MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison

ANGEL R. ORTIZ,[1] CHARLIE E.M. STRAUSS,[2] AND OSVALDO OLMEA[1]

[1]Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York University, New York, New York 10029, USA
[2]Biosciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

(RECEIVED May 21, 2002; FINAL REVISION August 12, 2002; ACCEPTED August 15, 2002)

Abstract

Advances in structural genomics and protein structure prediction require the design of automatic, fast, objective, and well benchmarked methods capable of comparing and assessing the similarity of low-resolution three-dimensional structures, via experimental or theoretical approaches. Here, a new method for sequence-independent structural alignment is presented that allows comparison of an experimental protein structure with an arbitrary low-resolution protein tertiary model. The heuristic algorithm is given and then used to show that it can describe random structural alignments of proteins with different folds with good accuracy by an extreme value distribution. From this observation, a structural similarity score between two proteins or two different conformations of the same protein is derived from the likelihood of obtaining a given structural alignment by chance. The performance of the derived score is then compared with well established, consensus manual-based scores and data sets. We found that the new approach correlates better than other tools with the *gold standard* provided by a human evaluator. Timings indicate that the algorithm is fast enough for routine use with large databases of protein models. Overall, our results indicate that the new program (MAMMOTH) will be a good tool for protein structure comparisons in structural genomics applications. MAMMOTH is available from our web site at http://physbio.mssm.edu/~ortizg/.

Keywords: Protein structural alignment; model evaluation; protein structure prediction; structural genomics

The challenge of structural genomics is to be able to extract useful biological information about the biochemical role of every protein in the organism (Brenner 2001; Mittl and Grutter 2001; Thornton 2001; Chance et al. 2002). Regardless of its final degree of success, structural genomics is beginning to shift structural biology research from a reductionist to a more integrative view (Teichmann et al. 2001; Burley and Bonanno 2002; Hurley et al. 2002). To fully realize its potential, structural genomics needs rigorous and fast methods to compare at large scale the vast number of both experimental protein structures and models, involving disparate resolution levels, that will be produced over the next years (Sali 1998; Vitkup et al. 2001). The ability to compare theoretical and low-resolution models with high-resolution experimental structures can be expected to be particularly relevant. Experimental methods will be providing high-resolution structures for a subset of proteins, but modeling techniques at different resolution levels will likely be used to obtain structural information for the bulk of sequences (Baker and Sali 2001). In addition, high-throughput determination approaches in X-ray crystallography (Adams and Grosse-Kunstleve 2000) and nuclear magnetic resonance (Prestegard et al. 2001; Al-Hashimi and Patel 2002) will deliver largely automatically generated structures, but at the expense of resolution, structural refinement, and manual checking. Therefore progress is dependent upon, among other factors, having tools to match structurally predicted conformations and low-resolution models with experimentally determined structures. The field of pro-

Reprint requests to: Angel R. Ortiz, 1 Gustave Levy Place, Box 1218, New York, NY 10029, USA; e-mail: ortiz@inka.mssm.edu; fax: (212) 860-3369.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.0215902.

tein structural alignment and fold classification is mature, and a number of excellent approaches are available for this task (Holm and Sander 1993; Gibrat et al. 1996; Holm and Sander 1996; Lackner et al. 2000; Yang and Honig 2000). However, comparing a predicted conformation with an experimental structure has, as we will show, a certain number of peculiarities that, in our view, deserve the application of a specialized tool.

Structural comparisons involving models will also be increasingly required in automated methods for function annotation (Baker and Sali 2001). The availability of large number of structures and sequences is fueling encouraging developments in ab initio protein structure prediction (Orengo et al. 1999; Ortiz et al. 1999; Simons et al. 1999) and related techniques. It may soon be possible to obtain functional annotations at genomic scale for new open reading frames following a sequence-structure-function paradigm (Thornton et al. 2000): First, structure prediction can be used to provide candidate folds for the query sequence (Ortiz et al. 1998a; Simons et al. 2001). Then, putative functions are inferred on the basis of structural alignments to proteins of known structure and function, with (Lichtarge and Sowa 2002; Madabushi et al. 2002) or without concomitant sequence analysis (Fetrow and Skolnick 1998; Fetrow et al. 1998; Ortiz et al. 1998b; Simons et al. 2001).

The comparison of structural models with experimental structures is intimately related to the problem of evaluating structure predictions. A common way to evaluate success in structure prediction is to study the structural similarity between predicted and experimental conformations. A usual measure of this similarity is given by the root mean square distance (RMS) between the positions of the corresponding atoms on these proteins, after the structures have been superimposed by an optimal three-dimensional rigid body rotation. Although this approach is adequate for comparing two closely related structures, it does not work well when the structures are remotely related. The reason for this is that one should find similar substructures in otherwise partly dissimilar conformations. Portions of those substructures that do not match tend to dominate the RMS value. This is an instance of the classical issue of outliers dominating a fitting measure. Additionally, commonly used least-squares superposition methods suffer from bias introduced into the comparison process by the choice of atoms employed in the superposition. This turns out to be a problem also in structure prediction, where there is usually a one-to-one correspondence at the sequence level between prediction and experiment, because reduced representations—normally employed in structure prediction force fields—blur these correspondences and often produce shifts in registration in different areas of the structure.

Alternative comparison metrics have been proposed, but no consensus has been established to date, as can be exemplified by the array of different approaches used in the successive CASP (critical assessment of methods of protein structure prediction, round III) meetings (Moult et al. 1999). To some extent, this is because the specific features to be compared dictate the similarity measure and algorithm of choice. Our aim here is to compare theoretical models of protein folds and to evaluate fold predictions, and therefore we are interested in determining structural similarity at the fold level. We started with the intuitive idea that a prediction is successful when the modeled structure is significantly more similar to the target fold than to any other known fold. Consequently, the evaluation method should be consistent with consensus classifications of experimental protein structures in folds, such as the manually derived SCOP database by Murzin and collaborators (Murzin et al. 1995; Lo Conte et al. 2000).

In what follows we elaborate on these ideas: First, we developed a fast structural alignment approach that (1) is sequence-independent, (2) focuses on model Cα coordinates, and (3) avoids references to sequence or contact maps. This allows possible registration shifts that tend to happen in secondary structure assignments and different resolution levels, and also takes into account the fact that similar models can have different contact maps. The method is also capable of considering only portions of the target protein, avoiding the need to model the complete chain of the target. Second, and following seminal work by Levitt and Gerstein (1998) and Abagyan and Batalov (1997), we assess the final structural alignment obtained with the algorithm by attaching a statistical significance to the similarity score in the form of a *P*-value, that is, the probability that a better score can occur by chance when comparing two unrelated folds provided by Nature. We then demonstrate the utility of this approach by analyzing models from the CASP3 contest (Moult et al. 1999). Finally, we briefly discuss the applicability of our approach in different areas of structural genomics.

Results

Structural alignments with MAMMOTH

First, we analyzed the quality of the structural alignments produced by MAMMOTH. For that, we compared the fraction of residues aligned using a set of protein pairs comprising some difficult cases. The set is described by Jung and Lee (2000) in their Table 2 and can be found in our Table 1. MAMMOTH provides structural alignments similar to those obtained by other approaches such as Dali (Holm and Sander 1993), Vast (Madej et al. 1995; Gibrat et al. 1996), ProSup (Lackner et al. 2000), and SHEBA (Jung and Lee 2000). Inspection of the superimposed structures confirmed the agreement between the different algorithms (not shown). There is, however, one exception: for the 1acx-1tnf_A pair, MAMMOTH fails to find the correct structural

Table 1. *Comparison of structural alignments obtained with MAMMOTH and with other methods*

Protein 1	Protein 2	MAMMOTH	SHEBA	OTHERS
1acx	1cob_B	77	69	90 (1)
1acx	1tmf–A	25	—	80 (1)
1pts_A	1mup	75	72	76 (1)
2gbl	1ubq	45	44	42 (2)
2gb1	4fxc	42	44	39 (2)
1ubq	4fxc	60	54	48 (2)
1plc	2rhe	64	49	50 (2)
1plc	1acx	51	57	48 (2)
1acx	1rbe	62	59	49 (2)
1aba	1trs	65	64	60 (3)
1aba	1dsb_A	36	52	47 (3)
1aba	1pbf	47	51	36 (4)
1mjc	5tss_A	42	54	50 (4)
1pgb	5tss_A	45	43	43 (4)
2tmv_P	256b_A	68	68	64 (4)
1tnf_A	1bmv_I	54	80	71 (4)
1ubq	1frd	60	56	48 (4)
2rsl_C	3chy	55	59	56 (4)
3chy	1rcf	99	89	75 (4)

This set is taken from Jung and Lee. Table 2 (Jung & Lee, 2000). For each pair, the number of aligned residues after optimal structural alignment, as obtained with the different programs, is shown.
(1) Dali (Holm & Sander, 1993); (2) MLC (Boutonnet et al., 1995); (3) VAST (Madej et al., 1995; Gibrat et al., 1996); (4) ProSup (Lackner et al., 2000)

alignment. SHEBA also has problems with this pair, but Dali is able to find a good solution. In Figure 1 we show two examples of typical structural alignments obtained with MAMMOTH.

A measure of the computational time needed by MAMMOTH as a function of problem size is given in Figure 2, which shows running times for various size protein pairs, as computed in a 500Mhz Alpha workstation. There are 10 million comparisons in this plot, averaged so that each point is the average of 250 comparisons. The double behavior of MAMMOTH running times is due to the different behavior of the MaxSub routine in different length regimes. There is a "phase transition" in the average number of cycles needed for convergence in MaxSub around the 10^4 residues boundary, apparently due to the increase in structural complexity. As can be observed, for a typical comparison of a pair of small proteins (~100 residues), computation takes ~0.02 sec (single processor). The algorithm is faster than most approaches and runs roughly at the same speed as SHEBA or PrISM (Yang and Honig 2000). However, it should be noted that MAMMOTH is more general, as it does not rely upon some of their approximations. SHEBA, for example, is not a pure structural alignment algorithm, since it establishes residue correspondences based on a previous sequence alignment of the proteins to align. PrISM, on the other hand, computes a prealignment using secondary structure vectors, and this makes it inadequate for low-resolution theoretical models or irregular proteins. Of course, additional gain in speed could be obtained in MAMMOTH if a secondary structure filter is applied. Overall, MAMMOTH is able to provide a good compromise between alignment quality, computational speed, and generality.

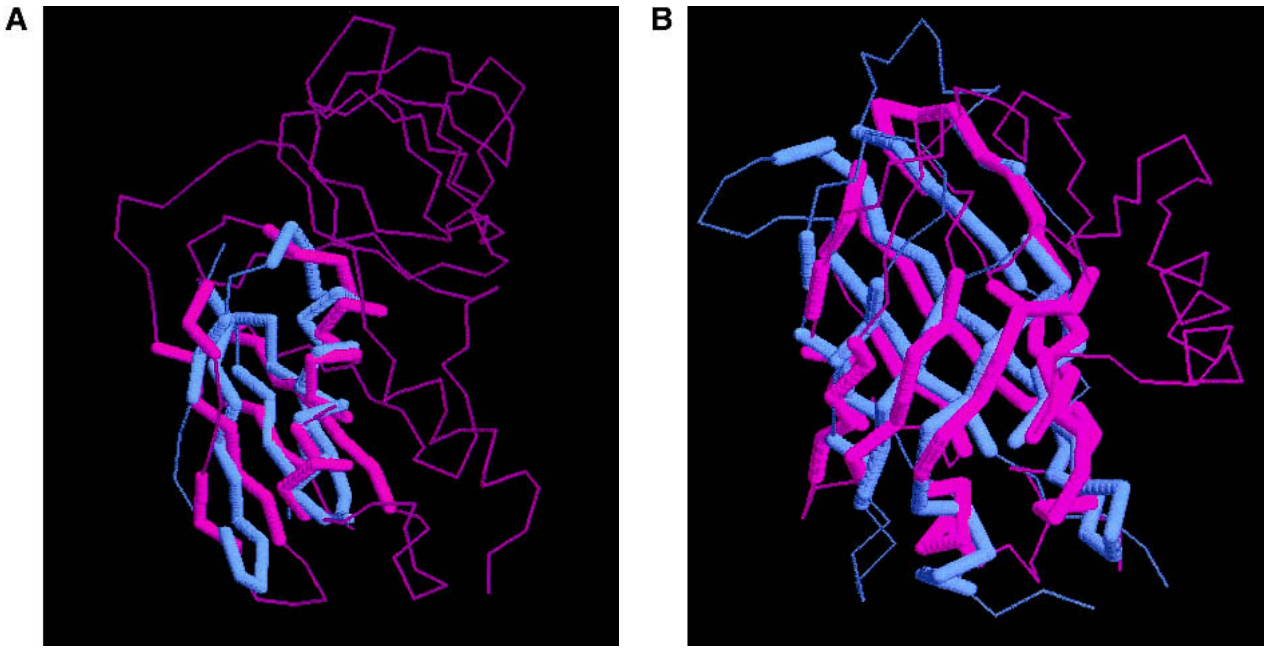

Fig. 1. Examples of structural alignments obtained with MAMMOTH. (*A*) Alignment of 1pts_A with 1mup. The structural alignment score is 9.52; (*B*) Structural alignment of 1pgb with 5tss_A. The score in this case is 6.29.

