How similar must a template protein be for homology modeling by side–chain packing methods ?

Su Yun Chung[1] and S Subbiah[2]

1. Department of Biochemistry, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814
2. Department of Structural Biology, Stanford University School of Medicine, Sherman Fairchild Building, Stanford, CA 94305

Abstract

Given the correct backbone coordinates of a globular protein, side–chain packing methods can be generally expected to predict the side–chain coordinates of the buried core residues accurately. In the context of a study in modeling a family of bacteriophage DNA–binding proteins, we observed that when the coordinates of the actual perfect backbone are not available, the side–chain packing methods are still of predictive value using homologous but imperfect backbones. This is the situation in practical homology modeling where a target protein sequence is modeled from template structures of known protein homologs. In order to assess the quality and degree of accuracy of such predictions and their dependence on the extent of homology, we have now extended these studies to a well characterized family of globin structures that span a much wider range of sequence–structure similarity. The collective results show a clear relationship that is independent of protein family between side–chain prediction accuracy and the level of similarity between the template and target proteins. We judge this similarity in terms of sequence identity and the backbone r.m.s. deviation of the template structure used for modeling and the actual target structure in cases where the target structures are available. In summary, as sequence identity drops from 100% to about 50%, or when the backbone r.m.s. deviation between template and target structures increases from 0 Å to about 1 Å, the overall average r.m.s. error for the buried–core residues rises from 1.2 Å to 1.5 Å while the $\chi 1$ prediction accuracy drops from 85% to 70–75% and the $\chi 2$ prediction accuracy drops from 80% to 60–65%. When the sequence identity drops below 50% or the backbone r.m.s. deviation rises above 1 Å , all 3 measures of prediction accuracy decrease rapidly. When the sequence identity edges to the so–called twilight zone of sequence similarity at around 22%, or when the backbone r.m.s. deviation exceeds 2 Å, the prediction accuracy approaches the values to be expected for random predictions, namely, 3.1 Å for average r.m.s. error, 22% and 29% for accuracy of $\chi 1$ and $\chi 2$ prediction. These observations provide a practical evaluation of the side–chain packing methods and are of value to the homology–modeler. The extent and degree to which the backbone topology of a protein fold can constrain internal side–chain orientation gives insight into the plasticity of the sequence–structure relationship found in the architecture of proteins.

1) Introduction

It is now well established that given a perfect protein backbone, the rotamer orientation of buried side–chain atoms can be accurately predicted by 'packing considerations' alone (1). By virtue of the constraining environment provided by the backbone, the conformations of the buried side–chains are practically uniquely defined. The question of how effectively such backbone constraints can allow prediction of side–chain orientation in the event where only imperfect but similar template backbones are available has been recently explored (2). By systematically homology modeling proteins from two structurally well–studied families – the bacteriophage repressor and cro proteins and the oxygen–binding globins – over a wide range of pair–wise sequence identities (13% to 100%) and pair–wise backbone r.m.s.(root–mean– square) deviation (0 Å to 1.99 Å), we illustrate the influence of this constraining effect of the backbone on prediction accuracy as a function of increasing r.m.s. deviation between the template and target backbones (target/template r.m.s. deviation). Since the backbone r.m.s. deviation is known to increase with decreasing sequence identity (3), we also consider the relationship between side–chain prediction accuracy and decreasing sequence identity between target and template. From such considerations, we find that when the sequence identity is above ~22% or when the target/template r.m.s. deviation is below ~2 Å, the assumption of a fixed, even if imperfect, backbone template against which to enumerate side–chain packing possibilities is useful. Beyond these limits, it is clear that this assumption, and by extension any method that relies on the same assumption, is not of sufficient predictive value to result in a homology model that could be of use to the experimental biologist. In particular, when the target/template r.m.s. deviation is in the range of 1 to 2 Å, typically corresponding to some 20% to 50% sequence identity (3–5), packing calculations can result in useful homology models with correct predictions of the buried core residues. These evaluations and other related guidelines for homology modeling success are derived by the systematic modeling and analysis described in the remainder of this manuscript.

