Relationship between multiple sequence alignments and quality of protein comparative models

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · OCTOBER 2004

Impact Factor: 2.63 · DOI: 10.1002/prot.20284 · Source: PubMed

Domenico Cozzetto

University College London

Relationship Between Multiple Sequence Alignments and Quality of Protein Comparative Models

Domenico Cozzetto and Anna Tramontano*
Department of Biochemical Sciences, University "La Sapienza" Rome, Italy

ABSTRACT **Comparative modeling is the method of choice, whenever applicable, for protein structure prediction, not only because of its higher accuracy compared to alternative methods, but also because it is possible to estimate *a priori* the quality of the models that it can produce, thereby allowing the usefulness of a model for a given application to be assessed beforehand.**

By and large, the quality of a comparative model depends on two factors: the extent of structural divergence between the target and the template and the quality of the sequence alignment between the two protein sequences. The latter is usually derived from a multiple sequence alignment (MSA) of as many proteins of the family as possible, and its accuracy depends on the number and similarity distribution of the sequences of the protein family.

Here we describe a method to evaluate the expected difficulty, and by extension accuracy, of a comparative model on the basis of the MSA used to build it. The parameter that we derive is used to compare the results obtained in the last two editions of the Critical Assessment of Methods for Structure Prediction (CASP) experiment as a function of the difficulty of the modeling exercise. Our analysis demonstrates that the improvement in the scope and quality of comparative models between the two experiments is largely due to the increased number of available protein sequences and to the consequent increased chance that a large and appropriately spaced set of protein sequences homologous to the proteins of interest is available. Proteins 2005;58:151–157. **© 2004 Wiley-Liss, Inc.**

Key words: comparative modeling; CASP experiments; multiple sequence alignment; protein structure evolution

INTRODUCTION

Proteins are linear polymers of amino acids whose sequence is selected by evolution for its ability to fold independently and spontaneously into a unique functional conformation. A protein structure is stabilized by a large number of relatively weak interactions, and therefore the number of 'correct' combinations of amino acids compatible with a sufficiently stable protein conformation is low compared to the number of possible ones. The consequences of this are far reaching; perhaps the most important is that proteins preserve their structure during evolution.[1]

The most frequent evolutionary events are single amino acid mutations, and in order to be accepted and transmitted to the progeny each of these has to be compatible with life. It follows that we can observe only those proteins derived from a common ancestor via the accumulation of evolutionarily accepted changes. A mutation can destabilize the protein structure, leading to the inability of the organism to survive, or it can be assimilated into the structure, producing only local structural changes. The remaining possibility, that the new protein sequence is compatible with a totally different but similarly stable and equally functional conformation, is highly unlikely and has not been observed thus far. This implies that evolutionarily related proteins have similar structures. Nevertheless, the local structural effects of mutations will accumulate, and the differences among the structures of evolutionarily related proteins will be greater for more distant evolutionary relationships.[2–4]

Chothia and Lesk[1] studied the correlation between evolutionary divergence and structural variation in a seminal work in which they compared the known structures of pairs of homologous proteins and related the extent of structural changes to their evolutionary distance, as measured by their sequence differences.

The distance between two sequences can be estimated by counting the number of different amino acids in corresponding positions, i.e. in structurally equivalent positions. In their work, Chothia and Lesk analyzed pairs of proteins of known structure, for which the correct identification of the pairs of corresponding positions (alignment) in any two structures is almost straightforward.[5–10]

The observed correlation between sequence and structure changes in homologous proteins forms the basis of comparative modeling, the most commonly used method of protein structure prediction.[11]

In this approach, the structure of a protein (target) is approximated by that of one or more homologous proteins of known structure (templates), and the correspondence between structurally equivalent amino acids is inferred from the optimal sequence alignment (usually the one that maximizes similarities) of the amino acid sequences.

Correspondence to: Anna Tramontano, Department of Biochemical Sciences, University "La Sapienza" P.le Aldo Moro, 5, 00185 Rome, Italy. Email: anna.tramontano@uniroma1.it

Received 20 May 2004; Revised 30 June 2004; Accepted 1 July 2004

Published online 19 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20284

Therefore the quality of the produced model will, to a first approximation, depend upon two factors: the expected structural divergence between the two protein structures as judged by their evolutionary distance, and our ability to reconstruct the correct structural alignment only using amino acid sequences or, in more recent approaches, amino acid sequences and the known structure of the template protein.

