt,G.J. (1999) CASP3 comparative modeling evaluation. *Proteins*, Suppl 3, 30–46.
- Lackner,P., Koppensteiner,W.A., Domingues,F.S. and Sippl,M.J. (1999) Automated large scale evaluation of protein structure predictions. *Proteins*, Suppl 3, 7–14.
- Levitt,M. and Gerstein,M. (1999) A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA*, **95**, 5913–5920.
- Marchler-Bauer,A. and Bryant,S.H. (1999) Comparison of prediction quality in the three CASPS. *Proteins*, Suppl 3, 218–225.
- Moult,J., Hubbard,T., Bryant,S.H., Fidelis,K., Pedersen,J.T. and Predictors, (1997) Critical assessment of methods of proteins structure prediction (CASP): round II. *Proteins*, Suppl 3, Dedicated Issue.
- Moult,J., Hubbard,T., Fidelis,K. and Pedersen,J. (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins*, Suppl 3, 2–6.
- Murzin,A.G. (1999) Structure classification-based assessment of CASP3 predictions for the fold-recognition targets. *Proteins*, Suppl 3, 88–103.
- Orengo,C.E., Bray,J.E., Hubbard,T., LoConte,L. and Sillitoe,I. (1999) Analysis and assessment of Ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins*, Suppl 3, 149–170.
- Sippl,M.J., Lackner,P., Domingues,F.S. and Koppensteiner,W.A. (1999) An attempt to analyze progress in fold recognition from CASP1 to CASP3. *Proteins*, Suppl 3, 226–230.
- Venclovas,C., Zemla,A., Fidelis,K. and Moult,J. (1999) Some measures of comparative performance in the three CASPs. *Proteins*, Suppl 3, 231–237.
- Zemla,A., Venclovas,C., Moult,J. and Fidelis,K. (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, Suppl 3, 22–29.