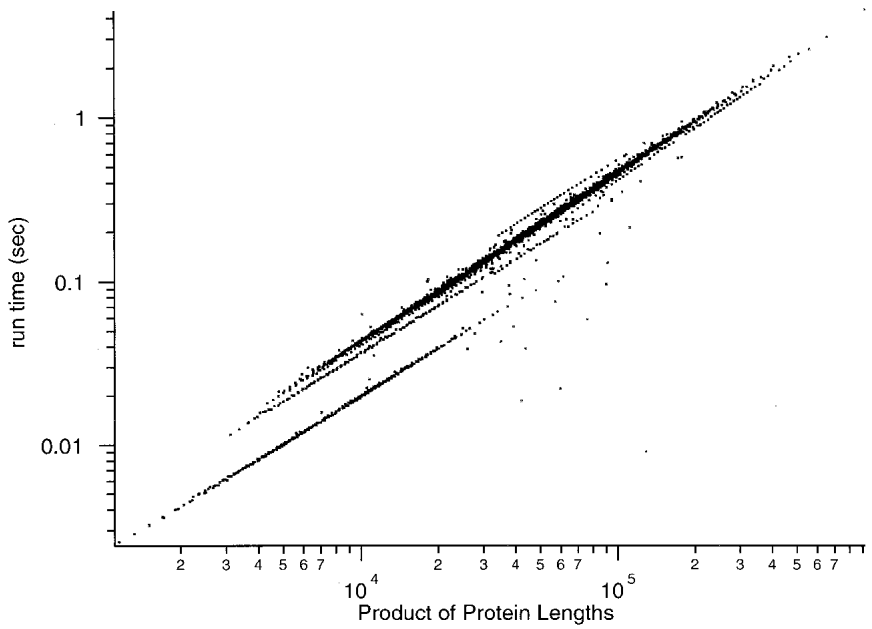

Fig. 2. Running time as a function of problem size. In the *x* axis, the product of the length of the two sequences being compared is shown, whereas in the *y* axis, the structural alignment time in seconds is plotted.

Statistical significance of MAMMOTH scores

An all-against-all comparison of different protein folds (Table 1A, Appendix) was carried out with MAMMOTH. The set of different folds compared was selected from the SCOP database as described in Materials and Methods. Figure 3 summarizes the results of this calculation as a plot of the relationship between length of the shortest protein being compared and percentage of structural identity (see Materials and Methods) after optimal structural fitting. The distribution of points in Figure 3 follows the familiar exponential decay observed by Sander and Schneider (1991) and Abagyan and Batalov (1997) in alignments of structurally unrelated sequences, suggesting a similar law for the background distribution of random structural alignments with MAMMOTH.

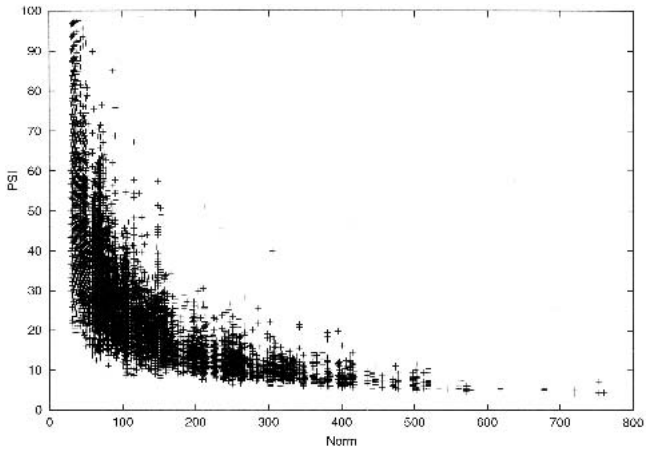

Fig. 3. Background distribution of random structural alignments. The percentage of structural similarity (*PSI*) after superimposing with MAMMOTH pairs of protein structures with different folds (see Materials and Methods and Table 1A in the appendix) is plotted as a function of the length of the shortest protein (Norm) being compared. All pairs of proteins in Table 1A are compared in the figure.

Thus, the raw data of percentage of superimposed residues (Fig. 3) were used to fit an extreme value distribution (EVD; Gumbel 1958) using a procedure similar to that put forward by Abagyan and Batalov (1997) and described in Materials and Methods. Figure 4 shows examples of fitting accuracy at two different sequence length intervals, comparing the frequency histogram obtained from the data and the fitted EVD curve. Figure 5 shows the curve fitting of these parameters to a power law of the length of the shortest protein being compared. This allows us to obtain the analytical *P*-value only from the knowledge of the length of the shortest protein and the percentage of superimposed residues. In order to test the accuracy of this *P*-value, a second test of all-against-all structural alignments was carried out (see Materials and Methods and Table 2A, Appendix). This time the analytical probability was compared to the calculated probability using the test set. Figure 6 shows an excellent agreement between both curves in the most relevant interval, up to the 95% confidence level. MAMMOTH is able to detect 50% of the true fold relationships in SCOP at the 99% confidence level, and 60% of them at the 95% confidence level. These numbers are comparable to results obtained with other automatic structure comparison meth-

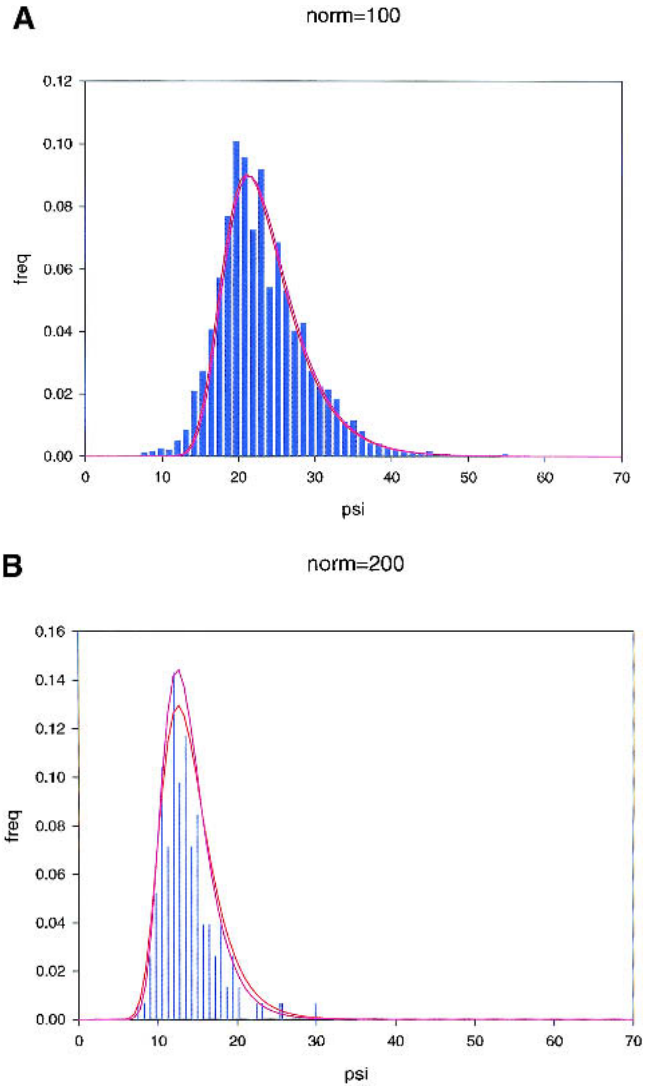

Fig. 4. Extreme value distribution (EVD) fit at different length intervals (Norm). In bars is the frequency histogram of *PSI* values; in red, the EVD curve using parameters derived from the frequency histogram; in magenta is the curve obtained using EVD parameters derived from a fitting to Norm (see text for details). (*A*) Norm = 100; (*B*) Norm = 200.

ods. For example, Yang and Honig (2000) reported 54% coverage at the 99% confidence level with PrISM. There are no published data regarding Dali. However, we have conducted similar tests using DaliLite (Holm and Park 2000), which indicated that DaliLite is able to detect 60% of SCOP relationships at the 99% confidence level, a slightly better performance, but similar to that obtained with MAMMOTH or PrISM. We conclude that MAMMOTH shows performance consistent with other structural alignment methods when comparing experimental protein structures, and that the *P*-value estimation provided by the EVD fitting is rather accurate.

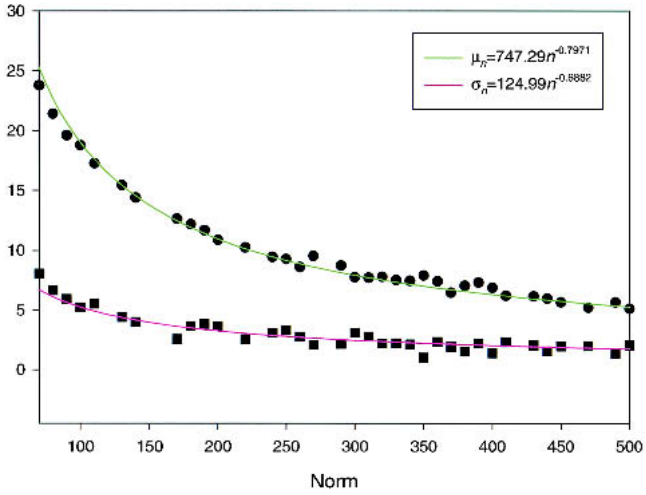

Fig. 5. Length-dependent estimate of EVD parameters. Parameters fitted at each sequence interval are in turn modeled as a function of the length of the shortest protein in the comparison.

Protein fold recognition with experimental structures

So far we have shown evidence that MAMMOTH partitions fold space in a way somewhat similar to that implicit in the SCOP database. We were also interested in testing the consistency and robustness of this partition, that is, the ability of the method to recognize entire families of members belonging to the same fold in SCOP. We selected fold families classified in SCOP with more than 15 members per family, and from each fold family we randomly picked one representative member, and then carried out comparisons with all other members of that fold family. We then studied how these families distribute as a function of MAMMOTH mean recognition ability (i.e., percentage of members above the

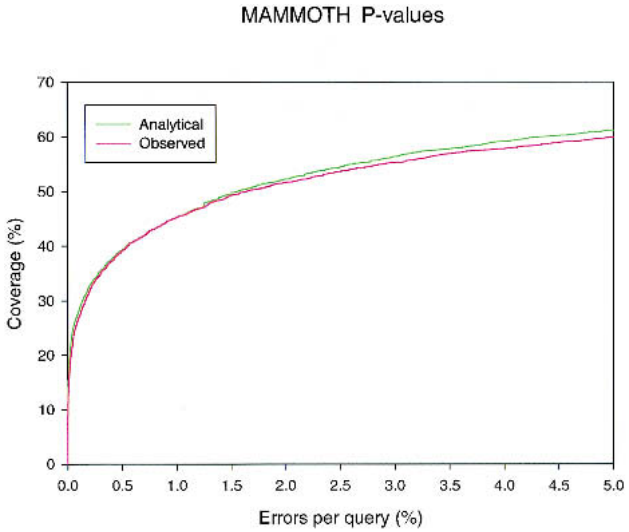

Fig. 6. Coverage-error plot for MAMMOTH scores. See text for details.

4.5 threshold score) and MAMMOTH mean scores (averages over all members in the family). Figure 7 shows the results in the form of a density contour plot, contoured at 0.01 density. The figure indicates that MAMMOTH scores provide consistent partitions: For most cases, more than 80% of members are recognized. And for those families with more than 80% of recognized members, the lowest level of the density curve is close to the boundary of $-\ln(P) = 4.5$ (P-value≈0.01), the threshold for statistical significance (99.0% confidence level). Thus, the cutoff for a statistically significant similarity is close to the boundary of mean similarity found among family members. This is another indication that, within MAMMOTH, the extent of fold space covered by members of each fold type is large, although highly variable depending on the specific group. A cutoff of $-\ln(P) = 4.5$ seems to be adequate to classify fold members together. In Figure 8, we have plotted the frequency of fold families as a function of the percentage of members recognized at this cutoff. Again, for most fold families, 80% or more of their members are recognized, and the proportion of false positives is small and evenly distributed (Fig. 8). Our view after these experiments is that the complete protein fold space appears to be quasi-discrete, with some overlap between different folds. That is, fold type attractors seem to be clearly defined in fold space, but the boundaries between some of the fold types are diffuse, populated by intermediate structures that may be indicative of evolutionary pathways. Other authors have arrived at similar conclusions through different analyses (Domingues et al. 2000; Yang and Honig 2000).