Intuitively, the expected accuracy of an alignment between two protein sequences depends on their sequence similarity. However, present model building procedures rarely make use of two sequences alone. Homology is transitive; therefore if two proteins are evolutionarily related to a third protein, they are also evolutionarily related to each other. This can be used to detect more distant evolutionary relationships in database searching strategies, by 'hopping' in sequence space from one homologous protein to the next and thus increasing the number of proteins that can be included in the family.[12,13]

Making use of sequences of proteins that are intermediate between target and template not only allows more distant evolutionary relationships to be detected but also improves our chances of obtaining a correct alignment. The alignment of two closely related sequences is less prone to errors and, once obtained, can be used to guide the alignment of a third, more distant, sequence iteratively until all known members of the protein family are aligned. This is the method of choice in all present comparative modeling experiments, and its advantages have been clearly demonstrated in many cases.

However, the use of many intermediate sequences between target and template makes it incorrect to evaluate the expected accuracy of a model solely on the basis of the sequence identity or similarity between target and template. The quality of the alignment between these two sequences will also depend on the number and the similarity distribution of all the sequences of the multiple sequence alignment (MSA).

This issue has several implications: for example, the 'quality' of the MSA needs to be taken into account when evaluating *a priori* the expected quality of a comparative model, an important practical issue since it permits us to decide whether the model will be of sufficient quality for the desired applications.

It also has important implications for the evaluation of prediction methods. There are several worldwide initiatives, the most popular being CASP,[14–18] for assessing the reliability of protein structure prediction methods and for verifying the extent to which the field is progressing. In CASP and other similar initiatives, predictors are asked to model the same set of target proteins so that different methods can be compared on examples of the same difficulty. However, understanding whether there has been progress between different editions of the experiment is a non-trivial issue, since for this purpose one needs to compare methods at different times, i.e. on different sets of targets.

Better results in subsequent experiments may be due to genuine method improvements or to differences in the difficulty of the targets between the two sets. As we said, a larger protein family may facilitate a correct alignment; therefore targets of the same apparent difficulty, as judged by their sequence similarity to the template, can become progressively easier as sequence databases continue to grow. Pair-wise sequence similarity between target and template is therefore not a good parameter to estimate the difficulty of the targets.

Here we describe a method to evaluate the difficulty of the pair-wise alignment between two sequences implied by a given MSA and show its application to the analysis of the results of two subsequent CASP experiments to answer the long-standing question of whether comparative modeling methods have improved.

MATERIALS AND METHODS

The targets of the CASP4 and CASP5 experiments are available from the CASP web site (http://predictioncenter.llnl.gov). The same site lists the optimal structural template for each target as determined by a structural search in the Protein Data Bank (PDB) database[19] at the time of the experiment.

In order to collect the sequences available to predictors to build a MSA, we performed a maximum of five iterations of PSI-BLAST[20] with default parameters on the non-redundant sequence database frozen at the time of the corresponding experiment. We discarded the targets for which the optimal structural template could not be detected by PSI-BLAST.

CLUSTALW[21] with default parameters was used to obtain a MSA, including the target, template and all sequences with E-values lower than the E-value of the target–template alignment. Sequences spanning less than 80% of the aligned region between target and template were removed from the alignment.

Structural superposition of targets and templates was obtained using LGA[22] and limited to the core region (defined as the set of residues with corresponding C_α distances shorter than 5 Å). Correctly aligned residues in the model are defined as those for which the closest residue in the target is the correct one, and the distance between them is less than 3.8 Å. The percent of correctly aligned residues is calculated with respect to the core.

RESULTS AND DISCUSSION

A typical comparative modeling experiment can be described as a stepwise procedure: (1) given the target protein sequence, run a database search program such as BLAST, FASTA or PSI-BLAST to find a statistically significant similarity to one or more proteins of known structure; (2) collect all sequences likely to belong to the common evolutionary family of the target and the template; (3) build a MSA using methods such as CLUSTALW[21] or T-Coffee[23]; (4) extract the target–template(s) alignment(s) from the MSA; (5) use the alignment(s) to guide model construction.

