PROTEINS: Structure, Function, and Genetics 53:548–560 (2003)

CAFASP3 in the Spotlight of EVA

Volker A. Eyrich,[1*] Dariusz Przybylski,[1,2,4] Ingrid Y.Y. Koh,[1,2] Osvaldo Grana,[5] Florencio Pazos,[6] Alfonso Valencia,[5] and Burkhard Rost[1,2,3*]

[1]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[2]*Columbia University Center for Computational Biology and Bioinformatics (C2B2), New York, New York*
[3]*North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[4]*Department of Physics, Columbia University, New York, New York*
[5]*Protein Design Group, Centro Nacional de Biotecnologia (CNB-CSIC), Cantoblanco, Madrid, Spain*
[6]*ALMA Bioinformatica, Contoblanco, Madrid, Spain*

ABSTRACT **We have analysed fold recognition, secondary structure and contact prediction servers from CAFASP3. This assessment was carried out in the framework of the fully automated, web-based evaluation server EVA. Detailed results are available at http://cubic.bioc.columbia.edu/eva/cafasp3/. We observed that the sequence-unique targets from CAFASP3/CASP5 were not fully representative for evaluating performance. For all three categories, we showed how careless ranking might be misleading. We compared methods from all categories to experts in secondary structure and contact prediction and homology modellers to fold recognisers. While the secondary structure experts clearly outperformed all others, the contact experts appeared to outperform only novel fold methods. Automatic evaluation servers are good at getting statistics right and at using these to discard misleading ranking schemes. We challenge that to let machines rule where they are best might be the best way for the community to enjoy the tremendous benefit of CASP as a unique opportunity for brainstorming. Proteins 2003;53:548–560.**

Key words: protein structure prediction; evaluation; secondary structure; inter-residue distances; contact maps; threading; fold recognition; automatic servers

INTRODUCTION

Continuous, Automated, Large Data Sets, Statistical Significance

The goal of EVA is to evaluate the sustained performance of protein structure prediction servers through a battery of objective measures for prediction accuracy.[1,2] EVA evaluates: (1) comparative modelling, (2) fold recognition and threading, (3) inter-residue contact, and (4) secondary structure predictions. Since May 2000, EVA has collected predictions from public servers in all four categories for over 10,000 different protein chains and has maintained an extensive database of predictions allowing for detailed statistical analysis using various criteria. EVA is a fully automatic assessment procedure and focuses on web-based prediction servers exclusively. Submission of targets to servers as well as data collection and evaluation are fully automated and are capable of scaling up to much larger numbers of prediction targets than CASP: despite the immense increase in the number of targets at CASP5, over 100 times more targets have been handled by EVA between CASP4 and CASP5 than at CASP5 alone. Details on the mechanism of data acquisition and the presentation of results are available from the web and our previous publications;[1,3,2] forthcoming publications will describe the evaluation procedure in more detail.

CASP and EVA

CASP[4-7] is unique as an assessment and evaluation experiment in the sense that the selection of targets is more or less unbiased; it is certainly *not* biased by the developers themselves. Experimentalists submit prediction targets (sequences) of yet unknown structure for assessment. While EVA also avoids the bias, most proteins analysed in the past could - in principle - have been used directly by the prediction servers at the time of submission by EVA. In other words, developers could cheat by first downloading the known structure, then perturbing the answer and returning the 'prediction' to EVA. We see no evidence that any group has attempted to explore this loophole. Furthermore, Phil Bourne (UCSD) and Helen Berman (Rutgers) have added a new detail to the submission of structures to the PDB: when experimentalists submit their structure to PDB, they usually attach a particular blocking date before the co-ordinates are to be made public by the PDB while the sequence is available by default. This will eventually close the loophole. The loophole of CASP is that so few targets are evaluated every two

Grant sponsor: Spanish Ministry of Science and Technology; Grant number: BIO2000-1358-CO2-01; Grant sponsor: TEMBLOR; Grant number: CSIC:LIFE/001/0957 UE:QLRT-2001-00015; Grant sponsor: SANITAS; Grant number: CSIC:LIFE/992/0553 UE:QLK3-CT-2000-00079; Grant sponsor: RAS; Grant number: CSIC:LIFE/991/0219 UE:QLK3-CT-1999-00875; Grant sponsor: National Institutes of Health: Grant numbers: 5-P20-LM7276 and RO1-GM63029-01.

*Correspondence to: Burkhard Rost, CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th St., BB 217, New York, NY 10032. E-mail: rost@columbia.edu

Received 28 February 2003; Accepted 19 May 2003

years that groups linked to the experimentalists will have advantages. Since CASP - due to the small numbers - can only focus on 'peak' and not on sustained performance, being part of a structural genomics consortium could for instance make groups become 'winners'. Again, there is no evidence that any group has explored this loophole of CASP.

While CASP addresses the question: "How well can experts predict protein structure?", CAFASP is meant to address the question "How well can machines predict?". Here we analysed all servers that subscribed to CAFASP on the limited number of CASP targets. For those secondary structure prediction servers that have been evaluated for a longer time by EVA, we also compared this CAFASP snapshot to their sustained performance.

DATA

Domain Versus Chain

CASP parses proteins into structural domains and assesses the performance on each structural domain separately. In contrast, the 'basic unit' of EVA is a protein chain. This deliberate choice in EVA was made because of the following two reasons. (1) Different delineations of protein domains agree for less than half of all the structural domains known today.[8,9] Given that the PDB is currently extremely biased toward single-domain proteins,[10] the number of domains that are not well-defined is likely to rise considerably in the future. (2) As a rule of thumb in sequence analysis: database comparisons and prediction methods are more accurate/sensitive when queried with relevant fragments than with full-length proteins. This implies that the parsing of full-length proteins into domains already solves some of the tasks of prediction methods. Hence, beginning from domains yields an over-optimistic perspective.

Sequence-Unique Versus Similar-to-Known Structure

The CASP categories of comparative modelling, fold recognition/threading, and novel fold overlap to some extent. Obviously, this paves the way for endless, confusing debates on which target belongs to which category. True to its spirit CASP takes the route of letting the expert assessors decide what to evaluate in which category. Since EVA is fully automatic, we have to make a more coarse-grained, objective choice: When evaluating fold recognition/threading, and predictions of contacts and secondary structure, EVA focuses on sequence-unique proteins, i.e. proteins for which structure could not be modelled 'trivially' by comparative modelling methods. Protein chains are considered as sequence-unique if either (1) an iterated PSI-BLAST detects any similarity to a known structure at an E-value $< 10^{-2}$ or (2) a dynamic programming detects a similarity above an HSSP-distance of 0.[11]

25 Sequence-Unique Targets for CASP5

For CASP5/CAFASP3, the above distinction (sequence-unique/not) implied that 25 proteins (T0129, T0130, T0132, T0134, T0135, T0136, T0138, T0139, T0146, T0147, T0148, T0149, T0156, T0157, T0159, T0161, T0162, T0168, T0170, T0173, T0174, T0181, T0187, T0193, T0194) could be used to assess the performance of methods other than comparative modelling. However, both for completeness and comparison, we also analysed the performance on the remaining 32 proteins (T0133, T0137, T0140, T0141, T0142, T0143, T0150, T0151, T0152, T0153, T0154, T0155, T0160, T0165, T0167, T0169, T0172, T0176, T0177, T0178, T0179, T0182, T0183, T0184, T0185, T0186, T0188, T0189, T0190, T0191, T0192, T0195).

Servers

We show results from the following 55 methods: 3D-PSSM,[12] 3D-SHOTGUN-3,[13] 3D-SHOTGUN-5,[13] 3D-SHOTGUN-N,[13] 3D-SHOTGUN-INBGU,[13] APSSP (GPS Raghava, unpublished), APSSP2 (GPS Raghava, unpublished), ARBY(Ingolf Sommer et al., unpublished), BasicB,[14] BasicC,[14] BLAST,[15] Bystroff (Chris Bystroff, unpublished), CMAPpro,[16] CORNET,[17] FAMS, FAMSD (both M Iwadate et al., unpublished), FFAS,[14] FFAS03 (L Jaroszewski et al., unpublished), FORTE1 (K Tomii at al., unpublished), FUGUE2,[18] FUGUE3 (K. Mizuguchi et al., unpublished), GenTHREADER,[19] INBGU,[20] JPred,[21] LIBELLULA,[22] LOOPP,[23] mGenTHREADER,[24] MPALIGN (T Akutsu et al., unpublished), ORFblast (L Rychlewski, unpublished), ORFeus (L Rychlewski, unpublished), Pcomb, Pcons2, Pcons3 (all three: Arne Elofsson, unpublished), PDGcon,[25] PHDsec,[26] Pmodeller, Pmodeller3 (both Arne Elofsson, unpublished), PROFking,[27] PROFphd (B Rost, unpublished), Prospect,[28] PROTINFO-CM, PROTINFO-FR (both Ram Samudrala, unpublished), PSI-BLAST,[29] PSIpred,[30] RAPTOR (J Xu et al., unpublished), ROBETTA (D Chivian et al., unpublished), SAM-T99/SAM-T99sec,[31] SAM-T02/SAM-T02sec (Kevin Karplus, unpublished), SSEARCH,[32] SSpro2,[33] SUPERFAMILY,[34] SUPFAM_PP (Julian Gough, unpublished).