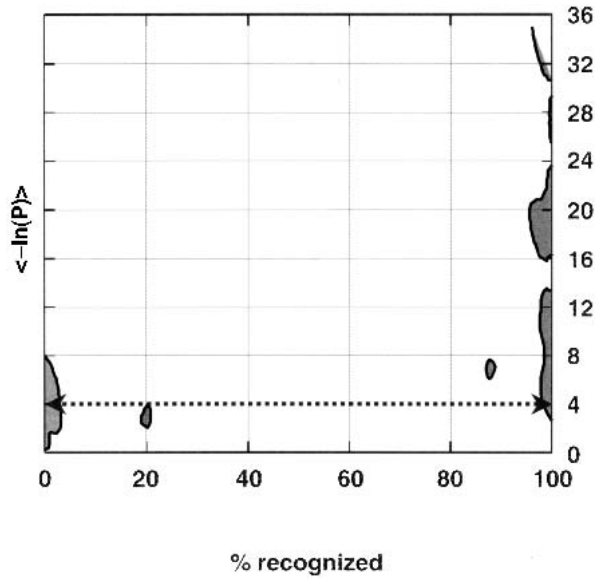

Fig. 7. Contour plot for family recognition. The percentage of family members recognized is plotted in the x axis; the y axis indicates the mean MAMMOTH score (−ln(P)) for that family. A density surface is contoured in the x–y plane using 0.015 as contouring threshold. See text for additional details.

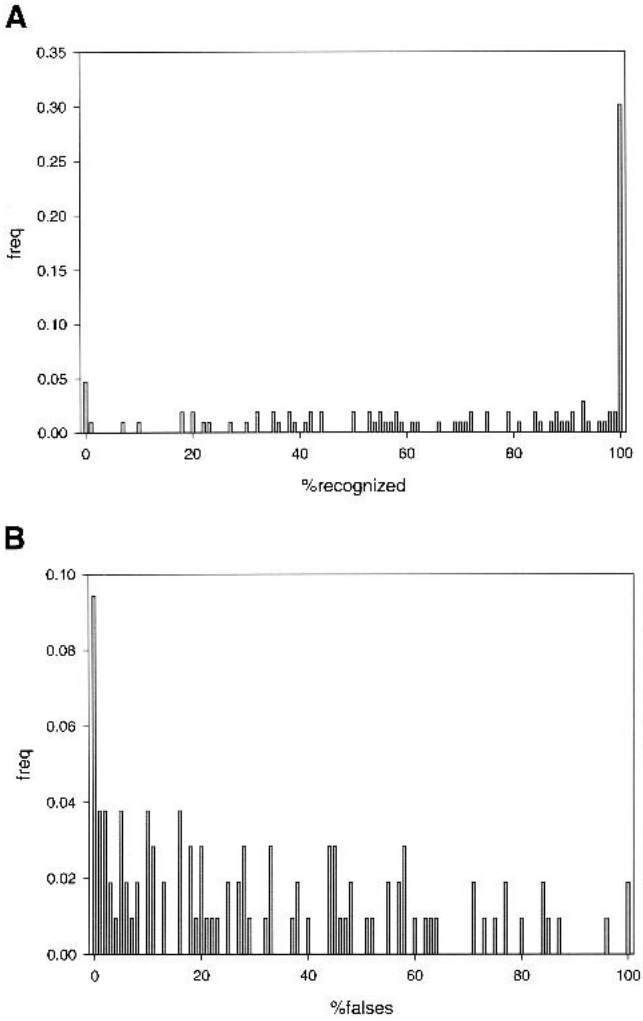

Fig. 8. Cumulative frequency of family recognition at the detection threshold. (*A*) Percentage of members recognized per family. (*B*) Percentage of false positives.

Finally, in Figure 9 we have plotted the mean fold family score as a function of the length of the representative protein family member (the length of the query protein described above). Fold families were separated, taking into account the average percentage of residues structurally aligned with that member in four classes. Protein families with an average percentage of aligned residues between $0 < PSI \leq 25$ are colored in red, giving regression equations

$$y = 0.0167x \text{ if } x \leq 200,\ y = 2.9996 \cdot 10^{-3}x + 2.74 \text{ if } x > 200.$$

Protein families with an average percentage of aligned residues in the interval $25 < PSI \leq 50$ for all family members are colored in green, giving regression lines

$$y = 0.0351x \text{ if } x \leq 200,\ y = 0.0226x + 2.5 \text{ if } x > 200.$$

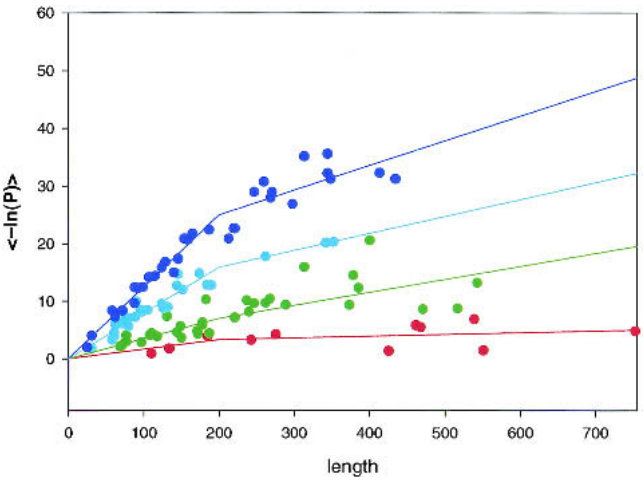

Fig. 9. Model quality using MAMMOTH scores. Each point is the mean *P*-value within each fold family as a function of the query protein length. Lines are a bilinear fitting using a cutoff at 200 residues ($x < 200$ and $x > 200$). Points correspond to individual families, and are colored as a function of PSI: red ($0 < \text{PSI} \leq 25$), green ($25 < \text{PSI} \leq 50$), cyan ($50 < \text{PSI} \leq 75$), blue ($75 < \text{PSI} \leq 100$).

Cyan represents families in the interval $50 < PSI \leq 75$, giving regressions

$$y = 0.0796x \text{ if } x \leq 200,\ y = 0.0294x + 10.4 \text{ if } x > 200.$$

Finally, protein families in the interval $75 < PSI \leq 100$ for all family members are colored in blue, with equations

$$y = 0.1252x \text{ if } x \leq 200,\ y = 0.0427x + 16.5 \text{ if } x > 200.$$

As expected, protein structures with low percentages of aligned residues have low MAMMOTH *P*-values. There is a bilinear dependency between MAMMOTH scores and protein length within each quality category, with a change in slope at the threshold of ~200 residues. The green line represents roughly the threshold for correct fold identification, and can be used to correctly assess a fold prediction taking protein size into account. Despite this dependency on protein size, it is interesting to note that, according to the regression equation, a typical pair of 150-residue proteins having about 50% of their residues aligned would have a score of 5.2, only slightly above the 4.5 threshold marking a statistically significant match. Again, this is an indication of the quasi-discrete distribution of protein structures in fold space.

Thus, as a summary, for a typical protein structure prediction in the 100–200-residue range, predictions with a score below ~4.00 can be considered definitely wrong. Predictions with a score between ~4.00 and ~5.25 on average are borderline, with some well predicted pieces in an overall wrong fold. Scores above ~5.25 are, on average, consistent with a correct fold prediction.

Benchmarking on CASP3 predicted models

Once the evaluation method has been described, we proceed to test it by comparing its performance with datasets of manual, consensus evaluations of predicted protein structures. In detail, evaluation performance was tested by comparing model rankings given by MAMMOTH *P*-values (more accurately the –ln(P) scores) with rankings produced by Murzin (1999) in his analysis of the fold recognition section in CASP3 (Moult et al. 1999). Figure 10 shows the relationship between the mean score by group given by Murzin and the mean score produced by MAMMOTH. Each point in the plot was obtained by computing the average over all models submitted by each different group participating in CASP3. MAMMOTH scores explain roughly 50% of the variance in the Murzin scores. Thus, there is a reasonably good correspondence between the mean MAMMOTH scores calculated within each predicting group participating in CASP3 and the mean Murzin evaluation score made by manual comparison. An interesting result to note from Figure 10 is the low value of the MAMMOTH mean scores, below 4.0 for most groups. Thus, the expected structural similarity between the models produced for most groups and the experimental structures is not much better than that the expected value obtained by randomly picking any pair of folds in the database. This is an important feature of the evaluation method: The scoring system is connected to our knowledge of protein structure. Figure 11 displays the results of another quality check with MAMMOTH. It shows all models submitted to CASP3 su-

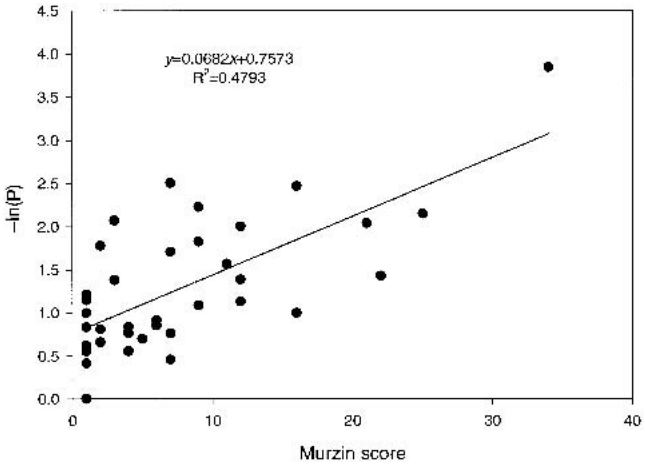

Fig. 10. Correlation between manual evaluation and automated scoring. Mean score by group given by Murzin against mean score produced by MAMMOTH. Each point is an average over all models submitted by each different group participating in CASP3.

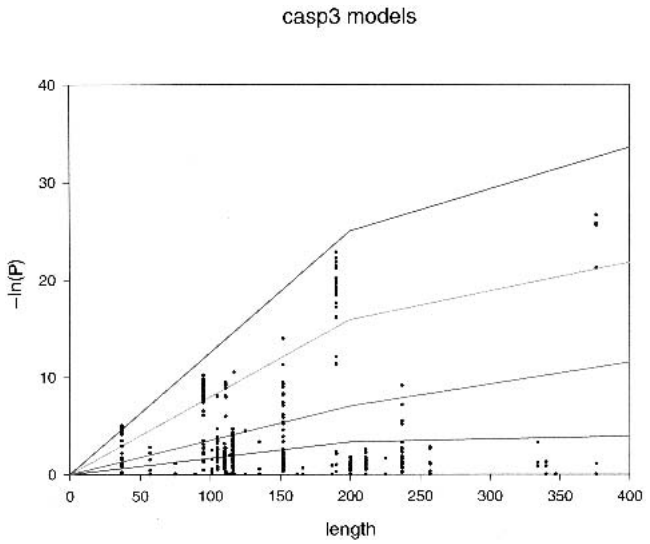

Fig. 11. Models submitted to CASP3 in the quality framework described in Figure 9. Each point is a model represented by the target length and the *P*-value obtained in the MAMMOTH superposition.

perimposed onto the trend lines derived in Figure 9. Only a few models with good quality were created.

Comparison with other methods in prediction evaluation

We also compared MAMMOTH with other previously proposed approaches to model evaluation (see Materials and Methods). Using the same set of predicted structures from CASP3, the Spearman's rank correlation coefficient was calculated between all pairs of different evaluation methods. We used rank correlation because of its inherent higher robustness (Langley 1970). From the rank correlation matrix we then derived the tree shown in Figure 12 by single linkage cluster analysis (Johnson and Wichern 1998) of the Spearman correlation coefficients. MAMMOTH is the evaluation method with its scoring scheme closest to Murzin's ranking, so that both of them have a similar behavior in comparison with the rest of the score systems. For example, scores computed within Dali are less similar to each other than Murzin's and MAMMOTH scores are between them.