Starting with our original work on side–chain packing optimization, many other methods have been used to demonstrate repeatedly that the side–chain coordinates of the residues in the buried core, corresponding roughly to between a third and a half of all the residues in a globular protein, can be predicted quite accurately by computational means if the accurate coordinates of the backbone atoms of a protein is given (6–15). In test case studies, the prediction accuracy at buried positions is typically about 1.2 Å r.m.s. deviation for the side–chain atoms of all predicted residues (15). Measured in terms of the absolute difference in the torsional side–chain χ angles, where similarity is judged on whether two angles are less than 40 degrees different, the success rate is around 82% for $\chi 1$ and 78% for $\chi 2$. More recently, in the context of an initial study involving the modeling of the DNA–binding repressor and cro proteins of bacteriophages 434 and P22, we have demonstrated that even when the available template backbone is less than perfect, significant levels of prediction accuracy approaching that for the perfect test–case

studies quoted above can be attained (2). The cross–modeling studies between the known X– ray structures of the 434 cro and 434 repressor structures and the NMR structure of P22 repressor suggested some intriguing reference points for prediction accuracy measured as a function of both the sequence identity and the target/template structural similarity. Nevertheless, by virtue of the familial similarity particular to these DNA–binding proteins, these reference points do not span a wide enough range of sequence or structure space to reflect adequately the general behavior of prediction accuracy as a function of sequence–structure similarity. We now rectify this by exploring a more standard family for which the structures of proteins with a wider range of sequence similarity exist in the database; namely, the globin family, myoglobins and hemoglobins, which is very well characterized with more than a dozen different high–resolution crystal structures of varying degree of sequence similarity available (16). Additionally, several hundred easily–aligned globin protein sequences with no corresponding crystal structures are also known (17). We have cross–modeled a selected series of these structures that have sequence–structure similarity levels complement to those from our previous cross–modeling of the bacteriophage repressor and cro proteins (2). Combining this data with the previous data has allowed us to ascertain some general rules–of–thumb regarding the suitability of employing side–chain packing methods that assume a fixed template backbone to practical homology modeling problems.

2) Materials and Methods

Modeling of the globin structures was carried out in a manner similar to our previous work with bacteriophage repressors (2). Altogether, 8 structures from the globin family were obtained from the Brookhaven database of protein structures (18–26). The sperm whale myoglobin structure, 1MBC (19), was always used as the target for cross–modeling the other 7 template structures, as well as for self–modeling itself from its own backbone. For each modeling exercise, the biological source of the proteins, their size, their sequence identity with the target sequence, and the r.m.s. deviation between the target/template backbones are tabulated in Table 1. This table also contains similar information for the proteins modeled in previous studies of the cro and repressor protein family of bacteriophage 434 and P22 (2, 27–29). There is little or no ambiguity in the sequence alignment of highly similar globin sequences and unique sequence alignments can be obtained using standard sequence alignment programs. When, as in our case, the structures of both proteins are known, these sequence alignments are typically almost identical to that derived from a structural alignment of the pair of backbones. However, as the sequence identity decreases, a sequence alignment derived from sequence information alone is fairly ambiguous, particularly in loop regions. Our previous work shows that the correct sequence alignment is absolutely critical to the success of homology modeling based on side–chain packing considerations (2). Our chief concern here is to assess how poor a template model can be for side–chain prediction by packing optimization and is not on sequence alignment, *per se*. So we have elected to use the sequence alignment derived from structural alignment of a given template structure with the target 1MBC structure, as the ideal starting point

Table 1. Modeling Globin and DNA-binding Protein Families

Template pdb entry (number of residues)		Template molecule	Target	Target/ template main-chain r.m.s.d. (Å)	Target/ template sequence identity (%)
1mbc	(153 aa)	Sperm whale myoglobin	1mbc	0	100
2mm1	(153 aa)	Human myoglobin mutant	1mbc	0.51	86
1ymb	(153 aa)	Horse myoglobin	1mbc	0.9	84
1lhs	(153 aa)	Loggerhead sea turtle myoglobin	1mbc	0.94	65
1myt	(146 aa)	Yellowfin tuna myoglobin	1mbc	1.25	42
1mba	(146 aa)	Sea hare myoglobin	1mbc	1.92	23
3sdh	(146 aa)	Ark clam hemoglobin I	1mbc	1.88	20
1ith	(141 aa)	Innkeeper warm hemoglobin I	1mbc	1.99	13
1r69	(63 aa)	434 repressor	1r69	0	100
2cro	(65 aa)	434 cro	2cro	0	100
1r69	(63aa)	434 repressor	2cro	0.78	53
2cro	(65 aa)	434 cro	1r69	0.78	53
1adr	(68 aa)	P22 repressor	1adr	0.00	100
1adr	(68 aa)	P22 repressor	1r69	1.70	33
1adr	(68 aa)	P22 repressor	2cro	1.80	32

for the modeling process. We performed the structural alignment using our program Structal, which automatically aligns equivalent Cα atoms (5). The Structal program ignores the prosthetic heme group. The backbone r.m.s. deviations reported in Table 1 as the 'target/template r.m.s. deviation' are based on the structural superpositions obtained in this manner. The values for the percentage sequence identity are also based on these structurally derived sequence alignments.