SECONDARY STRUCTURE PREDICTIONS

Servers

Since its inception EVA has evaluated approximately 15 secondary structure prediction servers and has accumulated over 50,000 individual predictions. The servers are: APSSP2, JPred,[21] PHDsec,[26] PHDpsi,[35] PROFking,[27] PROFsec /PROFphd, Prospect,[36,28] PSIpred,[30,24] SAM-T99sec,[31] SAM-T02sec, SSpro[33] and SSpro2.[37] Most methods that contributed to CAFASP[38,39] were also tested over a long period by EVA. Interestingly, many servers participated in CAFASP2 and CAFASP3. APSSP and APSSP2 appear to be updated version of PSSP, and SAM-T02 is the successor to SAM-T99 (both were tested by EVA); ROBETTA and Prospect represent new servers. This number is considerably smaller than the number of new servers in the fold recognition category.

Assessment

Assignment of secondary structure

EVA uses DSSP,[40] DSSPcont,[41] and STRIDE[42] to assign secondary structure from 3D co-ordinates. Here, we report only values for DSSP; we used the following conversion of the eight DSSP states into three classes DSSP(HGI)->H,

DSSP(EB)->E, DSSP(other)->L. For the 'new fold', fold recognition and comparative modelling methods, we assessed only the first model. For models without sidechains, we predicted these through MaxSprout.[43]

Raw scores

The assessment of secondary structure predictions was carried out using the scoring system defined by EVA. The measures include the familiar three-state accuracy (Q_3) and segment overlap scores (SOV),[44] the BAD score,[45] and a battery of other measures established in the field (for details http://cubic.bioc.columbia.edu/eva/). The three-state per-residue accuracy is defined as:

$$Q_3 = \left\langle 100 \cdot \frac{\text{number of residues correctly predicted in [HEL]}}{\text{number of residues in protein}} \right\rangle_{\text{all N proteins}} \quad (1)$$

where $\langle \ldots \rangle$ is the average over all N proteins in the data set. BAD is the percentage of residues observed in helices and predicted in strands, or observed in strand and predicted in helix. The information index *info* x

$$\text{info} = \ln\left\{\frac{P_{prd}}{P_{obs}}\right\} \quad (2)$$

where P_{obs} describes the probability for finding one particular string of N residues observed in class *i* (HEL) out of all combinatorially possible ones, and P_{prd} is the probability for a particular realisation of the prediction matrix {M} the element M_{ij} of which gives the number of residues observed in class i and predicted in class j. The precise definition of this score is explained in more detail elsewhere.[46]

All per-residue scores ignore correlations in the predictions of consecutive residues and hence the 'fact' that regular secondary structure forms segments. The three-state per-segment overlap SOV was defined as:

$$SOV = SOV_3 = \left\langle \sum_{i}^{HEL} \frac{1}{M_i} \cdot \sum_{S(i)} \frac{\text{MINOV(S1;S2)} + \text{DELTA(S1;S2)}}{\text{MAXOV(S1;S2)}} \right\rangle_{\text{all N proteins}} \quad (3)$$

where S1 is the observed and S2 the predicted secondary structure segment in class i (HEL), MINOV(S1;S2) the number of residues that the observed and predicted segments S1 and S2 overlap, and MAXOV(S1;S2) is the total number of residues over which residues from either S1 or S2 extend. Note that MINOV=MAXOV if S1 and S2 are identical, otherwise MINOV<MAXOV. The normalisation is:

$$M_i = \sum_{S(i)} \text{LEN(S1)} + \sum_{S'(i)} \text{LEN(S1)} \quad (3a)$$

where LEN(S1) is the length of segment S1, and S(i) is the number of all the pairs of segments {S1;S2} that have at least one residue in common in class i, and S'(i) is the number of segments S1 that do not overlap with any prediction. Finally, DELTA(S1;S2) is defined by:

$$\text{DELTA(S1;S2)} = \min\begin{cases} \text{MAXOV(S1;S2)} - \text{MINOV(S1;S2)} \\ \text{MINOV(S1;S2)} \\ \text{INT}(0.5 \cdot \text{LEN(S1)}) \\ \text{INT}(0.5 \cdot \text{LEN(S2)}) \end{cases} \quad (3b)$$

In contrast to the per-residue based scores, SOV reflects the observation that similar folds often differ primarily in the precise lengths of helices and strands. The SOV measure has been optimised to distinguish between pairs of proteins with similar folds (SOV->100%) and pairs with different folds (SOV->0%).[47,44]

The per-segment score is still local in the sense that it does not explicitly reflect a feature of the entire protein. In contrast, the difference between the content in predicted and observed regular secondary structure explicitly captures a more global aspect; it is defined by:

$$\text{contDX} = \frac{1}{N} \cdot \sum_{c}^{N} |\text{frac(X)}_c^{obs} - \text{frac(X)}_c^{prd}| \quad (4)$$

where X is helix or strand, N the number of proteins in the data set, and frac(X)_c^{obs}(frac(X)_c^{prd} are the fractions of residues observed (predicted) in class X for protein c.

Ranking

Ranking methods in order to declare 'winners' has been a major source of excitement during and after all five CASP meetings. Obviously, there are many social - as opposed to scientific/technical - reasons for such debates. However, there are also at least three major technical issues that fuel the perpetual, immobilising debate: (1) the number of targets at CASP has always been so small that differences between methods were based on numerical differences that had no statistical significance. (2) Given a plethora of measures for performance, almost every method appears best under one of these scores. (3) Comparing methods based on different data sets, i.e. comparing apples and oranges, can give rise to very misleading conclusions.[48] Therefore, EVA carefully analyses whether or not the rank can be distinguished between two methods. First, methods are never compared based on different data sets. Second, ranks are never differentiated between two methods if the sustained performance of the two is statistically insignificant. Currently, we are working on even more robust ranking schema derived from pairwise comparisons.[49]

RESULTS

Changes From CAFASP2/CASP4

The first observation from the CAFASP3 results (Table 1) appears to be that the field did stunningly better than two years ago. In fact, without information from outside the framework of CASP/CAFASP this optimistic scenario is the only conclusion an expert can walk away with.

TABLE I. EVA Results for All 21 Sequence-Unique Targets at CAFASP3[†]

Rank	Method	Q_3	SOV	BAD	info	contDH	contDE
1	APSSP2	77.3	75.6	2.9	0.41	3.4	5.6
	PHDsec	72.9	70.6	4.1	0.35	7.0	5.5
	PROFking	75.3	74.8	2.1	0.38	6.1	5.2
	PROFsec	77.2	76.3	2.6	0.41	3.1	4.1
	Prospect	73.5	72.1	3.7	0.36	5.6	5.9
	PSIpred	78.6	77.4	2.6	0.44	3.5	5.2
	ROBETTA	75.9	74.3	2.4	0.39	3.7	4.8
	SAM-T02sec	78.9	76.2	2.1	0.43	4.1	4.1
	SAM-T99sec	77.7	76.9	2.5	0.42	6.1	5.0
	SSpro2	76.5	72.9	3.3	0.40	4.8	5.3
2	APSSP	67.5	64.8	6.6	0.25	7.1	7.8

[†]Data Sets/Methods. All methods shown predicted for 21 of the 25 sequence-unique targets. Methods. APSSP: G Raghava, imtech.res.in/raghava/apssp/; APSSP2: G Raghava, imtech.res.in/raghava/apssp2/; PHDsec: Refs. 26 and 35, cubic.bioc.columbia.edu/predictprotein/; PROFking: Ref. 27, www.aber.ac.uk/~phiwww/prof/; PROFsec/PROFphd B Rost, cubic.bioc.columbia.edu/predictprotein/; Prospect: Refs. 28 and 36, PSIpred: Refs. 24 and 30, insulin.brunel.ac.uk/psiform.html; ROBETTA, Baker Lab; SAM-T99sec: Ref. 31, www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html; SAM-T02sec: K Karplus; SSpro2: Ref. 37, promoter.ics.uci.edu/BRNN-PRED/ Scores. Q_3, three-state per-residue accuracy (Eq. 1); SOV, three-state per-segment accuracy (Eq. 3);[44] BAD, percentage of residues confused between helix and strand;[45] info, information content of prediction,[46] err < 0.03; contDH (contDE), difference in content of helix (strand, Eq. 4). Ranking. Methods that do not differ based significantly in Q_3 get the same rank. Within a given rank, methods are sorted alphabetically.