It is important to take into account that Orengo made her evaluation of CASP3 models in a subset of all targets and groups evaluated by Murzin. In order to test whether there is a significant difference between both subsets, we performed a Wilcoxon's sum of ranks test (see Materials and Methods) using Murzin's, Dali, PrISM, and MAMMOTH scores (for which we had available both sets of numerical data). For all methods except Dali, differences in ranks are not significant, and can be explained by differences in sample size. This is not the case with Dali scores, however, although in this case ranking of correlations is still preserved. We conclude from this analysis that the tree shown in Figure 12 is robust and is not likely to change with an increase in sample size, although some of the branches could fluctuate to some extent, as seems to be the case for the Dali branch. All correlation coefficients used to build the tree on Figure 12 can be found in Table 2.

How can we explain the improvement in fold evaluation achieved by MAMMOTH? We have studied errors and successes of the different approaches to try to detect underlying patterns that could explain these differences, and will discuss some examples. Methods based on counting the number of fragments below a certain RMS threshold tend to fail, not surprisingly, when the predicted model is built from short fragments assembled in 3D. This is the case, for example, of some predictions for target t0071 in CASP3, where Group 217 made a threading model using structural fragments shorter than 25 residues. This model is ranked second both by Murzin and MAMMOTH; however, *Orengo-Lesk* failed to give it a high rank. We have also observed that assessment methods based on compatibility of 3D environments tend to fail if there are shifts in model registration, even in cases where the overall fold is preserved. For example, Group 5 submitted a threading model for target t0071 (Figure 13A), which is ranked in fifth place by Murzin and in fourth place by MAMMOTH. The structural alignment produced by MAMMOTH shows a considerable shift in registration, even though the overall fold is well reproduced. There are also problems associated with multidomain proteins, probably related to the way the similarity score is normalized. For example, the best model submitted for target t0071, according to Murzin's criteria, was ranked only fifth with PrISM. Finally, there are also con-

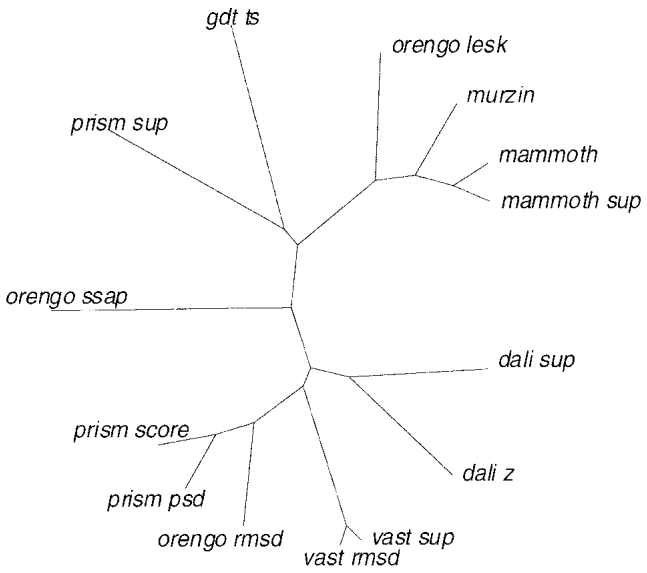

Fig. 12. Cluster analysis of the different evaluation methods. See text for details.

Table 2. *Correlation matrix between the different evaluation methods*

	Murzin	Dali Z	Dali sup	Mammoth	M-sup	O-Lesk	O-rmsd	Oreagoss	Vast sup	Vast rmsd	PrlSM sco	PrlSM sup	PrlSM psd	GDT-TS
Murzin	1	0.38	−0.19	0.84	0.78	0.6	0.53	0.01	0.53	0.53	0.07	0.34	0.08	0.49
Dali Z	0.38	1	0.6	−0.07	−0.06	0.06	0.19	0.14	0.52	0.52	0.53	−0.14	0.53	0.11
Dali sup	−0.19	0.6	1	−0.36	−0.41	−0.16	0.18	0.17	0.29	0.29	0.74	−0.16	0.52	0.26
MAMMOTH	0.84	−0.07	−0.36	1	0.91	0.54	0.56	−0.02	0.4	0.4	−0.13	0.23	−0.05	0.26
M-sup	0.78	−0.06	−0.41	0.91	1	0.72	0.41	0.06	0.41	0.41	−0.22	0.28	−0.12	0.17
O-Lesk	0.6	0.06	−0.16	0.54	0.72	1	0.23	0.46	0.24	0.24	−0.06	0.61	−0.09	0.42
O-rmsd	0.53	0.19	0.18	0.56	0.41	0.23	1	0.38	0.53	0.53	0.59	−0.04	0.72	0.07
O-ssap	0.01	0.14	0.17	−0.02	0.06	0.46	0.38	1	0.35	0.35	0.49	0.21	0.56	−0.16
V-sup	0.53	0.52	0.29	0.4	0.41	0.24	0.53	0.35	1	1	0.53	−0.23	0.52	0.05
V-rmsd	0.53	0.52	0.29	0.4	0.41	0.24	0.53	0.35	1	1	0.53	−0.23	0.52	0.05
P-score	0.07	0.53	0.74	−0.13	−0.22	−0.06	0.59	0.49	0.53	0.53	1	−0.009	0.88	−0.09
P-sup	0.34	0.14	0.16	0.23	0.28	0.61	−0.04	0.21	−0.23	−0.23	−0.009	1	−0.20	0.42
P-psd	0.08	0.53	0.52	−0.05	−0.12	−0.09	0.72	0.56	0.52	0.52	0.88	−0.20	1	0.07
GDT-TS	0.49	0.11	0.26	0.26	0.17	0.42	0.07	−0.16	0.05	0.05	−0.09	0.42	0.07	1

For each pair of evaluation scores, the Spearman correlation coefficient was computed using the data set of models shown in Table 3A of the Appendix.

siderable sources of error associated with distortions of secondary structure elements, particularly for ab initio models. Structure alignment programs designed to classify experimental protein structures, and not specifically to evaluate predictions, tend to suffer from artifacts arising from this. It is the case of the model submitted by Group 5 for target t0083 (Fig. 13B), considered by Murzin as the second-best model for this target. Whereas Dali did not find a significant structural similarity between target structure and model, MAMMOTH scored it with $-\ln(P) = 7.35$. Thus, the improvement achieved by MAMMOTH seems to be the result of a successful design to explicitly avoid some of these shortcomings.

Discussion

A new algorithm for protein structural alignment is described. As have other authors (Holm and Sander 1993; Madej et al. 1995; Gibrat et al. 1996; Shindyalov and Bourne 1998; Jung and Lee 2000; Lackner et al. 2000; Yang and Honig 2000), we resorted to the use of heuristics to cast the problem in a computationally tractable form. We divide the process into two steps: first, we compute the optimal similarity of the local backbone chain to establish residue correspondences between residues in both structures; in a second step, we then compute the largest subset of residues found within a given distance threshold in cartesian space. Insertions, deletions, and registration shifts between both structures are introduced in the first step. The approach is reminiscent of other structural alignment algorithms, although there are some clear differences. First, MAMMOTH uses unit-vector root mean square (Chew et al. 1999; Kedem et al. 1999) distances in the comparison of local structures, instead of the more widely used secondary structure elements. This is important when evaluating structure predictions because it avoids relying on secondary structure assignments, known to be very sensitive to the exact position of the backbone atomic coordinates (Labesse et al. 1997). It also allows the comparison of structures with a small percentage of defined secondary structure motifs, such as disulfide-rich small proteins, which cannot be handled by the more traditional methods. Second, the heuristic procedure used to search for the largest core with minimum RMS (Siew et al. 2000) is able to accelerate considerably the computation with respect to alternative approaches. The joint use of the above two features yields a fast, simple, deterministic, and yet completely general algorithm. This is demonstrated by the quality of the structural alignments and cores detected in difficult cases, with results comparable to other well known programs. Finally, the use of the EVD provides a rigorous score to evaluate structural alignments, particularly in structure prediction, as shown in the evaluation of CASP3 results.

In agreement with previous observations using percentage of sequence identity or sequence similarity with random sequence alignments, we show that the percentage of structural superimposition in random structural alignments also follows the well known EVD. This is not unexpected because a structural alignment, as well as its sequence counterpart, involves the optimization of a similarity score. The same distribution was reported by Levitt and Gerstein (1998) using a different metric to compute structural distances and a different optimization algorithm. A comparison between analytical and observed curves shows that MAMMOTH provides accurate estimates of the real *P*-values. On the other hand, the ability of MAMMOTH to reproduce SCOP fold classifications is similar to that of other available methods.

The structural comparison method described here has been successfully tested as an approach to evaluate models

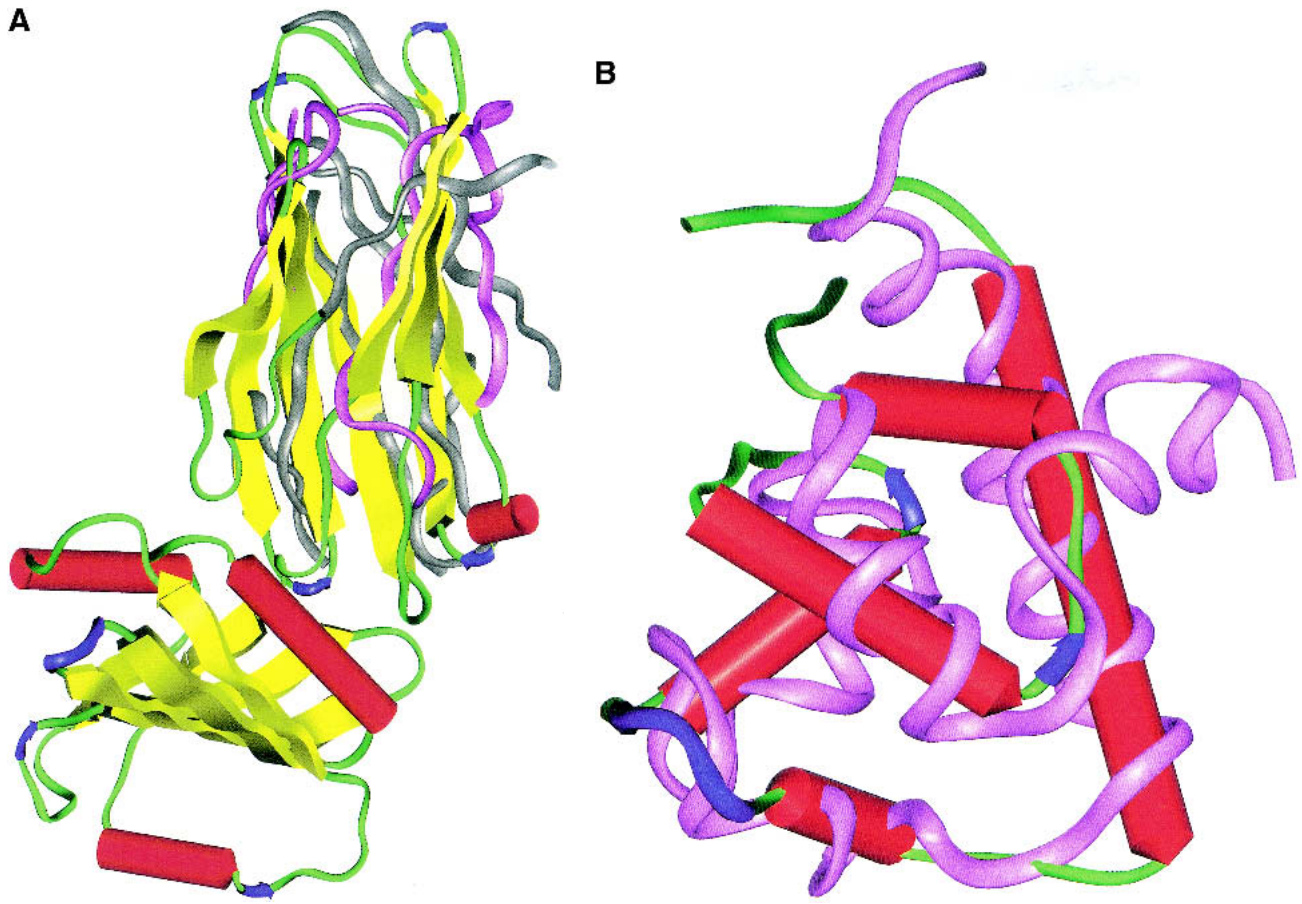

Fig. 13. Some typical "mistakes" in evaluation produced by other methods. The experimental structure is shown as a cartoon model. The matched portion of the theoretical model is shown in magenta, while the unmatched region is shown in gray. (*A*) t0071_g5; (*B*) t0083_g190.

generated by protein structure prediction methods. A comparison of different evaluation methods using the CASP3 benchmark indicates that MAMMOTH provides model quality rankings more consistent than those produced by other methods with the criteria provided by a human expert. It is instructive to compare the performance of different approaches when using experimental versus modeled structures. Although MAMMOTH, Dali, and PrISM, for example, show similar ability to recognize structural homologs based on experimental coordinates, there is a considerable difference when the objective is the comparison of modeled structures. In this case, MAMMOTH is considerably better than the other approaches. This highlights the fact that the problems involved in comparing modeled structures with their experimental counterparts and in comparing two experimental structures are different.