The quality of the final pair-wise alignment will depend on the similarity between the target and template sequences and also on the number and similarity distribution of homologous sequences in the MSA.

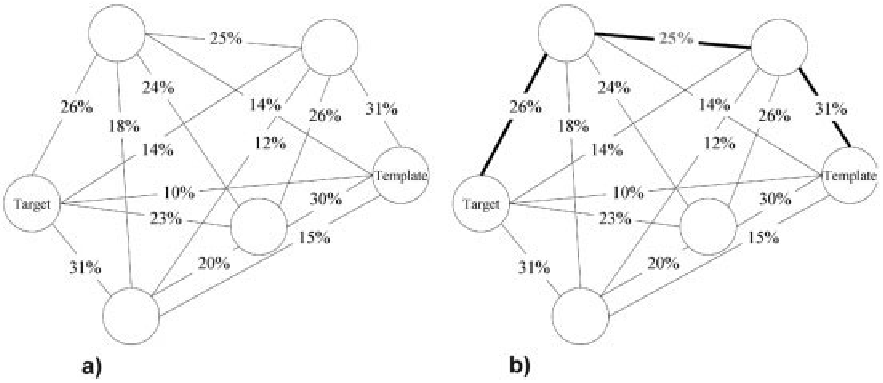

Fig. 1. Schematic representation of a graph deduced from a MSA. Each edge is labeled with the percent identity between the sequences that it connects (a). The path for which the shortest edge is maximal is shown as a thick line in (b). The value of the parameter μ in this example is 25.

In order to evaluate the difficulty of a pair-wise alignment extracted from a MSA, we calculated the pair-wise sequence identity between each pair of sequences and used the values to construct a graph similar to that shown in Figure 1(a), where each node represents one of the sequences in the MSA, and the lengths of the edges are proportional to the distance (inversely proportional to the percent of identity) between the connected nodes. The MSA is a path in the graph that includes all of the sequences.

Our aim was to estimate the upper bound of the difficulty of aligning target and template given the availability of intermediate sequences, and this depends upon which is the most difficult pair-wise alignment that we need to perform in order to go from the target to the template. Therefore, given all possible paths including target and template, we were interested in the one(s) for which the maximum distance between each pair of traversed nodes is minimal [Fig. 1(b)].

Once such a path is found, the maximum distance in the path, i.e. the sequence similarity between the two most diverse sequences in the set, which we call μ, is an estimate of the difficulty of aligning target and template, given the distribution of sequences in the MSA. We do not take into account the improvements that can be obtained by using sequence profiles throughout the alignment, as their effect would depend on the specific implementation of the alignment scoring method.

Our algorithm to find the path and the value of μ has a complexity $O(N \log N)$, where N is the number of sequence pairs in the alignment and its details will be described elsewhere. Here we show the application of the method to the comparison of results obtained in two subsequent CASP experiments. The best results obtained for targets of similar difficulty in the two experiments are compared to assess the extent to which methods for comparative modeling have improved in the intervening two years.

The CASP4[17] and CASP5[18] targets belonging to the comparative modeling category are listed in Table I, along with their best structural template at the time of the respective experiment and available on the CASP web site.

Figure 2 shows a plot of the root mean square deviation (RMSD) between the core of the target–template pairs after optimal superposition of their structures for CASP4 and CASP5 comparative modeling targets as well as for the original pairs of proteins used by Chothia and Lesk[1] for their analysis. Sequence identity was calculated on the basis of the structural alignment. The quality of the structural alignment is the upper limit that can be reached by any sequence-based alignment method: if it is correctly reproduced, the implied model should be at least as close to the experimental structure as the template.

Figure 2 highlights a few peculiar cases in which the structural difference between the target and the template is higher than would be expected on the basis of their sequence similarity. These are interesting illustrations of difficulties that can arise in a comparative modeling experiment and are independent of the ability of any method to detect the correct alignment.