Secondary Structure Prediction Methods Appear Best in Their Domain

Homology modelling and fold recognition methods also generate implicit predictions of secondary structure. While Krzysztof Fidelis and colleagues at the Lawrence Livermore Prediction Center[50] have provided cross-comparisons for these methods since CASP2, these data have not been systematically analysed so far, assuming that when applicable homology modelling and fold recognition methods are more accurate than secondary structure prediction methods. While the small data set (15 sequence-unique proteins in common) indicated that specialist methods tended to outperform non-specialists (Fig. 1), the best methods were less than two standard deviations off the average over all methods. Considering that most non-specialist models were based on regions of the most reliable prediction, the advantage of specialists became more striking (Fig. 2): for instance, the fold recognition META-servers maximally modelled 74% of the target residues (Fig. 1, Pmodeller), at levels around 72% accuracy. At the same level of coverage, the prediction accuracy for the best specialists exceeded 84%, i.e. more than a standard deviation higher, and over two standard deviations above the average over all methods. While the per-segment score SOV (Eqn. 3) correlated well with the per-residue score Q_3 (Eqn. 1; correlation > 0.95), most methods tended to perform worse when evaluated based on segments (Fig. 1, upper panel).

Methods Better on Comparative Modelling Targets

Twenty-two chains of the CASP5/CAFASP3 targets were considered as 'sequence-similar to known structures'. Not surprisingly, all methods performed on average better for these than for the sequence-unique targets (data not shown). While 22 are still too few for sustained conclusions, EVA confirmed that prediction methods are more accurate for proteins similar to known structures on a much larger data set (http://cubic.bioc.columbia.edu/eva/sec_homo/).

Sustained Performance of Methods

For six of 11 methods in CAFASP3 (PHD, PROFking, PROFsec, PSIpred, SAM-T99sec, SSpro2), EVA has been collecting significant amounts of data in the past (Table 2). These data suggest three main conclusions. Firstly, the 21 common CASP/CAFASP proteins were not representative. Instead, all methods performed better on these proteins than they did on average for more significant data sets. Thus, developers at CASP5/CAFASP3 appeared lucky to have unusually 'simple' targets. Secondly, if we ranked methods by their numerical values for the 21 targets, we should make serious mistakes! For example, PROFking and PHDpsi differ 2.5 percentage points in their Q_3 (Table 1), however, the larger set shows that there is no sustained difference between the two (Table 2). Conversely, the larger data set splits the field of best methods. Thirdly, on a set of 567 sequence-unique proteins (data not shown), PROFsec and PSIpred both reach a sustained Q_3 of 75.6%; this is significantly better than the best method of the first CASP meetings (PHDsec).

Many Ranking Schemes From CASP/CAFASP Would Yield Misleading Results

Five particular ways of ranking are popular in CASP/CAFASP: (1-2) number of scores/proteins for which method M is above average, (3-4) number of scores/proteins for which M is best, (5) average rank over target set. On the CASP/CAFASP data set, these techniques would make the following mistakes (compare to Table 2): (1) PSIpred on par with PHDpsi, (2) PHD better than SAM-T99, (3)

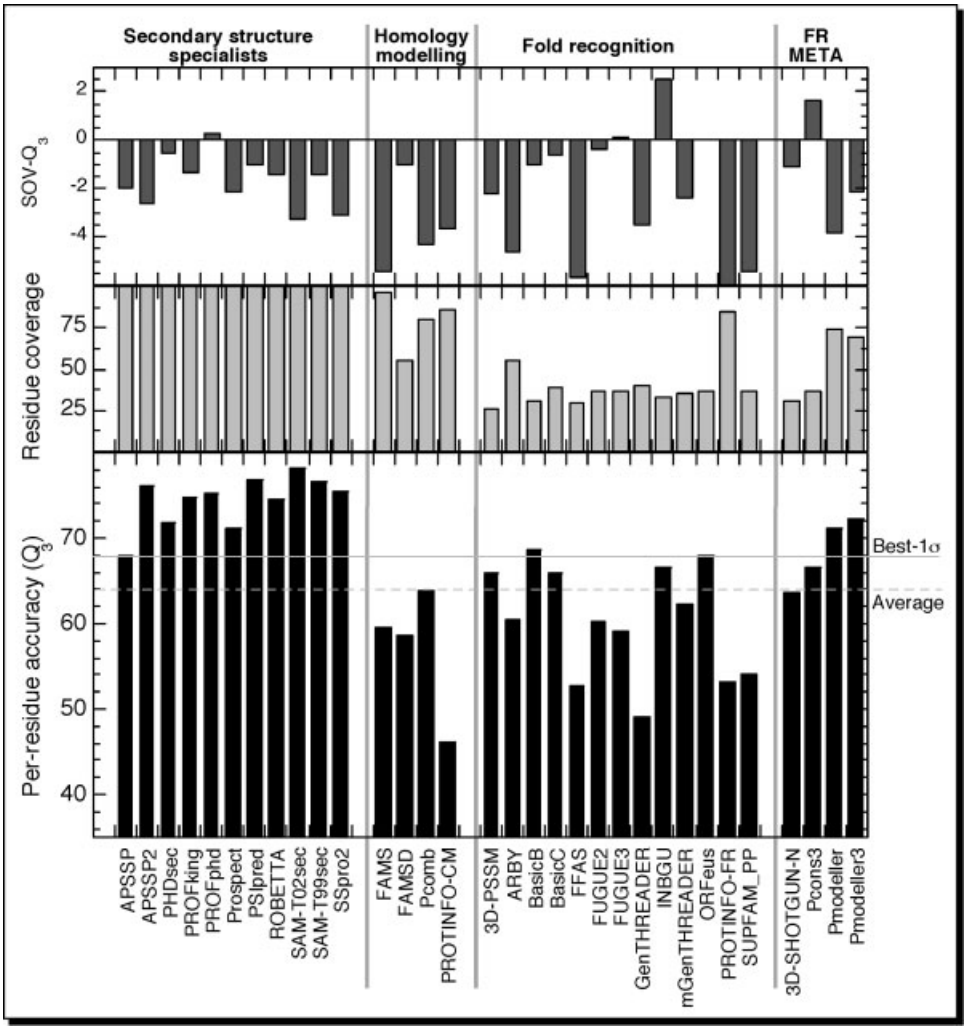

Fig. 1. Secondary structure prediction for all categories at CAFASP3. The data includes only 15 sequence-unique targets (Data, T0130, 32, 34, 36, 46-49, 57, 61, 68, 73-74, 81, 87) common to all methods shown. The average three-state per-residue accuracy (Eqn. 1) over all methods was about 65%, the best methods reached 78% accuracy with one standard deviation of 10% (note the base line marks a random prediction). Thus, the best specialist performed slightly better than average, and slightly better than most non-specialists. Most non-secondary structure prediction methods covered only partial regions (percentage coverage in middle panel, compare Fig. 2). The top panel illustrates that most methods performed better on a per-residue than on a per-segment base, the exceptions were (sorted by coverage in brackets): PROFphd (100%), FUGUE3 (36%), Pcons3 (36%), and INBGU (33%). Homology modelling methods are not designed for the targets tested here; it is then surprising that they appear similar to the fold recognition field (slightly lower accuracy but higher coverage).

SAM-T99sec second worst, (4) PHD better than SSpro2, (5) SSpro2 worse than PROFking.

INTER-RESIDUE CONTACTS/DISTANCES

Servers

We analysed the capacity for predicting inter-residue contacts for both specialised methods and all other models generated in the context of CAFASP. For the 'new fold', fold recognition and comparative modelling methods, we assessed only the first model. We considered two residues to be in contact when their C-beta atoms (C-alpha for Gly) were $\leq$ 8Å. For models without side-chains, we predicted these through MaxSprout.[43] In the case of the contact prediction servers, the list of predicted contacts pairs was sorted by their associated probability scores and separated evaluations were done for a number of pairs corresponding to: L/10, L/2, L/5, L, 2L, and 5L, where L is the protein length.