Due to its speed, insensitivity to differences in length, and rigorous evaluation score, MAMMOTH can be an important tool for protein structure comparison studies in structural genomics applications, particularly in those cases where partial or low-resolution models are of interest. For example, Baker and coworkers recently reported evidence that ab initio structure prediction followed by global structure comparison against the protein structure database can give insight into protein structure and function in cases where sequence-based methods alone fail (Simons et al. 2001). It can be reasonably expected that in the near future it will possible to apply this two-stage approach to small proteins at genomic scale. The good performance shown by MAMMOTH in this work makes it an ideal tool for the second part of this protocol, and recent results support this conclusion (Bonneau et al. 2002).

Additionally, MAMMOTH seems to be an adequate tool to be used in more fundamental studies of protein structure. For example, it allows finding and classifying, in a general way, recurrent structural motifs present in protein structures. These motifs are possibly responsible for the quasi-discreteness of fold space described by us in this paper and by others before us (Domingues et al. 2000). There is considerable interest in the structural biology community to derive a full inventory of these structural building blocks, and several approaches to the subject have already been

made (Holm and Sander 1998; Kleywegt 1999; Shindyalov and Bourne 2000; Reddy et al. 2001). Likewise, the ability of MAMMOTH to detect structural similarities using query substructures or building blocks can be of interest in approaches aimed at fitting models to electron density maps using databases of known protein structures (Diller et al. 1999a,b; Perrakis et al. 1999; Lamzin and Perrakis 2000; Jiang et al. 2001).

Finally, the high formal correspondence of MAMMOTH program structure to sequence alignment programs suggests that it should be straightforward to develop multiple structure alignment algorithms using MAMMOTH as a starting point. Several groups are actively addressing the problem of multiple structural alignment (Guda et al. 2001; Leibowitz et al. 2001a,b). With the current increase in the mean number of homologous protein structures in the database, it is important to develop more efficient algorithms for this problem. Work is in progress along these directions.

Materials and methods

MAMMOTH algorithm

The evaluation method focuses on model coordinates, avoiding references to sequence or contact maps while allowing registration shifts and different resolution levels. The method considers only the modeled portion of the target structure, avoiding the need to model the complete chain of the target. In common with other researchers, we reduce the complexity of the problem by using a heuristic approach: We first find the structural alignment that provides the optimal local similarity of the protein backbone (i.e., optimal local structure similarity of the complete amino acid sequence of both proteins) and then try to find the maximum subset of residues below a predefined distance in 3D space. The method consists of four basic steps:

(1) From the Cα trace, compute the unit-vector root mean square (URMS) distance between all pairs of heptapeptides of both model and experimental structure (Kedem et al. 1999). This is a measure sensitive to local structure, originally suggested by Chew et al. (1999). Consider a protein as described by its sequence of α-carbons (Cα). For each successive pair of Cα atoms along the backbone chain, we can record the unit vector in the direction from Cα *i* to Cα *i+1*. We can then place all recorded unit vectors at the origin, so that the backbone is mapped into vectors in the unit sphere. The URMS distance between two protein segments *A* and *B* (heptapeptides in our case) can then be computed by determining the rotation matrix which minimizes the sum of the squared distances between the corresponding unit vectors, using standard techniques (McLachlan 1979). The square root of the resulting minimum sum is defined as the URMS distance between heptapeptides *A* and *B*. It has been shown that the URMS metric provides an efficient detection of substructure similarities in proteins (Chew et al. 1999; Kedem et al. 1999).

(2) Use the matrix derived in step 1 to find an alignment of local structures that maximizes the local similarity of both the model and the experimental structure. First, URMS values need to be transformed to similarity scores. This is accomplished by noting that, as discussed by Chew et al. (1999), the expected minimum URMS distance between two random sets of *n* unit vectors ($URMS^R$) is:

$$URMS^R = \sqrt{2.0 - \frac{2.84}{\sqrt{n}}} \quad (1)$$

Thus, from eq. (1) we can then compute a similarity score (S_{AB}) between any two heptapeptides A and B as:

$$S_{AB} = \frac{(URMS^R - URMS^{AB})}{URMS^R}\,\Delta(URMS^R, URMS^{AB}) \quad (2)$$

Here, $\Delta(URMS^R, URMS^{AB}) = 10$ if $URMS^R > URMS^{AB}$ and $\Delta(URMS^R, URMS^{AB}) = 0$ otherwise. Therefore, S_{AB} provides a similarity scale between 0 and 10. Entries S_{AB} are used to build the similarity matrix **S** obtained by comparing all possible heptapeptides in both proteins. Dynamic Programming is then applied to this similarity matrix in order to build an alignment of both structures on the basis of their backbone (local) similarity. This alignment is produced using a global alignment method with zero end gaps (Needleman and Wunsch 1970). Internal gaps are penalized using an affine gap penalty function of the form $g(k) = \alpha+\beta k$, where k is the number of gaps and α and β are the opening and extension penalties, respectively. Trial and error tests (see below) indicated that values of $\alpha = 7.00$ and $\beta = 0.45$ gave good results.

(3) Find the maximum subset of similar local structures that have their corresponding Cα *close in cartesian space. Close* is considered here as a distance less than or equal to 4.0 Å. The method to find this subset is a small variant of the heuristic MaxSub algorithm (Siew et al. 2000; http://www.cs.bgu.ac.il/~dfischer/MaxSub/). Once the algorithm converges, the percentage of structural identity *(PSI)* is computed, defined as the percentage of corresponding residues below 4.0 Å in 3D space, measured with respect to the shortest structure.

(4) Calculate the probability of obtaining the given proportion of aligned residues (with respect to the shortest protein model) by chance (*P*-value). The *P*-value estimation is based on extreme-value fitting of the scores resulting from random structural alignments, following the work of Abagyan and Batalov (1997). The Type-I extreme value distribution based on the largest extreme, also known as the Gumbel distribution, has the following general form for its probability density function (Gumbel 1958):

$$f(x) = \frac{1}{b} e^{\frac{-(x-a)}{b}} e^{-e^{\frac{-(x-a)}{b}}} \quad (3)$$

where a is the so-called location parameter and b is the scale parameter. We are interested in the probability of having a t value greater than x, $P(t > x)$. This value can be found by integrating equation (3) from t to infinity, yielding:

$$P(t > x) = \int_t^\infty f(x)dx = 1 - e^{-e^{\frac{-(x-a)}{b}}} \quad (4)$$

In order to apply eq. (4) we need parameters a and b. For their derivation it is more convenient to work with the probability of having a value t smaller than or equal to x:

$$P(t \le x) = e^{-e^{\frac{-(x-a)}{b}}} \quad (5)$$

Taking logarithms in eq. (5) and setting $Q(x) = P(t \le x)$ and $P(x) = P(t > x)$, we have $Q(x) + P(x) = 1$. Equation 5 can then be transformed to the following linear form:

$$\ln(-\ln(1 - P(x))) = \frac{a}{b} - \frac{1}{b}x \quad (6)$$

Parameters a and b can now be estimated from a linear fitting between x, the percentage of aligned residues (*PSI*) obtained from the structural alignment algorithm in step 3, and $\ln(-\ln(1-P(x)))$, where $P(x)$ is computed as an accumulated sum of the observed frequencies with values greater than x. The reason for using $P(x)$ instead of $Q(x)$ in eq. 6 is in order to give a larger weight to the tail of the distribution, which contains the most critical part of the curve. Once a and b are found, expected values for the mean μ and variance σ^2 can be derived using the method of moments, giving relationships:

$$\mu = a + \gamma b \quad (7)$$

$$\sigma^2 = \frac{\pi^2}{6} b^2 \quad (8)$$

where $\gamma \approx 0.5772$ is the well-known Euler-Mascheroni constant (Gumbel 1958). Introducing eqs. 10 and 11 in eq. 4, the P-value as a function of z-score is obtained:

$$P(Z > z) = 1 - e^{-e^{-\left(\frac{\pi}{\sqrt{6}} z + \gamma\right)}} \quad (9)$$

Parameter optimization

Several parameters are used within the program: the length of the peptide in the URMS calculation, the similarity score derived from the URMS computation, the gap opening and extension penalties, the maximal distance between Cα, and the a and b parameters in the EVD. With the exception of the gap penalties and a and b parameters, no exhaustive optimization has been carried out. Gap penalties were optimized using a grid-like search, once the rest of parameters were fixed. The a and b parameters have been discussed previously. For the rest, values were initially established in order to avoid an undesirable combinatorial explosion in parameter space, based on the following considerations: (1) Number of residues for local similarity: This number has to be large enough to consider the different types of secondary structure. Four residues are required to define a helix turn and a β-turn. Thus, this would be a lower bound. However, calculations with ideal secondary structures (data not shown) indicated that helices and turns are difficult to distinguish by the URMS value with only four residues. Adding flanking residues provides a window of six residues, able to distinguish β-turns and helices. A seven-residue window was found to be more appropriate, however, probably because it can consider a complete two-helix turn. We observed that larger values begin to flat correct alignment pathways in similarity matrices, and therefore selected a heptapeptide. (2) Random URMS score: This value is established analytically, on the basis of the expected random values, and simply scaled between 0 and 10. Therefore there are no parameters to fit. (3) Maximal distance between Cα: Based on the value used in the MaxSub algorithm, together with visual observation of the results. The original MaxSub algorithm uses 3.5 Å. When dealing with models, a slightly larger value of 4 Å was deemed necessary.

Computation of coverage-error plots

From the all-versus-all comparison, we compute the coverage-error plot applying a procedure similar to that described by Levitt and Gerstein (1998): (1) For each pair we determine its P-value as computed by eq. 12, and note whether the pair is a true-positive or a true-negative; (2) We sort all pairs by increasing P-value; (3) We count down the list from best to worst and at each point in the list we find out the number of false positives and from that, the observed P-value; (4) We also compute the fraction of true positives that are more significant than the threshold P-value; this number defines the coverage, which should be as large as possible. On the other hand, observed and calculated P-values should be as close as possible.

Comparison with other evaluation methods

In order to assess the relative performance of MAMMOTH, we compared the evaluation scores provided by this approach with a set of 11 different evaluation methods previously used in CASP for structure comparison and model evaluation. All methods were benchmarked against the assumed gold standard given by Murzin's manual ranking of models submitted to the CASP3 meeting (Murzin 1999). When assessing the merits of the different approaches discussed here, it is important to keep in mind that some of these algorithms were not specifically developed to compare predicted models with their corresponding experimental structures, but rather to compare and classify pairs of experimental structures. The following sets of automatic criteria for assessment of the different models were compared with that used by MAMMOTH:

A. During CASP3, Orengo (Orengo et al. 1999) evaluated the ab initio predictions using three different criteria: the amount of nonoverlapping segments of 25 residues with an RMS value of 4.0 Å (Lesk 1997, referenced here as *orengo-lesk*); the similarity of the structural environment at each residue position (Taylor and Orengo 1989; *orengo-ssap*); and the largest fragment with an RMS of 4.0 Å (Orengo et al. 1999; *orengo-rmsd*). All three measures were compared with the MAMMOTH score.

B. Dali (Holm and Sander 1993) is a well known program for protein structure comparison. The Dali Z-score has been frequently used in the evaluation of structural predictions (Ortiz et al. 1998b; Simons et al. 2001). We have studied here both the Z-score (*dali*) and the percentage of superimposed residues (*dali-sup*) provided by the DaliLite package (Holm and Park 2000).

C. Vast (Madej et al. 1995; Gibrat et al. 1996) is another automatic method frequently used for protein structural alignment. Vast scores were used in both CASP2 and CASP3 to evaluate predicted structures. Here we used as scores the RMSD of the structural alignment (*vast-rmsd*) and the percentage of superimposed residues (*vast-sup*).

D. PrISM (Yang and Honig 2000) is a recently reported multipurpose program for protein modeling that also evaluates structural relationships between protein structures by using a new measure of protein structural distance. We used as scores the protein structural distance (*prism-psd*); the secondary structure alignment score (*prism-score*) and the percentage of superimposed residues and calculated by PrISM.