In the case of T0099 (CASP4), the target protein is a designed molecule. As we said before, comparative modeling is based on the assumption that the observed changes have accumulated during evolution and have been accepted. This is not the case for a designed protein, and therefore the relationship between sequence and structure similarity derived for naturally evolving proteins does not hold in this case.

Another case, interesting but difficult to solve, is illustrated by T0123 (CASP4). This is a case in which one of the two proteins has a domain swapping. In other words, two homo-dimeric proteins having two domains per chain swap the position of the domains from one chain to the other during evolution (Fig. 3). This is a really tricky case, in which a single mutation can easily destabilize one structure with respect to the other. It is very difficult to predict and impossible to detect by sequence-based methods.

We excluded these cases from our analyses, but it is worth stressing that the last example is one of the problems of comparative modeling to which special attention should be devoted in future methods.

If these cases are not taken into account, the data can be fitted with the exponential: $\text{RMSD} = 0.4\, e^{2(1\ -\ \text{fraction id})}$

TABLE I. CASP4 and CASP5 Comparative Modeling Targets

	Target	Template	% ID	Max % AL0	μ	Number of Sequences	N_{eff}	Notes
CASP4	T0089_2	1dga (1c0f_a)	11.99	72.34	20	829	18.09	1
	T0089_4	1dej_a	9.17	75.79	20	1070	15.76	
	T0090_2	1mut	13.99	57.55	—	—	—	2
	T0092	1d2c_a	4.03	64.44	10	368	19.99	
	T0099	1lck_a	40.35	82.98	—	—	—	3,4
	T0103	1ak9	7.64	58.75	11	56	13.7	
	T0111_1	1ebh	55.82	100.00	65	87	5.27	
	T0111_2	6enl	45.19	98.63	57	92	5.64	
	T0112_1	1bxz	10.81	88.37	26	139	12.95	
	T0112_2	1agn_a	11.97	92.68	30	222	14.59	
	T0113	1ahi_b	20.22	94.14	34	162	12.07	
	T0117	1vtk	11.07	85.88	14	161	16.5	
	T0121_1	1b0u_a	26.36	76.63	37	417	14.1	
	T0121_2	1b9n	—	73.75	—	—	—	5
	T0122	c29_a	31.62	91.03	40	13	2.92	
	T0123	1beb_a	64.20	80.43	—	—	—	6
	T0125	3lyn_b	17.65	77.78	21	8	3.86	
	T0128	1b06_a	54.07	100.00	54	5	1.7	
CASP5	T0130	1fa0_b	4.43	66.26	18	176	17	
	T0132	1lo7_a	—	87.27	—	—	—	5
	T0133	1hf8_a	10.00	74.19	17	52	7.04	
	T0136_1	1nzy_a	—	65.16	—	—	—	5
	T0136_2	1jxz_b	—	66.24	—	—	—	5
	T0137	1pmp_a	42.22	98.49	45	31	2.81	
	T0140_1	1b8a_a	—	63.08	—	—	—	1,4,6
	T0141	1aro_l	13.07	66.67	—	—	—	3
	T0142	1i9z_a	21.66	82.87	26	40	6.65	
	T0143_1	1agj_a	21.08	89.29	23	14	2.98	
	T0143_2	1qtf_a	32.67	100.00	34	12	2.36	
	T0149_1	1de0_b	—	52.05	—	—	—	5
	T0150	1jj2_f	21.90	94.74	33	82	8.64	
	T0151	1qvc_a	24.64	92.71	31	62	5.76	
	T0152	1kux_a	12.23	69.63	16	378	19.59	
	T0153	1euw_a	33.97	99.22	48	93	7.79	
	T0154_1	1iho_a	48.31	100.00	50	16	3.52	
	T0154_2	1iho_b	35.24	86.60	43	15	4.54	
	T0155	1dhn	33.33	100.00	39	16	3.52	
	T0159_1	1iit_a	—	56.60	—	—	—	5
	T0159_2	1eu8_a	—	44.90	—	—	—	5
	T0160	1grw_a	17.60	90.68	22	73	8.97	
	T0165	1qfs_a	7.12	76.65	12	415	20.31	
	T0167	1jeo_a	33.51	84.85	37	23	5.49	
	T0168_1	1pmd	—	49.23	—	—	—	5
	T0168_2	1ga0_a	—	52.32	—	—	—	5
	T0169	1bo4_a	11.18	95.65	19	393	19.03	
	T0172_1	1i9g_a	8.17	83.33	44	691	20.5	
	T0176	1jrm_a	19.05	65.79	—	—	—	3
	T0177	1lfp_a	27.64	85.78	36	30	6.1	
	T0178	1jcj_a	22.65	96.71	34	61	7.8	
	T0179_1	1jq3_a	37.50	100.00	45	8	2.8	
	T0179_2	1jq3_a	45.70	97.59	48	8	2.61	
	T0182	3mat_a	42.10	97.98	48	17	3.55	
	T0183	1jcl_a	22.90	94.95	33	60	7.73	
	T0184_1	1jfz_a	30.72	86.61	36	82	9.54	
	T0184_2	1di2_a	2.00	95.52	19	191	16.96	
	T0185_1	2uag_a	10.89	70.83	25	111	10.78	
	T0185_2	1jbv_a	5.00	79.00	17	415	16.32	
	T0185_3	2uag_a	7.33	73.45	19	111	11.26	