Assessment

Contacts

Contact predictions of models were evaluated in terms of (1) accuracy (Acc) and (2) improvement over random (Imp). Accuracy was defined as:

$$\mathrm{Acc} = \frac{\text{number of contacts correctly predicted}}{\text{all contacts predicted}} \quad (5)$$

The improvement over random is simply the quotient between the accuracy of the prediction and the accuracy a

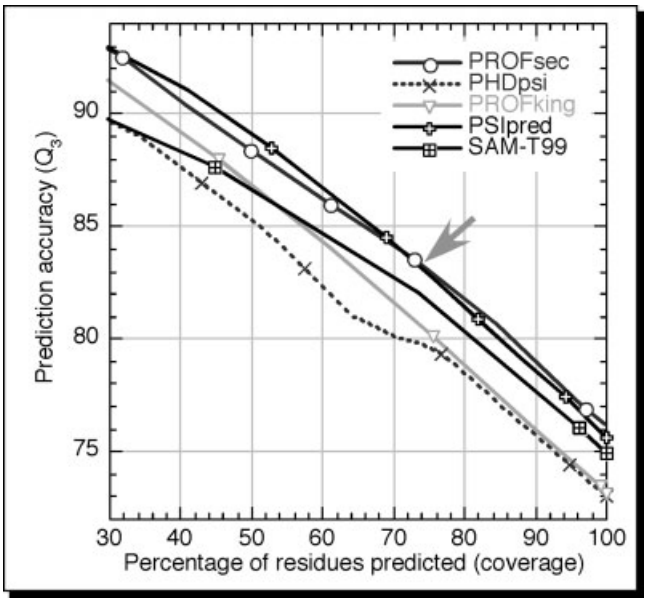

Fig. 2. Stronger secondary structure predictions more accurate. Many secondary structure specialists provide an index for the reliability of the prediction of each residue. Methods in the other categories typically model only the fraction that overlaps best; in analogy, secondary structure prediction methods could restrict their prediction to the most strongly predicted residues. Shown is the accuracy versus coverage for a subset of 205 sequence-unique proteins from EVA. For instance, when focusing on the most reliably predicted 74% of all residues, prediction accuracy (Q_3, Eqn. 1) rose to over 84% for PSIpred and PROFsec (arrows).

random prediction would yield. Our simple model of random is given by the average density of observed contacts: Imp = Acc / (C/L'), where C is the number of observed contacts, and L'=L-R. R describes the local sequence neighbourhood; we generated three different random models for R=6, 12, and 24. Contact predictions might be useful even if the observed contacting pair i/j is not predicted but the actual prediction is between (i + 1) and j. This is reflected by a delta evaluation[51] giving the percentage of correctly predicted contacts within delta residues of the experimental contact; we tested delta values from 0 to 5. This means that a predicted contact between two residues i and j is considered correct if there is a contact observed between (i-delta, i+delta) and (j-delta, j+delta).

Distances

We also evaluated a threshold independent performance that reflects the degree to which a model reproduced a distance map. In particular, Xd is a weighted harmonic average difference between the predicted and the observed inter-residue pair distances:

$$Xd = \frac{1}{15}\sum_{i=1}^{15}\frac{P_i^{prd} - P_i^{obs}}{d_i} \qquad (6)$$

where the sum runs over all 15 distance bins from 0 to 60 Å; d_i is the distance representing bin i (0-4, 4-8, . . . 56-60); P_i^{prd} is the percentage of pairs predicted in bin i, and P_i^{obs} the percentage of pairs observed in bin i. Negative values of Xd indicate that the inter-residue distances are predicted to be closer than observed.[52,53] Note, for the random background, Xd approaches 0, for 'good predictions' Xd values are positive.

Assessing contacts not local in sequence

We applied all scores for different thresholds in discarding sequence-local contacts. Namely, we separately assessed sequence separations of ≥6, ≥12 and ≥24 residues. For example, for a separation of 6, this implies that all contacts i/j with $|i-j|<6$ were ignored.

Reduction of data set

Some targets were not evaluated for technical reasons: T0134 and T0139 had numbering irregularities, T0131 had an unreliable structure, T0145 appeared to be a natively unfolded protein and T0144, T0158, T0163, T0164, T0166, T0171, T0175 and T0180 structures were not available on time. The raw evaluation results for contacts and distances are available at: http://www.pdg.cnb.uam.es/eva/cafasp3/.

RESULTS

Contact Prediction of All Categories at CAFASP3

Despite the small number of targets, it was clear that the contact prediction methods CMAPpro,[16] CORNET,[17] and Chris Bystroff's CASP contact predictions performed on par with 'homology modelling' servers, and that a few fold recognition methods performed slightly better (Fig. 3). Remarkably, methods from the 'novel fold' category performed clearly worse than the contact specialists (note the data for 'novel fold' is omitted from Fig. 3 due to a lack of common sets of reasonable size; data on web). To some extent, the higher accuracy of fold recognition methods may originate from modelling the native structures only partially (Fig. 3, coverage of models in light grey bars). However, even when modelling the entire native structure, contact specialists correctly predict only 20-40% of all observed long range contacts (Fig. 3, coverage of contacts in black bars). Obviously, homology modelling methods performed much better in their 'native' range of sequence similarity, i.e. for targets similar to known structures (40-60% accuracy, data on web). Although the CAFASP3 data sets were of limited size, we could not note any significant difference between the CAFASP3 and the EVA results for the few contact prediction servers for which we had sustained results.

Best Contact Specialist Similar

A detailed inspection of two genuine contact prediction methods (CMAPpro and CORNET, Fig. 4) revealed that both performed - on average - similarly. CMAPpro tended to be superior for sequence-unique targets; CORNET for those with homology to known structures. This observation may be explained by that CMAPpro was trained on a more recent, larger version data set of known structures.

TABLE II. EVA Results on 247 Sequence-Unique Proteins Not Used at CASP (10-2002)†

Rank	Method	Q_3	SOV	BAD	info	contDH	contDE
1	PROFsec	75.4	71.0	2.2	0.34	6.9	4.5
	PSIpred	75.3	70.8	2.2	0.38	6.4	5.1
	SAM-T99sec	75.1	69.4	1.8	0.36	7.1	4.7
2	JPred	73.7	67.6	1.9	0.33	8.0	6.2
	PHDpsi	73.6	68.0	2.6	0.28	7.5	5.2
	PROFking	73.1	67.1	2.5	0.31	7.8	6.5
3	PHDsec	70.2	65.3	4.0	0.24	9.3	6.3

†All abbreviations as in Table I.

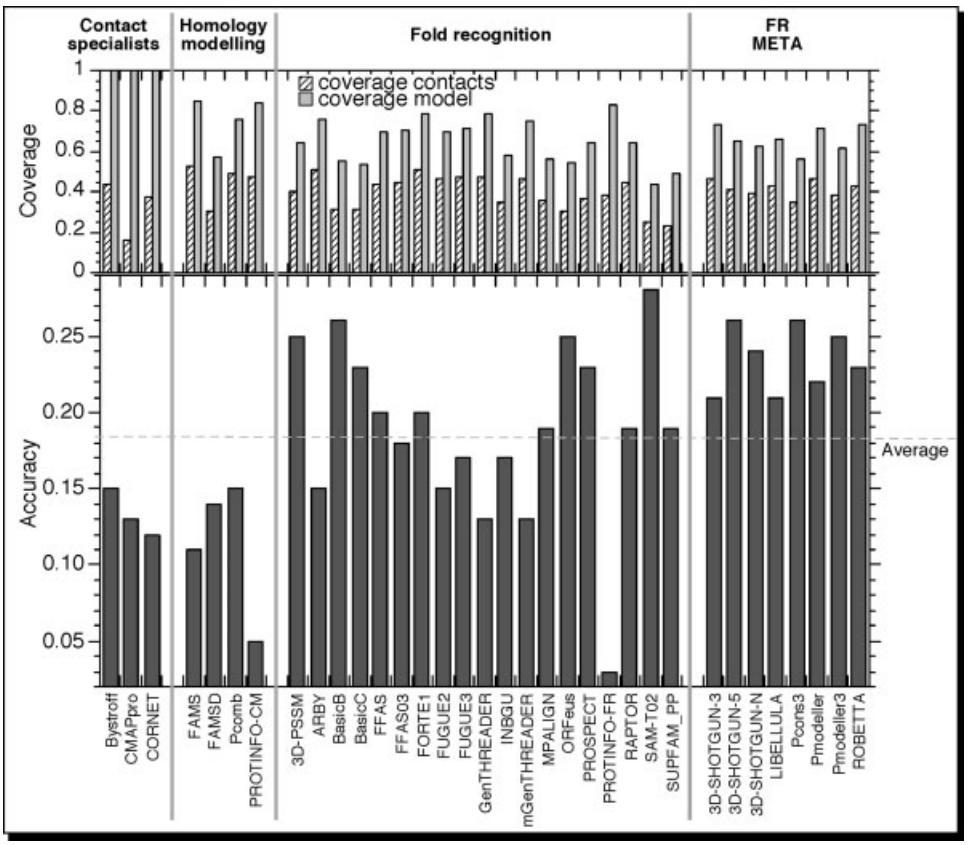

Fig. 3. Long-range contacts for all categories at CAFASP3. The data includes only the 15 sequence-unique targets common to all methods shown. Given is the average contact prediction accuracy (lower panel, Eqn. 5) for all contacts between all pairs of residues that are separated by at least 24 residues in sequence, i.e. are non-local in sequence. The top panel shows two aspects of coverage, namely the coverage of the model (residues modelled as percentage of residues in native structure, light grey bars), and the coverage of predicted contacts (contacts correctly predicted as percentage of contacts observed, stippled bars).