E. Finally, the GDT method (Zemla et al. 1999) was also included. The score (*gdt-ts*) is obtained from the global distance test (Zemla et al. 1999). It takes into account the percentage of residues that can be found within a given distance threshold between model and target. The *gdt-ts* measure is an average of

percentages obtained at 1, 2, 4, and 8 Å and has been used in previous assessments of CASP results by the Zemla team.

Comparisons were restricted to groups and targets evaluated jointly both by Murzin and Orengo during CASP3. These models are a subset of all models evaluated by Murzin during CASP3. The set of models included in the evaluation is listed in Table 3A of the Appendix. In order to test whether this subset is representative enough of the results that could have been obtained by using all models evaluated by Murzin, we used the Wilcoxon's sum of ranks test (Langley 1970) using the Murzin, Dali, MAMMOTH, and PrISM scores (for which we had all scores for both sets). We then compared the set of Murzin evaluations (all models) with the set evaluated jointly by Orengo and Murzin (subset). Our null hypothesis was that there are no significant differences in score distribution between both sets of models, so that results of the subset are representative of the complete set in CASP3. The test is as follows (Langley 1970): First, the scores of both samples (Murzin's set and the Orengo-Murzin subset) are pooled together. Then, the combined set of scores is sorted, and for each measurement a rank value is assigned. The smallest rank total R is then defined as the smaller of the sum of ranks coming from each sample. If distributions come from different underlying populations, unequal rank totals are expected. The probability of getting unequal rank totals as a consequence of chance variation can then be determined from R. The significance of the smaller rank total is found by calculating the statistic z given by the equation:

$$z = \frac{2R - n_R(n_A + n_B + 1)}{\sqrt{\frac{n_A n_B (n_A + n_B + 1)}{3}}} \quad (10)$$

where n_R is the number of measurements in whichever sample possesses the smaller rank total. The z-statistic distributes normally under the null hypothesis, and therefore the significance of z can finally be calculated using a normal distribution (Langley 1970).

Selection of structural datasets

Fold set selected to compute the background random distribution (Table 1A, Appendix)

This set was used to fit the EVD and to obtain the *P*-value estimation. It comprises a set of different folds without significant sequence identity (25% cutoff in sequence identity), selected by combining the *pdb_select* list from Hobohm and coworkers (Hobohm et al. 1992; Hobohm and Sander 1994) with the SCOP database.

A test set selected to compute coverage error plots (Table 2A, Appendix)

In this test set we first selected a representative set of proteins of different folds as in the previous case, but in addition we incorporated for each fold a second representative.

Fold families

All fold families from SCOP with more than 15 members per family were selected. We were able to select families belonging to 115 different folds, with 22 of them from the all-α class, 24 from the all-β class, 20 from the α/β class, and 21 from the α+β class. The rest (18 folds) belongs to other classifications in SCOP.

Table 1A. *PDB ID of the set of protein strutures used in* P*-value parameter estimation*

12asA	16pk	1a12A	1a1iA	1a1x	1a1yI	1a28A	1a2xB	1a34A	1a3aD
1a49A	1a6dA	1a6m	1a73A	1a92A	1aa7A	1aaf	1aba	1abv	1abz
1ad2	1adn	1aep	1afoA	1afp	1agrH	1ah7	1aho	1ahsA	1aie
1aihB	1ail	1aiw	1ajj	1ajsA	1ak4C	1ako	1am9A	1aohB	1aol
1ap8	1apj	1arb	1auz	1axn	1ay7B	1b0mA	1b0nB	1b0yA	1b10A
1b2nA	1b2pA	1b2vA	1b33N	1b3aA	1b4fB	1b4vA	1b5eA	1b63A	1b67A
1b77A	1b79A	1b7yB	1b93A	1b9qA	1ba4	1baq	1bb1C	1bd0A	1bd7A
1bdo	1be3A	1beg	1behB	1benB	1bfg	1bgf	1bk0	1bk7A	1bkf
1bkrA	1ble	1bm0A	1bm4A	1bm8	1bmqA	1bmtA	1bor	1bouA	1bp7A
1bs4A	1bsmA	1bvp1	1bwzA	1bx4A	1bxaA	1by1A	1byqA	1byrA	1c06A
1c24A	1c25	1c3mA	1c4zA	1c8zA	1cbn	1cby	1cc8A	1ce4A	1cem
1cex	1cfe	1cfh	1chd	1cipA	1cjdB	1clh	1cm5A	1cnzA	1co4A
1colA	1cozA	1cq3A	1csh	1csn	1ctf	1ctqA	1cw5A	1cwxA	1cy5A
1cyo	1czfA	1d0bA	1d1rA	1d2eA	1d3bE	1d3vA	1d6gA	1d7pM	1dciA
1dd5A	1dd8A	1devB	1dfeA	1dfnA	1dfuP	1dgi2	1dhn	1dj8A	1dlxA
1dmzA	1dozA	1dpsD	1dptA	1dxgA	1eaiC	1ebdC	1eciB	1ecmA	1ecrA
1efvB	1ehs	1euhA	1evhA	1extB	1fc2C	1fct	1fipA	1fltV	1fre
1fua	1furA	1fvkA	1fvl	1gdoB	1guxB	1gw4	1gyfA	1h2rL	1h2rS
1hcrA	1hfc	1hfeL	1hfeS	1hfi	1hiwS	1hmjA	1hoe	1hpy	1hqi
1htrP	1huuA	1hwtH	1hxn	1icfI	1ido	1ifc	1iibA	1iieA	1ixh
1jdw	1jhgA	1jsuC	1kapP	1kdxA	1khmA	1kifA	1kjs	1knyA	1kpf
1kptA	1latA	1lba	1lbeB	1lghB	1lkkA	1lktA	1lxtA	1lyp	1mabA
1mfmA	1mhdA	1mkaA	1mknA	1mml	1mnmA	1mof	1moq	1mrj	1mroB
1msc	1msi	1msk	1mtyG	1mun	1mut	1nkd	1nkl	1nls	1nox
1npoC	1nre	1nseA	1nxb	1oaa	1opd	1orc	1ospO	1othA	1ovaA
1p35B	1pcfA	1pdo	1peh	1pfo	1phf	1pmc	1pne	1poa	1poiA
1ppn	1pprM	1ppt	1psm	1ptq	1pty	1puc	1pyaA	1qa5A	1qauA
1qb2A	1qb7A	1qbhA	1qc7A	1qddA	1qdlB	1qexA	1qfqB	1qh4A	1qh5A
1qh8A	1qhfA	1qhkA	1qhvA	1qipA	1qj2C	1qj4A	1qkfA	1qklA	1qlaC
1qlmA	1qovH	1qoyA	1qp6A	1qq8A	1qqhA	1qqtA	1qr0A	1qreA	1qrvB
1qs1A	1qsaA	1qstA	1qvaA	1r2aA	1ra9	1rgeA	1rie	1rl6A	1roo
1rsy	1rvv1	1rzl	1seiA	1sfp	1sgpI	1sknP	1smpI	1stu	1svfA
1t1dA	1tac	1tbaA	1tfe	1thv	1tkaA	1t12A	1tml	1tmzA	1tpn
1trlA	1ttbA	1tx4A	1u9aA	1uae	1ubpA	1uby	1ugiD	1unkA	1uteA
1utg	1vcc	1vfyA	1vhh	1vid	1vii	1vns	1vpc	1vpu	1whi
1xxaB	1yacA	1yagG	1ycc	1ycqA	1ytbA	1ytfB	1ytfC	1yveI	1zpdA
1zto	256bA	2a3dA	2abd	2bbkL	2bpr	2cba	2chsA	2cpgA	2end
2erl	2fmtA	2gp8A	2hgf	2hhmA	2hp8	2igd	2ilk	2jhbA	2lisA
2napA	2occH	2por	2pspA	2pth	2pvbA	2rn2	2sicI	2spcA	2tbd
2tnfA	2tysB	2vgh	3bbg	3btoA	3chbD	3cla	3cyr	3daaA	3ezmA
3fib	3gcc	3lzt	3pte	3pviA	3sil	3vub	4eugA	4mt2	4pah
4pgaA	5hpgA	5pti	6gsvA	6pfkA	7a3hA	7atjA	7rsa	8abp	8rucI

Table 2A. *PDB ID of the set of protein structures used in the computation of the coverage-error plot (Figure 6)*