TABLE I. (Continued)

	Target	Template	% ID	Max % AL0	μ	Number of Sequences	N_{eff}	Notes
CASP5	T0186_1	1eyi_a	—	81.43	—	—	—	5
	T0186_2	1k70_a	—	48.97	—	—	—	5
	T0188	1eol_a	25.00	87.38	29	19	6.46	
	T0189	1j5v_a	8.51	67.16	20	367	17.86	
	T0190	2trh_a	27.12	95.45	34	66	6.75	
	T0191_2	1d4f_d	—	72.50	—	—	—	5
	T0192	1qsm_d	14.79	82.64	25	93	13.22	
	T0193_2	1evj_d	—	59.35	—	—	—	5
	T0195	1f0n_a	11.04	77.82	16	96	14.23	

The first column reports the CASP code for each target structure. The second is the PDB code of the optimal template followed by the chain identifier. The third column reports the percent of pair-wise sequence identity between target and template. The column named "max % AL0" is the maximum percent of correctly aligned residues obtained by any prediction submitted to CASP.

Notes: (1) The optimal template could not be found by PSI-BLAST, but a similarity with a protein of known structure (code in parentheses) could be detected. (2) The template structure has been determined by NMR. (3) The target structure has been determined by NMR. (4) Synthetic protein. (5) The optimal template could not be found by PSI-BLAST. (6) Domain swap with respect to the best template.

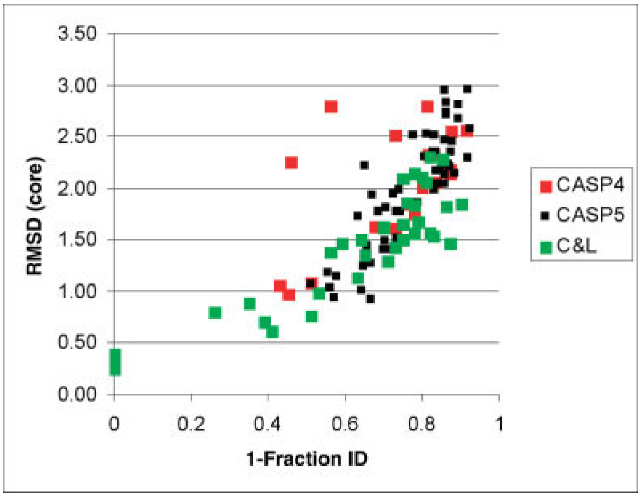

Fig. 2. The relationship between sequence identity and structural similarity for CASP4 (red) and CASP5 (black) targets. The green points refer to the pairs of structures used by Chothia and Lesk in their original analysis.[1]

with a correlation of 0.88, to be compared with the equation reported by Chothia and Lesk for a smaller dataset: RMSD = $0.4\,e^{1.9(1\,-\,\text{fraction id})}$.