FOLD RECOGNITION

Servers

Many servers subscribed to CAFASP3; only some returned predictions for all targets. While there are various ways around this problem for a comparative analysis of their performance, here we evaluated only those servers that returned predictions for all the sequence-unique targets (Data); all data are on the web at http://cubic.bioc.columbia.edu/eva/cafasp3/.

For comparison purposes we used data from servers being currently evaluated by EVA-FR: 3D-PSSM,[12] FUGUE2,[18] LIBELLULA,[22] LOOPP,[23] Prospect,[28] SAM-T99,[54] SUPERFAMILY[34] and 3 locally installed alignment methods (BLAST,[15] PSI-BLAST,[29] and SSEARCH)[32] in the fold recognition category.

Assessment

Scores

EVA-FR uses many measures to assess the quality of modelling the backbone. Both alignment-dependent and independent scores are employed[55-57] (details on the web). Here we concentrate on evaluating three main aspects of

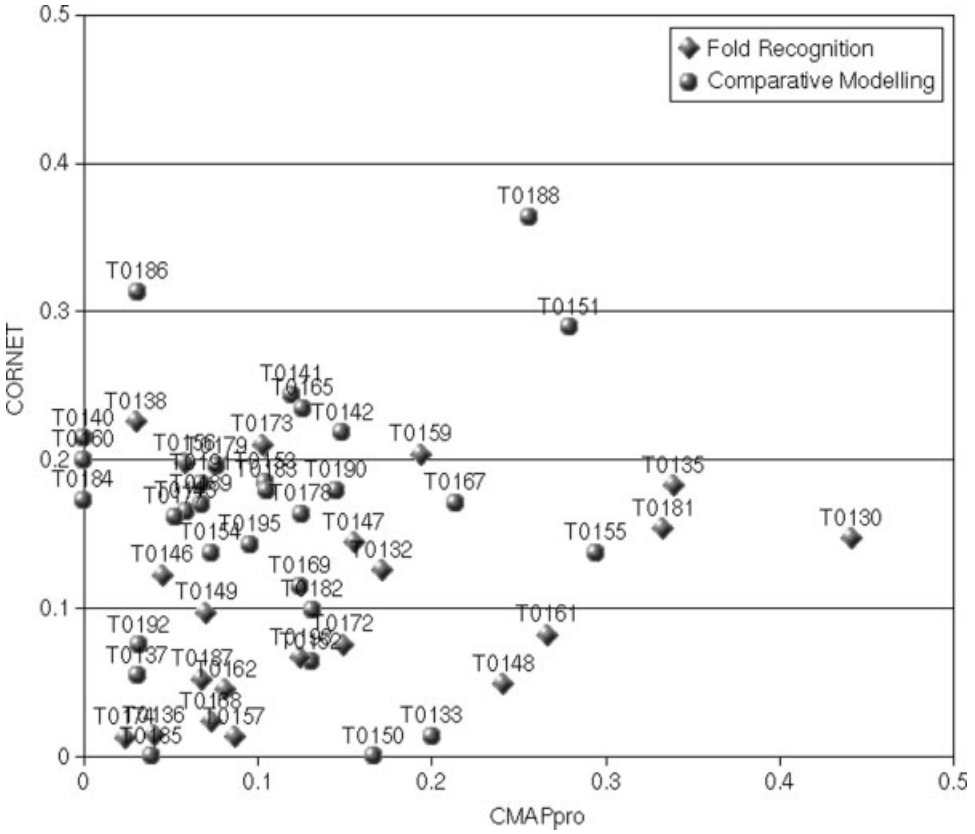

Fig. 4. Direct comparison between CMAPpro and CORNET. The contact prediction accuracy for long-range contacts (sequence separation $>$ 24 residues) is compared for all proteins predicted by the two best contact prediction specialists.

servers performance: (1) the ability to create models that resemble the known structure (alignment-independent scores), (2) the ability to create models locally similar to the known structure (alignment-dependent scores), and (3) the global quality of models generated (alignment-dependent scores). We used the MAMMOTH program[57] to measure alignment-independent performance. MAMMOTH finds the superposition between the model and the known structure that maximises the number of residues below 4Å RMS distance; results are reported as a probability to find such a match by chance in random structural alignments (P-values). We also employed MAMMOTH to establish whether or not the fold was correctly recognised. In particular, we used the threshold suggested by Ortiz et al. (negative logarithm of P-value $\geq$4.5). We used the LGscore program[56] to measure the alignment-dependent performance. LGscores finds the statistically most significant local alignment-dependent superposition of model and experimental structure (P-values). Finally, we evaluated the global quality by globally superimposing all modelled residues with the equivalent residues in the known structure and minimising the C-alpha RMS distance. Toward this end, we used the program ProFit (Andrew Martin, unpublished). We reported the percentage of target residues modelled below 3.5Å. This score complements the other two by distinguishing between models that were consistently good from those that had some correctly and some incorrectly predicted regions.

Ranking

We refrained from explicitly ranking individual methods for the following two reasons. Firstly, according to our analysis the set of CAFASP proteins was too small to reliably estimate the performance of the participating servers. In fact, for those servers that are currently evaluated on much larger data sets by EVA-FR the average differed considerably between the small CAFASP3 and the large EVA-FR data sets. Secondly, bootstrap estimates indicated that at the 66% confidence level most of the methods did not differ from each other. We randomly picked 1000 sets of 25 proteins out of the set of 25 sequence-unique ones. Then we compiled the standard deviations for the distributions of the average scores between the 1000 random samples. We used this standard deviation to estimate the 66% confidence level (one sigma).

Results

No single server significantly best

We performed a bootstrap analysis to establish the significance of differences in performance amongst servers. At a significance level of one standard deviation