119l	153l	19hcA	1a02F	1a02N	1a0aA	1a0fA	1a0tP	1a15B	1a17
1a1w	1a2pA	1a2zA	1a32	1a4iB	1a4mA	1a4pA	1a53	1a5r	1a65A
1a6bB	1a6s	1a79A	1a7vA	1a8e	1a8rA	1a9nA	1a9v	1ac6A	1acp
1adoA	1adr	1ads	1ae1B	1afrA	1ag4	1agjA	1ah1	1ah9	1aisB
1aj2	1akr	1al3	1alvA	1aly	1amf	1amp	1amx	1anf	1aocA
1aoiD	1aoiE	1aoiF	1aoiG	1aojA	1aoxA	1aoy	1aozA	1ap0	1apq
1apyA	1aq0A	1aq5A	1aqb	1aqe	1aqr	1aqzB	1arv	1ash	1atg
1atlA	1atx	1atzA	1aub	1auoA	1auq	1auuA	1avaC	1avgI	1avoA
1avpA	1avqA	1avyA	1awcB	1awj	1awz	1axh	1ayfA	1ayj	1ayoA
1azeA	1azh	1azsA	1b00A	1b0uA	1b16A	1b1cA	1b22A	1b2nB	1b33A
1b34A	1b35A	1b35C	1b3kA	1b3tA	1b3uA	1b5qA	1b5tB	1b64	1b66A
1b6cB	1b6e	1b71A	1b72B	1b7fA	1b87A	1b8oA	1b8wA	1b94A	1b9hA
1b9uA	1b9wA	1ba5	1babB	1baj	1bak	1bazA	1bb9	1bba	1bbhA
1bbpA	1bbzE	1bc8C	1bcpB	1bcpD	1bcpF	1bd3A	1bd8	1bdmB	1bdyA
1be1	1be3C	1be3D	1be3F	1be3G	1be3H	1be3J	1be9A	1bea	1bebA
1befA	1bf0	1bf4A	1bf6A	1bfrA	1bftA	1bg2	1bgvA	1bh9A	1bh9B
1bhe	1bhgA	1bhi	1bj7	1bk5A	1bkzA	1bl0A	1bl1	1bm9A	1bmy
1bnb	1bnxA	1bo4A	1boeA	1bouB	1bpoA	1bpyA	1bqcA	1bqhG	1bqk
1bqsA	1bqv	1br0A	1br9	1brt	1bs0A	1bs9	1bt7	1btn	1bu2A
1bu7A	1bueA	1bupA	1bvb	1bvoA	1bvyF	1bw4	1bw5	1bw6A	1bw9A
1bx7	1bxwA	1bxyA	1byb	1byeA	1byfA	1bykA	1byla	1bywA	1bzbA
1bzg	1c02A	1c0aA	1c1lA	1c1yB	1c20A	1c2aA	1c3d	1c3qA	1c3zA
1c4eA	1c52	1cb8A	1cc5	1ccvA	1cczA	1cd1A	1cdcA	1ce0A	1cewI
1cf7A	1cf7B	1cf9A	1cfr	1cg2A	1cg8B	1cgo	1ch6A	1chmA	1ci3M
1cjgA	1cjwA	1ck9A	1ckaA	1ckmA	1cktA	1cl4A	1cl8A	1cliA	1cmbA
1cmoA	1cmr	1cnoA	1cnt3	1cnv	1cokA	1coo	1couA	1cp2A	1cpq
1cpzA	1cqkA	1cqqA	1cqyA	1crb	1cseI	1csyA	1cunA	1cur	1cv8
1cxc	1cxzB	1cydA	1cyx	1cz1A	1cz4A	1czpA	1d1dA	1d2fA	1d2nA
1d2zA	1d2zB	1d3bL	1d3cA	1d4aA	1d4bA	1d4vB	1d66A	1d7uA	1d8bA
1d9cB	1d9yA	1dabA	1dbgA	1dbwB	1dcqA	1ddf	1dec	1delA	1dfjI
1dfoB	1dgvA	1dgwA	1dgwX	1dhr	1diiC	1dipA	1dlfH	1dlfL	1dmd
1dmuA	1dokA	1dosA	1dpe	1dpgA	1dun	1dupA	1dynA	1e2aA	1eal
1eayC	1ecd	1eceA	1ecpA	1edg	1edmB	1edt	1eerA	1eerB	1egaA
1egeA	1egr	1eh2	1eit	1epaA	1erd	1erp	1erv	1ery	1esc
1eur	1exh	1fadA	1fas	1faxL	1fcdA	1fcdC	1fce	1fdm	1fgjA
1fleI	1flp	1fltY	1fmb	1fna	1fppB	1fsb	1ft1A	1ftrA	1fus
1fvpA	1fxd	1fzcA	1g31A	1gca	1gcf	1gd1O	1gdhA	1ghk	1gks
1gky	1gllO	1gotB	1gotG	1gp2G	1gpc	1gpeA	1gph1	1gpr	1gpt
1gpx	1gseA	1gvp	1hce	1hcnA	1hcnB	1hdj	1hkbA	1hlb	1hloB
1hmt	1hnr	1hrzA	1hsbB	1hstA	1htp	1hulA	1hymA	1hyp	1i16
1iab	1iakA	1iakB	1ica	1idaA	1ihfB	1ihvA	1ilr2	1isuA	1itbB
1ithA	1ixxB	1jer	1jetA	1jfrA	1jkmA	1jli	1jlyA	1jmcA	1jrhI
1junA	1jvr	1kacB	1kb5B	1kevA	1kigI	1koe	1kp6A	1kte	1kuh
1kwaA	1kzuB	1lab	1lcl	1lea	1lfb	1lfdA	1lghA	1lki	1lmb4
1louA	1lpbA	1lrv	1lxa	1mai	1mba	1mdc	1mdyA	1mek	1mfiA
1mgtA	1mhlA	1mjc	1mnmC	1molA	1mpgA	1mpyA	1mroC	1msfC	1mspA
1mtx	1mtyB	1mtyD	1mup	1nar	1nbcA	1ncs	1ncu	1nddB	1ndoB
1neb	1neq	1nfiE	1nglA	1ngr	1nksA	1nmtA	1np4A	1npk	1nsj
1nwpA	1nzyB	1obpA	1obr	1obwA	1ofgA	1ohk	1omb	1onc	1opc
1opr	1ordA	1osa	1otcA	1otcB	1otfA	1otgA	1ounA	1oxa	1pba
1pbv	1pdc	1pdkB	1pea	1pfiA	1pfsA	1pft	1phnA	1pht	1picA
1pij	1pjcA	1plc	1pls	1pluA	1pmi	1pnbA	1pnkA	1ponB	1pot
1psf	1psrA	1pud	1pytA	1qa6A	1qa7A	1qaxB	1qbjC	1qcxA	1qd1A
1qeyA	1qf9A	1qfdA	1qfhA	1qg3A	1qghA	1qgiA	1qgoA	1qj8A	1qjdA
1qjkA	1qjoA	1qk6A	1qk7A	1qk9A	1qksA	1qlaB	1qlbA	1qleD	1qloA
1qmhB	1qo0E	1qo3D	1qo6A	1qorA	1qpoA	1qq2A	1qq4A	1qqp1	1qqp2
1qqp3	1qqrB	1qr4A	1qrjB	1qrqA	1qrrA	1qsdA	1qsfD	1qsgA	1qsmA
1qtsA	1qtwA	1qu0C	1quqB	1quuA	1qyp	1r1bA	1r2fB	1r69	1rcb
1rcf	1regY	1res	1rip	1rkd	1rlw	1rmg	1rmvA	1rrpB	1sacA
1sbp	1scjB	1scy	1sek	1semB	1sfcD	1sfcF	1sfcI	1sfcK	1shcA
1sis	1skf	1sltB	1smeA	1smtA	1spf	1sra	1sro	1srrA	1stfI
1stmA	1sur	1svfB	1svpA	1tafA	1tafB	1taxA	1tc3C	1tca	1ten
1tf4A	1tfi	1tgoA	1tgsI	1tgxA	1tib	1tif	1tig	1tiiD	1tit
1tkiA	1tlfA	1tmy	1tnt	1toaA	1tph1	1tsg	1tsk	1tud	1tul
1tupA	1tvdA	1tvxB	1tyfA	1u2fA	1ubpB	1urnA	1uroA	1vcaA	1vcbA
1vcbC	1vdfA	1vie	1vih	1vls	1vpsB	1vsrA	1vtx	1wab	1wapB
1wdcB	1wdnA	1who	1whtA	1wit	1wkt	1wwcA	1xbrB	1xdtR	1xikA
1xnb	1xsm	1xvaA	1yagA	1yaiC	1yprA	1ystL	1yujA	1zagB	1zap
1zaq	1zfd	1zmeC	2a0b	2a93A	2a93B	2aaiB	2abk	2acy	2adx
2af8	2afpA	2ahjD	2alcA	2arcB	2aw0	2ayh	2baa	2bb8	2bbkH
2bby	2bds	2bi6H	2bjxA	2bopA	2bosA	2bpa1	2bpa2	2brz	2cblA
2ccyA	2cpkE	2ctc	2cyp	2dorA	2dri	2drpD	2eboA	2ercA	2ezi
2ezl	2ezzA	2fapB	2fcpA	2fivA	2fmr	2fnbA	2fxb	2gdm	2gmfA
2gsaA	2hbg	2hddB	2hfh	2hmzA	2hsp	2i1b	2ifo	2irfG	2itg
2kaiA	2knt	2lefA	2liv	2mcm	2msbB	2mtaC	2myr	2nacA	2nbtA
2ncm	2new	2nlrA	2nmbA	2nsyA	2oatA	2occA	2occC	2occD	2occE
2occF	2occG	2occI	2occJ	2occK	2occL	2occM	2omf	2pii	2plc
2ptl	2qwc	2sak	2sas	2scpA	2scuB	2sh1	2sn3	2sns	2stv
2tbvC	2tgi	2tmdA	2tpsA	2trxA	2u1a	2u2fA	2xbd	3bamA	3chy
3crd	3dapA	3dfr	3gcb	3lhbA	3ncmA	3pcgA	3pvaA	3sdhA	3stdA
3tdt	3thiA	3tmkD	451c	4cpaI	4crxA	4kbpA	4sbvC	4sgbI	4tgf
4xis	5hck	5icb	5nul	5ptd	5tmpA	6cel	6mhtA	6paxA	6prcC
7fd1A	8prkA	9wgaA							

Dataset of predicted models (Table 3A, Appendix)

Models were downloaded from the CASP web site at: http://predictioncenter.llnl.gov/casp3/Casp3.html.

Acknowledgements

Mount Sinai School of Medicine start-up funds are acknowledged. We thank Fabien Champagne, Carlos Pérez, and Dmitry Lupyan

Table 3A. *Set of CASP3 models used to compare the different evaluation methods.*

T0043AL009_1	T0043AL033_1	T0043AL040_1_1	T0043AL061_1	T0043AL066_1	T0043AL074_1	T0043AL076_1	T0043AL090_1
T0043AL143_1	T0043AL166_1	T0043AL168_1	T0043AL176_1	T0043AL201_1	T0043TS005_1	T0043TS009_1	T0043TS023_1
T0043TS035_1	T0043TS045_1	T0043TS061_1	T0043TS105_1	T0043TS156_1	T0043TS217_1	T0044AL009_1	T0044AL017_1
T0044AL019_1	T0044AL028_1	T0044AL033_1	T0044AL040_1	T0044AL061_1_1	T0044AL074_1	T0044AL076_1	T0044AL156_1
T0044AL168_1	T0044AL176_1	T0044AL201_1	T0044TS005_1	T0044TS009_1	T0044TS045_1	T0044TS061_1_1	T0044TS136_1
T0044TS217_1	T0044TS224_1	T0053AL003_1	T0053AL009_1	T0053AL017_1	T0053AL019_1_1	T0053AL028_1	T0053AL033_1
T0053AL061_1	T0053AL074_1	T0053AL076_1_1	T0053AL090_1	T0053AL147_1	T0053AL162_1	T0053AL166_1	T0053AL168_1
T0053AL176_1	T0053AL201_1	T0053AL212_1	T0053AL273_1	T0053TS005_1	T0053TS009_1	T0053TS045_1	T0053TS061_1
T0053TS072_1	T0053TS217_1	T0053TS224_1	T0056AL003_1	T0056AL009_1	T0056AL017_1	T0056AL019_1	T0056AL028_1
T0056AL040_1	T0056AL076_1	T0056AL085_1	T0056AL142_1	T0056AL147_1	T0056AL166_1	T0056AL168_1	T0056AL201_1
T0056AL273_1	T0056TS005_1	T0056TS009_1	T0056TS023_1	T0056TS035_1	T0056TS061_1	T0056TS072_1	T0056TS105_1
T0056TS163_1	T0056TS185_1	T0056TS190_1	T0056TS217_1	T0057AL009_1	T0057AL017_1	T0057AL019_1	T0057AL061_1
T0057AL066_1_1	T0057AL076_1	T0057AL168_1	T0057AL201_1	T0057AL222_1	T0057AL273_1	T0057TS009_1	T0057TS061_1
T0057TS072_1	T0057TS136_1	T0057TS217_1	T0059AL003_1	T0059AL009_1	T0059AL019_1	T0059AL028_1	T0059AL040_1
T0059AL061_1	T0059AL066_1	T0059AL074_1	T0059AL076_1	T0059AL085_1	T0059AL090_1	T0059AL142_1	T0059AL162_1
T0059AL166_1	T0059AL168_1	T0059AL201_1	T0059AL222_1	T0059AL273_1	T0059TS005_1	T0059TS009_1	T0059TS035_1
T0059TS045_1	T0059TS053_1	T0059TS061_1	T0059TS072_1	T0059TS163_1	T0059TS190_1	T0059TS217_1	T0059TS224_1
T0061AL003_1	T0061AL009_1	T0061AL019_1	T0061AL028_1	T0061AL033_1	T0061AL040_1	T0061AL061_1	T0061AL076_1
T0061AL168_1	T0061AL176_1	T0061AL201_1	T0061AL273_1	T0061TS009_1	T0061TS035_1	T0061TS045_1	T0061TS061_1
T0061TS072_1	T0061TS105_1	T0061TS185_1	T0061TS190_1	T0061TS217_1	T0061TS224_1	T0063AL009_1	T0063AL019_1
T0063AL028_1	T0063AL040_1	T0063AL061_1	T0063AL066_1_1	T0063AL074_1	T0063AL090_1	T0063AL143_1	T0063AL147_1
T0063AL156_1	T0063AL168_1	T0063AL176_1	T0063AL201_1	T0063AL212_1	T0063TS005_1	T0063TS009_1	T0063TS035_1
T0063TS045_1	T0063TS061_1	T0063TS072_1	T0063TS163_1	T0063TS224_1	T0067AL009_1	T0067AL017_1	T0067AL019_1
T0067AL028_1	T0067AL033_1	T0067AL040_1	T0067AL061_1	T0067AL066_1	T0067AL074_1	T0067AL085_1	T0067AL142_1
T0067AL147_1	T0067AL156_1	T0067AL166_1	T0067AL168_1	T0067AL273_1	T0067TS005_1	T0067TS009_1	T0067TS045_1
T0067TS061_1	T0067TS072_1	T0067TS105_1	T0067TS217_1	T0067TS224_1	T0068AL009_1	T0068AL017_1	T0068AL019_1
T0068AL028_1	T0068AL040_1	T0068AL061_1	T0068AL066_1	T0068AL076_1	T0068AL085_1	T0068AL090_1	T0068AL143_1
T0068AL147_1	T0068AL162_1	T0068AL166_1	T0068AL168_1	T0068AL176_1	T0068AL201_1	T0068AL212_1	T0068AL222_1
T0068TS005_1	T0068TS009_1	T0068TS028_1	T0068TS061_1	T0068TS072_1	T0068TS074_1	T0068TS217_1	T0068TS224_1
T0071AL003_1	T0071AL009_1	T0071AL019_1_1	T0071AL040_1	T0071AL066_1_1	T0071AL090_1	T0071AL142_1	T0071AL147_1
T0071AL162_1	T0071AL166_1	T0071AL168_1	T0071AL212_1	T0071AL273_1	T0071TS005_1_1	T0071TS009_1	T0071TS035_1
T0071TS045_1_1	T0071TS217_1	T0071TS224_1	T0075AL019_1	T0075AL028_1	T0075AL040_1	T0075AL061_1	T0075AL066_1
T0075AL076_1	T0075AL143_1	T0075AL147_1	T0075AL156_1	T0075AL162_1	T0075AL166_1	T0075AL168_1	T0075AL201_1
T0075AL273_1	T0075TS005_1	T0075TS035_1	T0075TS045_1	T0075TS061_1	T0075TS072_1	T0075TS156_1	T0075TS163_1
T0075TS190_1	T0075TS217_1	T0075TS224_1	T0077AL009_1	T0077AL017_1	T0077AL019_1_1	T0077AL028_1	T0077AL033_1
T0077AL040_1	T0077AL061_1	T0077AL074_1	T0077AL142_1	T0077AL143_1	T0077AL147_1	T0077AL156_1	T0077AL166_1
T0077AL168_1	T0077AL176_1	T0077AL201_1	T0077AL212_1	T0077AL273_1	T0077TS005_1	T0077TS009_1	T0077TS035_1
T0077TS045_1	T0077TS061_1	T0077TS072_1	T0077TS105_1	T0077TS163_1	T0077TS190_1	T0077TS217_1	T0079AL009_1
T0079AL017_1	T0079AL019_1_1	T0079AL028_1	T0079AL040_1	T0079AL061_1	T0079AL066_1	T0079AL076_1	T0079AL085_1
T0079AL090_1	T0079AL156_1_1	T0079AL166_1	T0079AL168_1	T0079AL176_1	T0079AL201_1	T0079AL212_1	T0079AL222_1
T0079AL273_1	T0079TS005_1	T0079TS009_1	T0079TS023_1	T0079TS035_1	T0079TS045_1	T0079TS061_1	T0079TS072_1
T0079TS105_1	T0079TS156_1	T0079TS163_1	T0079TS185_1	T0079TS190_1	T0079TS217_1	T0079TS224_1	T0080AL009_1
T0080AL017_1	T0080AL019_1_1	T0080AL028_1	T0080AL033_1	T0080AL061_1	T0080AL074_1	T0080AL076_1	T0080AL085_1
T0080AL090_1	T0080AL147_1	T0080AL168_1	T0080AL201_1	T0080AL273_1	T0080TS009_1	T0080TS045_1	T0080TS061_1
T0081AL009_1	T0081AL019_1	T0081AL028_1	T0081AL040_1	T0081AL061_1	T0081AL074_1	T0081AL076_1	T0081AL090_1
T0081AL166_1	T0081AL176_1	T0081AL201_1	T0081AL212_1	T0081AL273_1	T0081TS005_1	T0081TS009_1	T0081TS035_1
T0081TS045_1	T0081TS061_1	T0081TS105_1	T0081TS217_1	T0081TS224_1	T0083AL009_1	T0083AL017_1	T0083AL019_1
T0083AL028_1	T0083AL033_1	T0083AL040_1	T0083AL061_1	T0083AL176_1	T0083AL201_1	T0083AL273_1	T0083TS005_1
T0083TS009_1	T0083TS035_1	T0083TS045_1	T0083TS061_1	T0083TS072_1	T0083TS190_1	T0083TS224_1	T0085AL009_1
T0085AL017_1	T0085AL019_1_1	T0085AL028_1	T0085AL061_1_1	T0085AL066_1_1	T0085AL076_1	T0085AL085_1	T0085AL142_1
T0085AL162_1	T0085AL166_1	T0085AL168_1_1	T0085AL176_1	T0085AL222_1	T0085AL257_1	T0085AL273_1	T0085TS005_1
T0085TS009_1	T0085TS045_1	T0085TS061_1_1	T0085TS085_1	T0085TS105_1			