Next, we searched the non-redundant sequence database for each of the domains. For each set of targets, we used the release of the database available at the time of the experiment and ran five iterations of PSI-BLAST. If the best structural template could not be detected with such a procedure, the target was not analyzed further. Similarly, we excluded cases in which either the template or the target structure was determined by NMR, as the detection of the correctly aligned residues can be ambiguous in some regions. In the remaining cases, we collected all sequences in the PSI-BLAST output whose length spanned at least 80% of the region of superposition between target and template in the MSA and constructed a graph similar to that shown in Figure 1 for each of them.

We subsequently calculated the parameter μ, the maximum distance between two sequences of the alignment that had to be included to connect the target to the

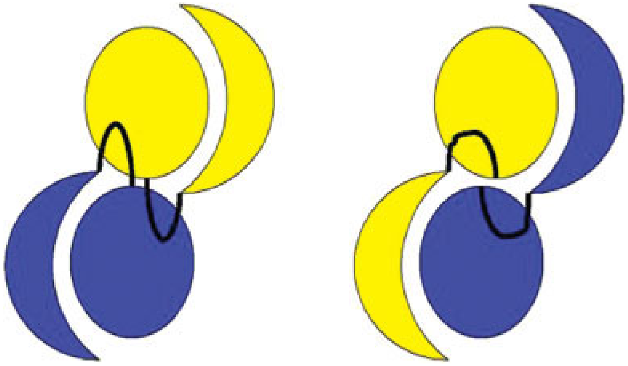

Fig. 3. Schematic illustration of a swap domain. Two evolutionarily related proteins can have a different relative positioning of one of their domains. Note that, because of the symmetry, most of the interactions, with the exclusion of those involving the black connector, are the same in both cases.

template (Fig. 1). We also computed, for each target, the percent sequence identity between the target and the template and the parameter N_{eff}, an estimate of the alignment diversity defined as the average number of different symbols in all columns of the alignment containing less than 50% deletions.[24]

The values of μ, of the pair-wise sequence identity between target and template, and of N_{eff} are shown in Table I.

Figure 4(a) shows a plot of the quality of the best alignment obtained in CASP4 and CASP5 for each of the targets as a function of the pair-wise sequence identity between target and template. The analysis of the figure highlights some of the problems connected with the use of pair-wise sequence identity for such analyses.

For example, it would appear that comparative modeling could be applied to a pair of proteins sharing only 2% sequence identity, and this is clearly impossible.

Furthermore, it seems that the alignment of targets sharing similar sequence identity with their templates can be of different quality for different targets. Because of this problem, it is quite difficult to judge whether there has really been a change in the accuracy of the alignment between two experiments.[25]