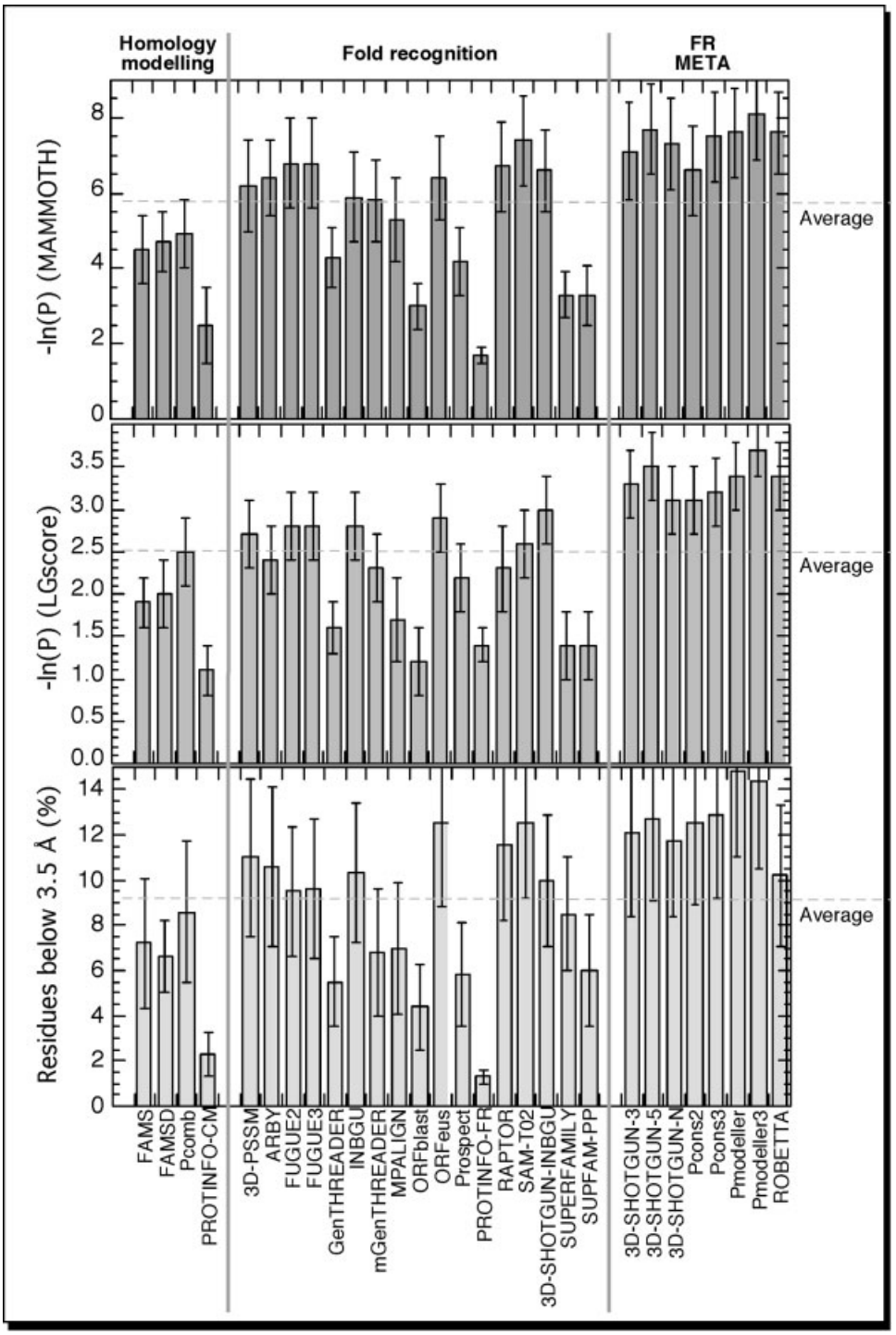

Fig. 5. Comparison between fold recognition and comparative modelling methods. The data was compiled for the 25 sequence-unique targets from CASP5 (Data); for each method only the first hit was considered. Scores used: average negative natural logarithm of P-values reported by MAMMOTH[57] (top panel), LGscore[56] (central panel), and the average percentage of residues below 3.5Å in an optimal global superposition between model and native structure. The error bars were derived from bootstrap analysis and correspond to one standard deviation of a statistic. Overall, the results between LGscore and MAMMOTH correlated to 0.95, those between LGscore and the global percentage to 0.90, and those between MAMMOTH and the global percentage to 0.93.

unequivocal 'winners' cannot be identified (Fig. 5). Thus, based on the CAFASP data set we cannot point to a best fold recognition server. However, we can point to two servers that were one standard deviation above the average over all methods for all three scores we employed, namely Pmodeller and Pmodeller3 (Arne Elofsson, unpub-

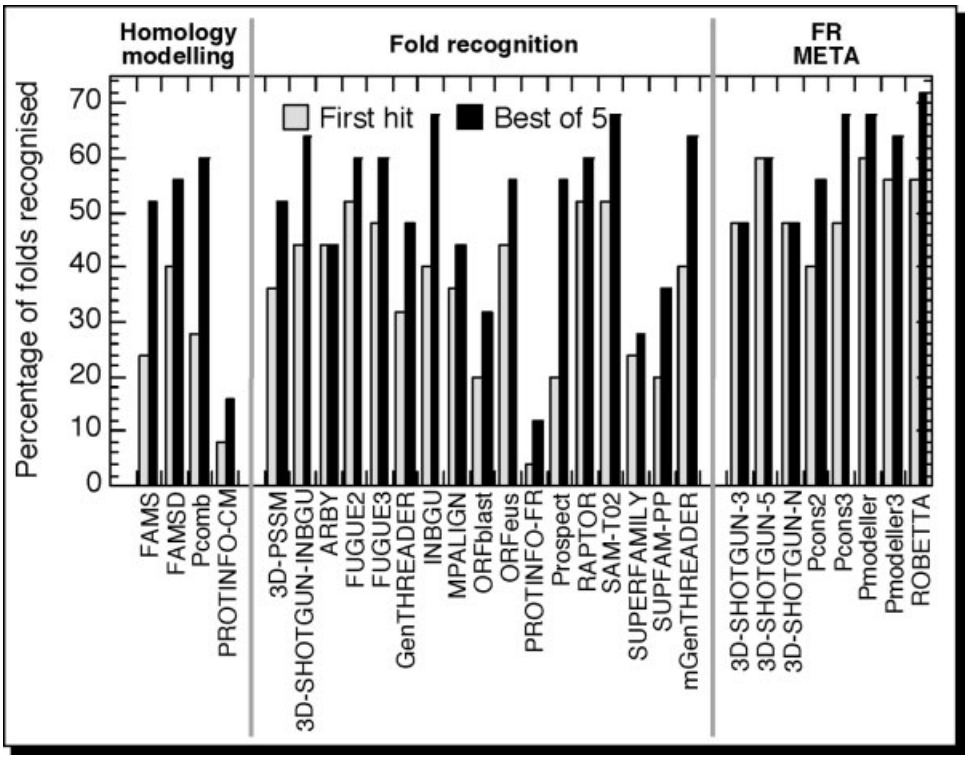

Fig. 6. Potential improvement through optimal internal scoring. The percentage of models that correctly recognised the fold (negative natural logarithm of MAMMOTH P-value >4.5) varied considerably between first hits and the best-of-5. For instance INBGU reaches a similar level of 'correct fold in first 5' as the best META-servers; however, the META-servers have more correct first hits than INBGU. Overall, META-servers appear characterised by that they reduce the difference between the two values in comparison to the fundamental method used as input to the META-servers.

lished). Three servers were one standard deviation below the average for all three scores (Fig. 5).

For many servers performance depended sensitively on what was measured

How well a given method performed depended in many cases on a particular measure. For example, the best META-servers were close to be significantly better than single servers when evaluating residues modelled locally correctly (LGscore, Fig. 5). However, the difference was much less pronounced when evaluating alignment-independent similarity and the global quality of models (top and bottom panels of Fig. 5). Nevertheless, no META-servers fall below the average over all methods of any score shown. In contrast, some original servers performed well above the average by one measure and below by another.

Correct ranking of models could yield significant improvements

For CAFASP3 servers could return up to five models for each target. Most of servers used the opportunity and returned more than one prediction. Trivially, the odds of correctly recognising the fold as 1-in-5 are higher than those of getting the first hit correct (Fig. 6). Nevertheless, the comparison between 'first correct' and '1-in-5 correct' highlighted interesting differences: homology modelling servers (for targets that are traditionally not their own realm) had the least successful internal scores to sort their alignments (largest difference between two bars in Fig. 6), while the META-servers had the most successful internal scores (smallest differences). Thus, the major achievement of META-servers originates from successful internal scoring functions.

Still room for improvement

The best servers correctly recognised the fold for about 60% of the proteins (Fig. 6). If we could identify the best model from all servers, we would recognise the correct fold as the first hit for 88% of the targets and for 96% of the targets if we could identify the best model amongst all first-5 of all servers. This high rate also points to one feature of MAMMOTH: it uses the shorter of two aligned segments to compile the percentage overlap; for 3 of the 'novel folds' MAMMOTH suggests a similarity to known folds. These were T0129 with 1lja (INBGU rank 2), T0139 with 1doq (Pcomb model 2), and T0161 with 2a0b (3D-PSSM model 2). Interestingly, for the first two, LGscore2 also suggests a 'real relation'. To some extent this 'mistake' highlights what has been heavily debated at CASP all along: homology modelling and fold recognition methods frequently pick up structural similarities that are very short, when can we objectively conclude that this local similarity is not meaningful? And more generally: will the concept of a 'fold' turn out to be a temporary concept that will disappear once the PDB will be sufficiently representa-

tive of all structures to demonstrate that structure space is continuous? Despite this minor problem with MAMMOTH, the high rate of 'recognition of fold' for all methods suggested that current META-servers still miss out on some opportunities (highest score for first-5 was 72%, i.e. significantly lower than 96%). On the other hand, when we considered the global quality of models the picture appeared less impressive: even if we could identify the best in all first-5 from all servers, only about 26% of all residues were modelled below 3.5 Å RMS distance (37% below 5 Å). When applying a local structural alignment algorithm, namely CE,[55] 37% of all residues were aligned below 3.5 Å RMS distance (48% below 5 Å).