Models are available at http://predictioncenter.11n1.gov/casp3/Casp3.html

for their help in setting up the MAMMOTH server, and Federico Gago for carefully reading the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Abagyan, R. and Batalov, S. 1997. Do aligned sequences share the same fold? *J. Mol. Biol.* **273:** 355–368.

Adams, P.D. and Grosse-Kunstleve, R.W. 2000. Recent developments in software for the automation of crystallographic macromolecular structure determination. *Curr. Opin. Struct. Biol.* **10:** 564–568.

Al-Hashimi, H.M. and Patel, D.J. 2002. Residual dipolar couplings: Synergy between NMR and structural genomics. *J. Biomol. NMR* **22:** 1–8.

Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294:** 93–96.

Bonneau, B., Strauss, C., Rohl, C., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., Baker, D. 2002. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* **322:** 65.

Boutonnet, N.S., Rooman, M.J., Ochagavia, M.E., Richelle, J., and Wodak, S.J. 1995. Optimal protein structure alignments by multiple linkage clustering: Application to distantly related proteins. *Protein Eng.* **8:** 647–662.

Brenner, S.E. 2001. A tour of structural genomics. *Nat. Rev. Genet.* **2:** 801–809.

Burley, S.K. and Bonanno, J.B. 2002. Structural genomics of proteins from conserved biochemical pathways and processes. *Curr. Opin. Struct. Biol.* **12:** 383–391.

Chance, M.R., Bresnick, A.R., Burley, S.K., Jiang, J.S., Lima, C.D., Sali, A., Almo, S.C., Bonanno, J.B., Buglino, J.A., Boulton, S., et al. Structural genomics: A pipeline for providing structures for the biologist. *Protein Sci.* **11:** 723–738.

Chew, L., Huttlenlocher, D., Kedem, K., and Kleinberg, J. 1999. Fast detection of common geometric substructure in proteins. *J. Comput. Biol.* **6:** 313–325.

Diller, D.J., Pohl, E., Redinbo, M.R., Hovey, B.T., and Hol, W.G. 1999a. A rapid method for positioning small flexible molecules, nucleic acids, and large protein fragments in experimental electron density maps. *Proteins* **36:** 512–525.

Diller, D.J., Redinbo, M.R., Pohl, E., and Hol, W.G. 1999b. A database method for automated map interpretation in protein crystallography. *Proteins* **36:** 526–541.

Domingues, F.S., Koppensteiner, W.A., and Sippl, M.J. 2000. The role of protein structure in genomics. *FEBS Lett.* **476:** 98–102.

Fetrow, J.S. and Skolnick, J. 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **281:** 949–968.

Fetrow, J.S., Godzik, A., and Skolnick, J. 1998. Functional analysis of the Escherichia coli genome using the sequence- to-structure-to-function paradigm: Identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282:** 703–711.

Gibrat, J.F., Madej, T., and Bryant, S.H. 1996. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6:** 377–385.

Guda, C., Scheeff, E.D., Bourne, P.E., and Shindyalov, I.N. 2001. A new

algorithm for the alignment of multiple protein structures using Monte Carlo optimization. *Pac. Symp. Biocomput.* 275–286.
Gumbel, E. 1958. *Statistics of extremes*. Columbia University Press, New York.
Hobohm, U. and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Sci.* **3:** 522–524.
Hobohm, U., Scharf, M., Schneider, R., and Sander, C. 1992. Selection of representative protein data sets. *Protein Sci.* **1:** 409–417.
Holm, L. and Park, J. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics* **16:** 566–567.
Holm, L. and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233:** 123–138.
———1996. Mapping the protein universe. *Science* **273:** 595–603.
——— 1998. Dictionary of recurrent domains in protein structures. *Proteins* **33:** 88–96.
Hurley, J.H., Anderson, D.E., Beach, B., Canagarajah, B., Ho, Y.S., Jones, E., Miller, G., Misra, S., Pearson, M., Saidi, L., Suer, S., Trievel, R., and Tsujishita, Y. 2002. Structural genomics and signaling domains. *Trends Biochem. Sci.* **27:** 48–53.
Jiang, W., Baker, M.L., Ludtke, S.J., and Chiu, W. 2001. Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* **308:** 1033–1044.
Johnson, R. and Wichern, D. 1998. *Applied multivariate statistical analysis*. Prentice Hall, Upper Saddle City, New Jersey.
Jung, J. and Lee, B. 2000. Protein structure alignment using environmental profiles. *Protein Eng.* **13:** 535–543.
Kedem, K., Chew, L., and Elber, R. 1999. Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins* **37:** 554–564.
Kleywegt, G.J. 1999. Recognition of spatial motifs in protein structures. *J. Mol. Biol.* **285:** 1887–1897.
Labesse, G., Colloc'h, N., Pothier, J., and Mornon, J.P. 1997. P-SEA: A new efficient assignment of secondary structure from C α trace of proteins. *Comput. Appl. Biosci.* **13:** 291–295.
Lackner, P., Koppensteiner, W.A., Sippl, M.J., and Domingues, F.S. 2000. ProSup: A refined tool for protein structure alignment. *Protein Eng.* **13:** 745–752.
Lamzin, V.S. and Perrakis, A. 2000. Current state of automated crystallographic data analysis. *Nat. Struct. Biol.* **7:** 978–981.
Langley, R. 1970. *Practical statistics. Simply explained*. Dover, New York.
Leibowitz, N., Fligelman, Z.Y., Nussinov, R., and Wolfson, H.J. 2001a. Automated multiple structure alignment and detection of a common substructural motif. *Proteins* **43:** 235–245.
Leibowitz, N., Nussinov, R., and Wolfson, H.J. 2001b. MUSTA—A general, efficient, automated method for multiple structure alignment and detection of common motifs: Application to proteins. *J. Comput. Biol.* **8:** 93–121.
Lesk, A. 1997. CASP2: Report on ab initio predictions. *Proteins* **1:** 151–166.
Levitt, M. and Gerstein, M. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci.* **95:** 5913–5920.
Lichtarge, O. and Sowa, M.E. 2002. Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12:** 21–27.
Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., and Chothia, C. 2000. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **28:** 257–259.
Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316:** 139–154.
Madej, T., Gibrat, J.F., and Bryant, S.H. 1995. Threading a database of protein cores. *Proteins* **23:** 356–369.
McLachlan, A.D. 1979. Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128:** 49–79.
Mittl, P.R. and Grutter, M.G. 2001. Structural genomics: Opportunities and challenges. *Curr. Opin. Chem. Biol.* **5:** 402–408.
Moult, J., Hubbard, T., Fidelis, K., and Pedersen, J.T. 1999. Critical assessment of methods of protein structure prediction (CASP): Round III. *Proteins* **l:** 2–6.
Murzin, A.G. 1999. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins* **l:** 88–103.
Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.
Needleman, S. and Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.
Orengo, C., Bray, J., Hubbard, T., LoConte, L., and Sillitoe, I. 1999. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins* **37:** 149–170.
Ortiz, A.R., Kolinski, A., and Skolnick, J. 1998a. Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.* **277:** 419–448.
Ortiz, A.R., Kolinski, A., and Skolnick, J. 1998b. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc. Natl. Acad. Sci.* **95:** 1020–1025.
Ortiz, A.R., Kolinski, A., Rotkiewicz, P., Ilkowski, B., and Skolnick, J. 1999. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins* **(Suppl 3):** 177–185.
Perrakis, A., Morris, R., and Lamzin, V.S. 1999. Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* **6:** 458–463.
Prestegard, J.H., Valafar, H., Glushka, J., and Tian, F. 2001. Nuclear magnetic resonance in the era of structural genomics. *Biochemistry* **40:** 8677–8685.
Reddy, B.V., Li, W.W., Shindyalov, I.N., and Bourne, P.E. 2001. Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins. *Proteins* **42:** 148–163.
Sali, A. 1998. 100,000 protein structures for the biologist. *Nat. Struct. Biol.* **5:** 1029–1032.
Sander, C. and Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9:** 56–68.
Shindyalov, I.N. and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11:** 739–747.
Shindyalov, I.N. and Bourne, P.E. 2000. An alternative view of protein fold space. *Proteins* **38:** 247–260.
Siew, N., Elofsson, A., Rychlewski, L., and Fischer, D. 2000. MaxSub: An automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **16:** 776–785.
Simons, K.T., Bonneau, R., Ruczinski, I.I., and Baker, D. 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **37:** 171–176.
Simons, K.T., Strauss, C., and Baker, D. 2001. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306:** 1191–1199.
Taylor, W. and Orengo, C. 1989. Protein structure alignment. *J. Mol. Biol.* **208:** 1–22.
Teichmann, S.A., Murzin, A.G., and Chothia, C. 2001. Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.* **11:** 354–363.
Thornton, J. 2001. Structural genomics takes off. *Trends Biochem. Sci.* **26:** 88–89.
Thornton, J.M., Todd, A.E., Milburn, D., Borkakoti, N., and Orengo .A. 2000. From structure to function: Approaches and limitations. *Nat. Struct. Biol.* **7:**, 991–994.
Vitkup, D., Melamud, E., Moult, J., and Sander, C. 2001. Completeness in structural genomics. *Nat. Struct. Biol.* **8:** 559–566.
Yang, A. and Honig, B. 2000. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* **301:** 665–678.
Zemla, A., Venclovas, C., Moult, J., and Fidelis, K. 1999. Processing, and analysis of CASP3 protein structure predictions. *Proteins* **Suppl (3)** 22–29.