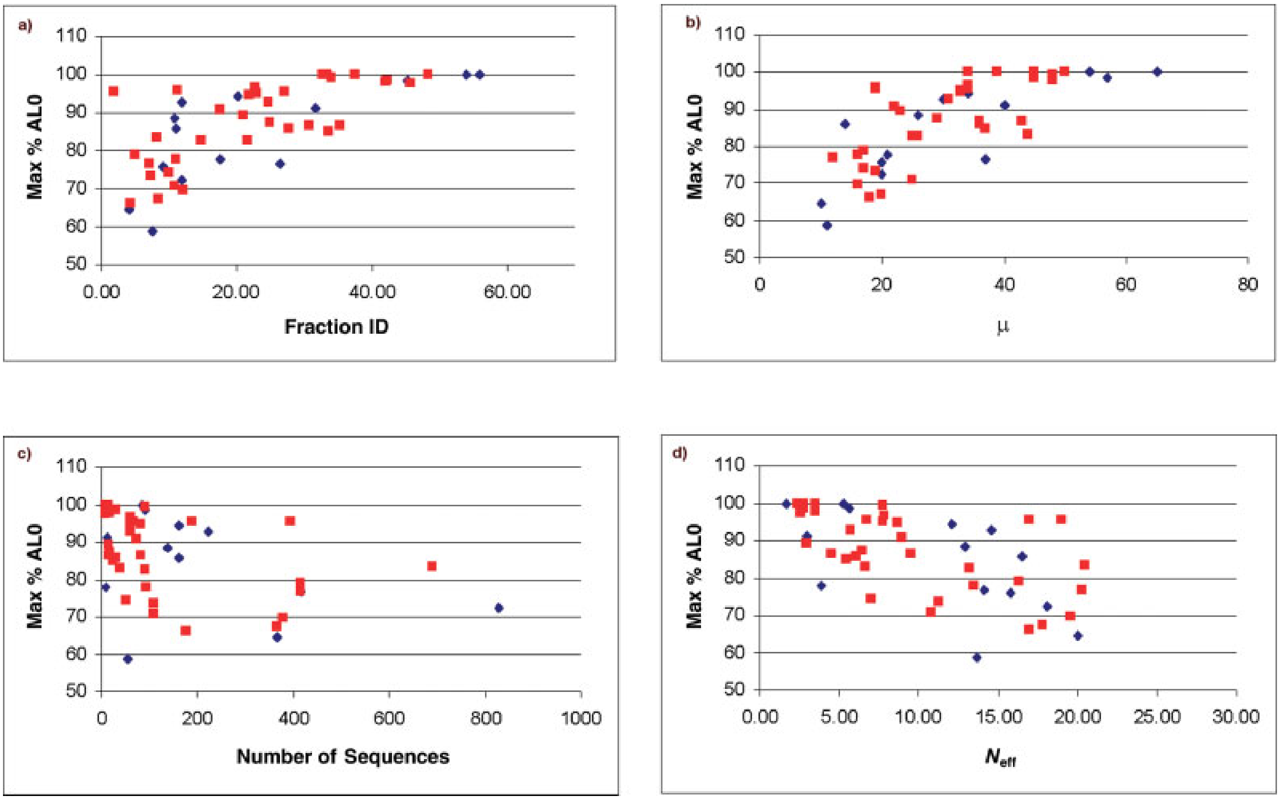

Fig. 4. Scatter plot of the maximum alignment quality obtained in CASP4 (blue) and CASP5 (red) as a function of the percent sequence identity between target and template (a), of the parameter μ (b), of the number of sequences in the alignment (c) and of the parameter N_{eff} (d).

In Figure 4(b), we report the maximum accuracy of alignment achieved in CASP4 and CASP5 for each target as a function of the parameter μ. As a comparison, Figures 4(c and d) show the same data plotted as a function of the number of sequences in the alignment and of N_{eff}.[24]

The parameter μ appears to be a better tool to analyze the difficulty of a target. For example, a comparative model is produced only if target and template can be connected via pair-wise alignments not including pairs of sequences sharing less than $\approx$10% sequence identity, a low but much more realistic threshold. Also the Pearson correlation coefficient moderately increases with respect to that obtained using the fraction of identical residues (0.73 vs. 0.70). The results obtained by considering only the predictions submitted by automatic servers are very similar (data not shown).

Notably, the parameter μ has a better predictive power of the expected accuracy of a model. The data can be fitted using the equation $\max\%\text{AL0} = 19 \ln(\mu) + 23$, $R^2 = 0.58$. None of the other parameters (sequence identity, N_{eff}, number of sequences in the alignment) allows to fit the percent of correctly aligned residues with a correlation coefficient above 0.45.

Finally, by separately fitting the data for CASP4 and CASP5 we obtain the following: $\max\%\text{AL0}_{\text{CASP4}} = 19 \ln(\mu_{\text{CASP4}}) + 21$ and $\max\%\text{AL0}_{\text{CASP5}} = 18 \ln(\mu_{\text{CASP5}}) + 25$. This implies that the average improvement in alignment accuracy for targets of equivalent difficulty in the two experiments is below 1% for $\mu > 20$ and never above 2% for all comparative modeling targets.

Although it is true that, as noted by the CASP5 comparative modeling assessor,[25] the range of applicability of the method has increased between CASP4 and CASP5, we can now confirm the hypothesis that this is due to the larger set of sequences available for some families rather than to genuine improvement of alignment methods.

By using the parameter μ as an estimate of the difficulty of the target, a quantitative relationship between the quality of the alignment and the difficulty of the target can be established, and this can be used more effectively to estimate the expected quality of a model given a MSA including target and template.