Each of the seven cross–modeling exercises started with knowing only the template backbone coordinates, the coordinates of the template heme group, and the structurally–derived sequence alignment of the target globin sequence (Table 1). Relative insertions and deletions in the sequence–structure alignment were treated as follows. Target residues that were not aligned to the template structure were not modeled and simply ignored. Residues in the template structure that had no target sequence aligned with them were left in a sidechain–less state at all times. Outside of a few residues inserted or deleted at loops only one significant insertion/deletion was encountered in all 8 modeling exercises. Relative to other globins in the set, the yellowfin tuna myoglobin lacks an entire short D helix (23).

Modeling was performed using the program Look (Molecular Application Group, 1995), which contains an implementation of side–chain packing methods (6,12). Based on the sequence alignment, at each residue position the target residue was added to the backbone in a completely random side– chain conformation (i.e. selected random torsion χ angles). The side–chain packing method performs a self–consistent ensemble optimization in a coarsely sampled side–chain torsion space and only considers a simple van der Waals constraint. This constraint is effectively provided by the fixed template backbone and by the side–chain–side–chain stearic exclusion between side–chain atoms from the moving residues. No other energy terms are considered explicitly. Unlike other methods, no database–derived rotamer libraries are used, and so the method is truly *ab initio*. Solvent effects are ignored. Since the packing criterion is strong for the fully or partially buried core residues, the method can typically predict such buried conformations to 1.2 Å r.m.s. error (6). Conversely, since there is nothing to pack against at the surface, the surface exposed residues are poorly predicted, with typical r.m.s. error approaching the 3.1 Å value, the expected r.m.s. error averaging over all amino acid types for completely random side–chain predictions (6).

When self–modeling, the side–chain coordinates of the known structure were stripped off, and the amino acid sequence was modeled using the remaining backbone coordinates and the prosthetic heme group jointly as a template. Both the main–chain and the heme group were held rigid at all times. The initial side–chain coordinates were chosen randomly by Look. In cross–modeling, the coordinates of the side–chain heavy atoms were predicted using the main–chain atomic coordinates from a different but homologous protein. Again, both this backbone and its accompanying prosthetic heme group were jointly used as a rigid template. Since the buried residues of the predicted models are more reliably predicted, the burial analysis calculation implemented in Look was carried out to rank order all residues in terms of their burial. Look calculates the percentage of burial by employing a

fine grid of points over atoms of each side–chain, and then assesses at each grid–point whether the point is within van der Waals reach of another protein atom, or not. For the purposes of this study two different cutoff values for burial were considered. The first set included the 50 most buried residues. Since the typical globin has about 150 residues, this corresponds to a stringent classification where a third of the protein (buried 1/3 residues in Fig. 1) is defined to be the buried core (30). A more lenient cutoff that includes the 80 most buried residues, corresponding to about half the protein (buried 1/2 residues in Fig. 1) being classified as a less–buried core residue set, was also used.

Comparison of the predicted model with the known reference X–ray structure was done after superposition of the two structures using Structal (5). The prediction accuracy was quantified in three different ways. Side–chain r.m.s. deviations were computed for each compared residue and then averaged to give a non–linear measure of overall prediction accuracy. To put the non– linearity into context, while a value of around 1 Å should be considered as exceptionally good, as stated earlier, if all the side–chain conformations of a typical protein are completely randomly predicted with no attention paid to stearic clashes, the side–chain r.m.s. error can be expected to be about 3.1 Å averaging over all amino acid types (6).