Sustained performance of fold recognition methods

For seven of the CAFASP3 methods (3D-PSSM, FUGUE2, LIBELLULA, LOOPP, Prospect, SAM-T99, SUPERFAMILY), EVA has been collecting significant data sets in the past. These data suggest that the 25 sequence-unique CASP/CAFASP proteins were not representative. Instead, six of the seven methods performed better on these proteins than they did on average for more significant data sets.

DISCUSSION AND CONCLUSIONS

Secondary Structure Prediction Still Advancing

Secondary structure prediction methods have evolved from the long-time best PHDsec. Most of the improvement (from PHD~70.2 to PHDpsi~73.6, Table 2) originated from larger sequence databases (PHDsec and PHDpsi differ only in the alignments: MaxHom[58] against SWISS-PROT[59] for PHDsec vs. iterated PSI-BLAST[29] against BIG (PDB[60] + TrEMBL + SWISS-PROT[59]) for PHDpsi. However, an impressive additional improvement originated from more refined algorithms (PSIpred, SAM-T99sec, and PROFsec in Table 2). Particularly striking is that the number of 'very bad' errors, i.e. the confusions between helix and strand has almost been halved. The improvement in the accuracy of predicting the overall content in secondary structure appears numerically less striking, however, the sustained error rates for the best methods remain below impressive averages of 7% of helix and below 6% for strand. While the small CAFASP3 data sets suggested that secondary structure prediction experts are more reliable in their own realm than methods from the homology modelling and fold recognition category (when 'mis-used' to only predict secondary structure, Fig. 1). The analysis of more reliably predicted residues highlights the advance of secondary structure prediction: the secondary structure predictions for most non-specialists covered less than 50% of the proteins at levels below 70% accuracy (Fig. 1); at this coverage, the best specialists reach levels around 88% accuracy (Fig. 2). The best method of the first CASP meeting achieved 88% for fewer than 35% of all residues.[61] In other words, more than 15% of the residues are now predicted at levels that resemble the similarity between similar structures.[47,41] Undoubtedly, five CASP meetings provided the incentive to advance a crucial tool from good to better. Two new promising methods appeared just before CASP5 (APSSP2, SAM-T02sec); we still do not have sufficient data from EVA to assess whether or not their high performance is sustained.

Inter-residue Contact Prediction Methods Under-rated

For sequence-unique proteins, contact prediction specialists performed on par with many methods from all other categories (Fig. 3). In fact, the best specialists appeared more useful in predicting relevant constraints at CAFASP3 than the 'new fold' methods that attracted more attention at the recent CASP meetings. Furthermore, for some targets (T0148, T0161, and T0181, data on the web), the specialists were superior to all other methods. While three targets are too few to establish any conclusion about the meaning of this result, it clearly illustrated that the specialists capture additional information relevant for structure prediction.

Fold Recognition META-Servers Successful, But Not THE Winners

EVA-FR is still not consolidated enough to provide a comprehensive picture of the sustained performance for the field of fold recognition. However, for a few methods for which we had results from both small CAFASP3 (25 proteins) and larger EVA (74-262 proteins depending on the server) sets, the average scores differed considerably. Hence, detailed conclusions from the CAFASP3 results are meaningless. Nevertheless, the following three tendencies may hold in general. (1) Methods differed in the aspect of fold recognition performance for which they were better or worse (Fig. 5). (2) META-servers succeeded partially in harvesting the potential created by sub-optimal internal scoring functions used by the fundamental fold recognition methods: the differences in the percentage of correctly predicted first and first-5 hits was considerably lower for META-servers than for the fundamental methods exploited by META-servers (Fig. 6). While the best META-servers now appear to recognise the correct fold in over 50% of the CAFASP3 proteins as the first hit (Fig. 6), the fold was correctly predicted by one of the methods for 96% of the proteins. Hence, META-servers still do not exploit the full potential. The best fundamental methods correctly recognised the fold as first hit in almost half of all CAFASP3 proteins (Fig. 6). If this performance can be sustained, it appears that fundamental methods have improved. (3) The quality of the models created in the fold recognition domain appears still rather low: less than 15% of all residues of the first model superposed globally to $<$3.5Å C-alpha RMS distance (Fig. 5, bottom); even if we could identify the best model of the first five, this number would still not rise above 26% of all globally aligned residues.

CASP5/CAFASP3 Sequence-Unique Targets Were Not Representative

All secondary structure prediction methods performed above average for the 21 sequence-unique targets at CASP5/CAFASP3 (Table 1, Table 2). We found the same to be true for the few fold recognition methods for which we

also had larger sets from EVA. Thus, the sequence-unique CASP5 targets were obviously not representative for secondary structure prediction and fold recognition. While we do not have a good comparison for the even smaller sets (six domains) used to evaluate 'novel fold' methods, the observation that contact specialists predicted contacts more accurately than 'novel fold' methods, somehow may slightly damp the optimism about progress. Nevertheless, we were stunned to which extent the large data sets from EVA supported some of the conclusions that cautious observers may have brought home from CASP5.

Is Ranking at CASP/CAFASP Scientific?

The evaluation of secondary structure prediction methods is perceived as a rather straightforward task. One reason may be that there are many large-scale, well-studied analyses of these methods. We observed that when basing the analysis only on the CAFASP3 targets there is ample cause for debate which method won even for this category. We challenge that secondary structure prediction methods should continue to be an essential component of CASP because of two main reasons: (1) secondary structure prediction methods have been improved significantly since the first CASP. We doubt this would have happened without CASP assessing this category. (2) Without assessing CAFASP3 secondary structure predictions, we would have only limited evidence to conclude that this year's targets suggested over-optimistic estimates. Additionally, secondary structure prediction methods constitute an ideal test-bed for choosing appropriate ranking schemata. From CASP4, we learned that ranking must be based on identical subsets;[1] from CASP5, we learned that 'most often better than average' or 'most often best' also paint completely distorted pictures. There may be a way to get from too small data sets to scientifically founded declarations of winners. If so, we have not found it, yet.

Where From Here: Servers Take Over CASP/CAFASP?

The CAFASP assessment of automatic servers has largely been taken over by automatic servers like LiveBench[62,63] and EVA. The data that we presented here suggested many reasons for running automatic servers on the subset of CASP targets separately. Not the least is that CASP constitutes an important incentive for continued development and that what became ready just in time for CASP may not be ready for large-scale evaluation, yet. EVA has some protocol enabling to test a new method X in the context of others without releasing the results to anyone but the authors of X. However, method X still has to cope with the data flow (about 3000 rather than 100 targets during the CASP5 prediction season). For some methods, this may be prohibitive. Nevertheless, LiveBench and EVA could seamlessly handle automatic servers without any additional headaches for developers; expert assessors could benefit from comparing the background performance; together assessors, the teams from the PredictionCenter and from LiveBench and EVA could refine the looking glass for assessment from year to year. Many problems with CASP originate in inappropriate interpretations, rankings, and blowing up tiny details. In our dream of combined resources, the CASP idea could survive the next decades without causing such unscientific stress.

ACKNOWLEDGMENTS

Thanks to Jinfeng Liu and Megan Restuccia (Columbia) for computer assistance; to Leszek Rychlewski (BioInfoBank, Poland) for collecting some of the data used here through his META-server. Thanks to Marc Marti-Renom, Mallur Madhusudhan, and Andrei Sali, (UCSF) for running EVA-CM and their tremendously valuable contributions to the whole EVA project. Thanks to Angel Ortiz and Dmitry Lupyan (MSSM, New York) for support in using MAMMOTH, and to Andrew Martin (Reading University, England) for the ProFit program used for the global structural superposition. Thanks also to the members of the Protein Design Group (PDG, Madrid) and in particular to David Juan and Ramon Alonso-Allende for the continuous support and interesting discussions; the contribution of the PDG is supported in part by the grants BIO2000-1358-CO2-01 from the Spanish Ministry of Science and Technology, CSIC:LIFE/001/0957 UE:QLRT-2001-00015 from TEMBLOR, CSIC:LIFE/992/0553 UE:QLK3-CT-2000-00079 from SANITAS, and CSIC:LIFE/991/0219 UE:QLK3-CT-1999-00875 from RAS. . IK was supported by the grant 5-P20-LM7276 from the National Institute of Health (NIH), DP and BR were supported by the NIH grant RO1-GM63029-01. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