CONCLUSIONS

Comparative modeling, whenever applicable, is the method of choice for producing models of proteins, both because of its higher accuracy compared to other methods of prediction and because it allows the quality of the model, and therefore the scope of its applications, to be estimated *a priori*.

Common considerations regarding the expected quality of the produced models are generally based on the percent of sequence identity or similarity between target and template, although it is common knowledge that an inspection of the quality of the multiple alignment is indispensable to estimate the reliability of the final result.

So far, to the best of our knowledge, no parameter has been proposed to directly link the number and distribution of sequences in the MSA to the expected quality of the model. Here we present a simple way of calculating such a parameter, which takes into account the distribution of the sequence identity between the sequences included in the MSA and demonstrates that it represents a good estimator of the expected quality of a comparative model.

The availability of such a parameter will allow comparative modeling results obtained on different datasets to be compared, as demonstrated here by the analysis of the CASP4 and CASP5 results. Hopefully, it will also improve current methods for the selection of target proteins in structural genomics projects.

ACKNOWLEDGMENTS

The authors are indebted to Krysztof Fidelis, Ceslovas Venclovas and Adam Zemla for providing CASP data and the sequence database versions available at the time of the CASP experiment and to Professor Rossella Petreschi for continuous support and encouragement.

This project was supported by the Finanziamento Ateneo dell'Università di Roma "La Sapienza" and by the EU funded BioSapiens Network of Excellence.

REFERENCES

1. Chothia C, Lesk A. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
2. Chothia C, Lesk A. The evolution of protein structures. Cold Spring Harb Symp Quant Biol 1987;52:399–405.
3. Hubbard T, Murzin A, Brenner S, Chothia C. SCOP: a structural classification of proteins database. Nuc Acid Res 1997;25:236–239.
4. Orengo C, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
5. Lesk A. Extraction of well-fitting substructures: root-mean-square deviation and the difference distance matrix. Fold Des 1997;2:S12–S14.
6. Rose J, Eisenmenger F. A fast, unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm. J Mol Evol 1991;32:340–354.
7. Shapiro A, Botha J, Pastore A, Lesk A. A method for multiple superposition of structures. Acta Crystallogr 1992;A48 (Pt 1):11–14.
8. Taylor W, Flores T, Orengo C. Multiple protein structure alignment. Protein Sci 1994;3:1858–1870.
9. Maiorov V, Crippen G. Size-independent comparison of protein three-dimensional structures. Proteins 1995;22:273–283.
10. May A, Johnson M. Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. Protein Eng 1994;7:475–485.
11. Tramontano A. Homology modeling with low sequence identity. Methods (San Diego, CA) 1998;14:293–300.
12. Teichmann SA, Chothia C, Church GM, Park J. Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL. Bioinformatics (Oxford, England) 2000; 16:117–124.
13. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for iscovery in protein databases. Trends Biochem Sci 1998;23:444–447.
14. Moult J, Pedersen J, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins 1995;23:ii–v.
15. Moult J, Hubbard T, Bryant S, Fidelis K, Pedersen J. Critical assessment of methods of protein structure prediction (CASP): round II. Prot Suppl 1997;1:2–6.
16. Moult J, Hubbard T, Fidelis K, Pedersen J. Critical assessment of methods of protein structure prediction (CASP): round III. Prot Suppl 1999;3:2–6.
17. Moult J, Fidelis K, Zemla A, Hubbard, T. Critical assessment of methods of protein structure prediction (CASP): round IV. Prot Suppl 2001;5:2–6.
18. Moult JAZ, Fidelis K, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. Prot Suppl 2003;6: 334–339.
19. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer-based archival file for macromolecular structures. Eur J Biochem 1977;80:319–324.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acid Res 1997;25:3389–3402.
21. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. Meth Enzymol 1996;266:383–402.
22. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucl Acid Res (Online) 2003;31:3370–3374.
23. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217.
24. Sadreyev RI, Grishin NV. Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs. Bioinformatics 2004;20:818–828.
25. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Prot Suppl 2003;53(6):352–368.