REFERENCES

1. Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Fiser, A., Pazos, F., Valencia, A., Sali, A. and Rost, B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17:1242–1243, 2001.
2. Koh, I., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Fiser, A., Pazos, F., Valencia, A., Sali, A. and Rost, B. EVA: evaluation of protein structure prediction servers. Nucleic Acids Research 2003.
3. Eyrich, V. A. and Rost, B. META-PP: single interface to selected web servers. Nucleic Acids Research 2003.
4. Moult, J., Pedersen, J. T., Judson, R. and Fidelis, K. A large-scale experiment to assess protein structure prediction methods. Proteins: Structure, Function, and Genetics 23:ii–iv, 1995.
5. Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K. and Pedersen, J. T. Critical assessment of methods of protein structure prediction (CASP): Round II. Proteins: Structure, Function, and Genetics Suppl 1:2–6, 1997.
6. Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K. and Pedersen, J. T. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins: Structure, Function, and Genetics Suppl 3:2–6, 1999.
7. Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins: Structure, Function, and Genetics Suppl 5:2–7, 2001.
8. Jones, S., Stewart, M., Michie, A. D., Swindells, M. B., Orengo, C. A. and Thornton, J. M. Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein Sci. 7:233–242, 1998.
9. Yang, A. S. and Honig, B. An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J. Mol. Biol. 301:691–711, 2000.
10. Liu, J. and Rost, B. CHOP proteins into structural domain-like fragments. J. Mol. Biol. submitted 2003-03-25, 2003.

11. Rost, B. Twilight zone of protein sequence alignments. Protein Eng. 12:85–94, 1999.
12. Kelley, L. A., MacCallum, R. M. and Sternberg, M. J. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299:499–520, 2000.
13. Fischer, D. 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor. Proteins 51:434–41, 2003.
14. Rychlewski, L., Jaroszewski, L., Li, W. and Godzik, A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9:232–41, 2000.
15. Altschul, S. F. and Gish, W. Local alignment statistics. Methods in Enzymology 266:460–480, 1996.
16. Pollastri, G. and Baldi, P. Prediction of contact maps by GIO-HMMs and recurrent neural networks using lateral propagation from all four cardinal corners. Bioinformatics 18:62S–70, 2002.
17. Fariselli, P., Olmea, O., Valencia, A. and Casadio, R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng. 14:835–843, 2001.
18. Shi, J., Blundell, T. L. and Mizuguchi, K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310:243–57, 2001.
19. Jones, D. T. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287:797–815, 1999.
20. Fischer, D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 119–30, 2000.
21. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. and Barton, G. J. JPred: a consensus secondary structure prediction server. Bioinformatics 14:892–893, 1998.
22. Juan, D., Grana, O., Pazos, F., Fariselli, P., Casadio, R. and Valencia, A. A neural network approach to evaluate fold recognition results. Proteins 50:600–8, 2003.
23. Meller, J. and Elber, R. Linear programming optimization and a double statistical filter for protein threading protocols. Proteins 45:241–61, 2001.
24. McGuffin, L. J., Bryson, K. and Jones, D. T. The PSIPRED protein structure prediction server. Bioinformatics 16:404–405, 2000.
25. Olmea, O., Rost, B. and Valencia, A. Effective use of sequence correlation and conservation in fold recognition. J. Mol. Biol. 293:1221–1239, 1999.
26. Rost, B. PHD: predicting one-dimensional protein structure by profile based neural networks. Methods in Enzymology 266:525–539, 1996.
27. Ouali, M. and King, R. D. Cascaded multiple classifiers for secondary structure prediction. Protein Sci. 9:1162–1176, 2000.
28. Xu, Y. and Xu, D. Protein threading using PROSPECT: Design and evaluation. Proteins 40:343–354, 2000.
29. Altschul, S., Madden, T., Shaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Research 25:3389–3402, 1997.
30. Jones, D. T. Protein secondary structure prediction based on position- specific scoring matrices. J. Mol. Biol. 292:195–202, 1999.
31. Karplus, K., Barrett, C. and Hughey, R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 14:846–856, 1998.
32. Smith, T. F. and Waterman, M. S. Identification of common molecular subsequences. J. Mol. Biol. 147:195–197, 1981.
33. Baldi, P., Brunak, S., Frasconi, P., Soda, G. and Pollastri, G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 15:937–946, 1999.
34. Gough, J. and Chothia, C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Research 30:268–272, 2002.
35. Przybylski, D. and Rost, B. Alignments grow, secondary structure prediction improves. Proteins: Structure, Function, and Genetics 46:195–205, 2002.
36. Xu, Y. and Uberbacher, E. C. A polynomial-time algorithm for a class of protein threading problems. Comput. Appl. Biosci. 12:511–517, 1996.
37. Pollastri, G., Przybylski, D., Rost, B. and Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47:228–235, 2002.
38. Fischer, D., et al. CAFASP-1: Critical assessment of fully automated structure prediction methods. Proteins 209–217, 1999.
39. Fischer, D., Elofsson, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., Ortiz, A. R. and Dunbrack, R. L. CAFASP2: The second critical assessment of fully automated structure prediction methods. Proteins 171–183, 2001.
40. Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
41. Andersen, C. A. F., Palmer, A. G., Brunak, S. and Rost, B. Continuum secondary structure captures protein flexibility. Structure 10:175–184, 2002.
42. Frishman, D. and Argos, P. Knowledge-based protein secondary structure assignment. Proteins: Structure, Function, and Genetics 23:566–579, 1995.
43. Holm, L. and Sander, C. Database Algorithm for Generating Protein Backbone and Side- Chain Coordinates from a C-Alpha Trace Application to Model- Building and Detection of Coordinate Errors. J. Mol. Biol. 218:183–194, 1991.
44. Zemla, A., Venclovas, C., Fidelis, K. and Rost, B. A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. Proteins: Structure, Function, and Genetics 34:220–223, 1999.
45. Defay, T. and Cohen, F. E. Evaluation of current techniques for ab initio protein structure prediction. Proteins: Structure, Function, and Genetics 23:431–445, 1995.
46. Rost, B. and Sander, C. Prediction of protein secondary structure at better than 70- percent accuracy. J. Mol. Biol. 232:584–599, 1993.
47. Rost, B., Sander, C. and Schneider, R. Redefining the goals of protein secondary structure prediction. J. Mol. Biol. 235:13–26, 1994.
48. Rost, B. and Eyrich, V. EVA: large-scale analysis of secondary structure prediction. Proteins: Structure, Function, and Genetics 45 Suppl 5:S192–S199, 2001.
49. Marti-Renom, M. A., Madhusudhan, M. S., Fiser, A., Rost, B. and Sali, A. Reliability of assessment of protein structure prediction methods. Structure 10:435–440, 2002.
50. Fidelis, K., Veclovas, C. and Zemla, A. Protein structure prediction center. Lawrence Livermore, 2000.
51. Ortiz, A. R., Kolinski, A., Rotkiewicz, P., Ilkowski, B. and Skolnick, J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins: Structure, Function, and Genetics Suppl 3:177–185, 1999.
52. Pazos, F., Helmer-Citterich, M., Ausiello, G. and Valencia, A. Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. 271:511–523, 1997.
53. Pazos, F., Olmea, O. and Valencia, A. A graphical interface for correlated mutations and other protein structure prediction methods. Computer Applications in Biological Science 13:319–321, 1997.
54. Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L. and Hughey, R. Predicting protein structure using only sequence information. Proteins: Structure, Function, and Genetics S3:121–125, 1999.
55. Shindyalov, I. N. and Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11:739–747, 1998.
56. Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L. and Elofsson, A. A study of quality measures for protein threading models. BMC Bioinformatics 2:5, 2001.
57. Ortiz, A. R., Strauss, C. E. and Olmea, O. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. Protein Sci. 11:2606–2621, 2002.
58. Sander, C. and Schneider, R. Database of homology-derived structures and the structural meaning of sequence alignment. Proteins: Structure, Function, and Genetics 9:56–68, 1991.
59. Boeckmann, B., et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31:365–370, 2003.
60. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. The Protein Data Bank. Nucleic Acids Research 28:235–242, 2000.
61. Rost, B. and Sander, C. Progress of 1D protein structure prediction at last. Proteins: Structure, Function, and Genetics 23:295–300, 1995.
62. Bujnicki, J. M., Elofsson, A., Fischer, D. and Rychlewski, L. LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci. 10:352–361, 2001.
63. Rychlewski, L., Elofsson, A. and Fischer, D. LiveBench and CAFASP. Proteins special CASP5 issue:2003.