Quantitative differences in the side–chain torsional χ angles ($\chi 1$ and $\chi 2$) between the modeled structure and the reference structure were calculated at each residue position. As only absolute differences were considered, the maximum possible difference is 180 degrees. Since electrostatic effects are not considered by Look, amino–acids like Arg, Tyr, Phe, Asp, Asn, Glu, and Gln can be just as well–packed in two different, but symmetry–related, conformations. For such residues, this was taken into account by selecting the lower of the two possible values for both the side–chain r.m.s. deviation and the absolute difference in $\chi 2$ angle. Two χ angles were judged to be similar if the absolute difference was less than 40 degrees. The $\chi 1$ prediction accuracy was simply defined to be the percentage of buried–core $\chi 1$ angles that were similar between model and reference for all residues other than Ala, Gly and Pro. A baseline for measuring prediction success can be obtained by crudely estimating the percentage similarity that could be expected for a completely random prediction of $\chi 1$ angles. An absolute difference of 40 degrees allows a 80 degree window for defining success, while the total range for prediction is 360 degrees. Thus, random prediction of $\chi 1$ angles can result in a success rate of (80/360) x 100 = 22%. The $\chi 2$ angle similarity was also judged in the same manner using a 40 degree cutoff over the 13 residue–types that have χ angles. The estimate for the success rate for random χ prediction is a little higher than 22%, since 4 residues – Tyr, Phe, Asp, Asn – are from the packing method point of view symmetric about their $\chi 2$ angles and so have their range

restricted from 360 to 180 degrees. Therefore, the adjusted computation , (4(80/180) + 9(80/360)) x 100 , gives 29%.

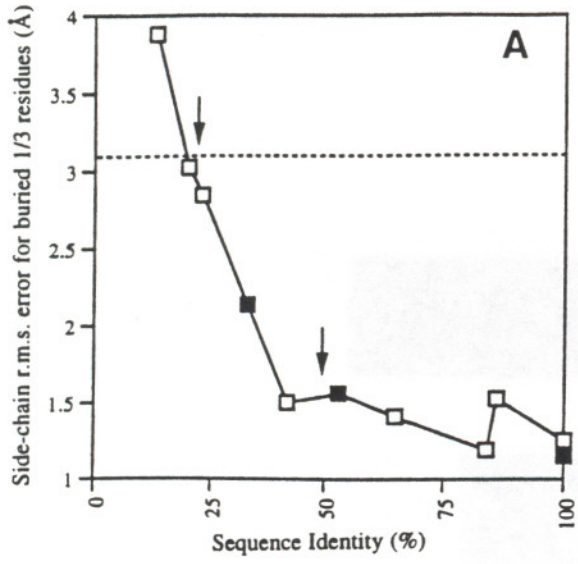
3) Results and Discussion

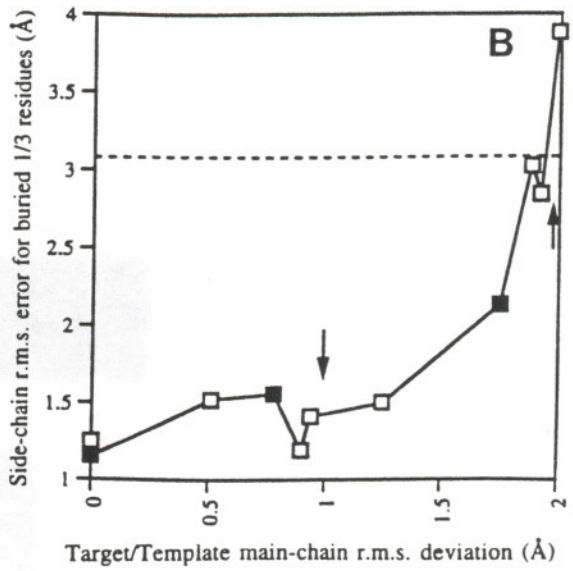
Prediction accuracy was assessed as a function of sequence identity over the sequence alignment corresponding to the correct structural alignment. It was also assessed as a function of the r.m.s. deviation between the backbones of the template being used to model the target sequence and the known target structure itself (i.e. the target/template r.m.s. deviation described earlier). The prediction accuracy itself was measured in three different ways: a) the r.m.s. deviation of all buried side–chain atoms for which coordinates had been predicted, b) the percentage of predicted side–chains for which the model– derived $\chi 1$ angles were within 40 degrees of the known structure, and c) an analogous percentage for the $\chi 2$ angle. These three measures of accuracy, when analyzed as a function of the two measures of similarity, result in six plots. All calculations were conducted for two different sets of buried–core residues, the most–buried third (Fig. 1A–1F) and the most–buried half (Fig. 1G–1L) of all residues, resulting in 12 plots in Fig. 1. The data points themselves come from two sources. Three data points come from our recent work on the homology modeling of bacteriophage DNA–binding proteins (2) and are summarized in Table 1. One of these comes from the mean results for the cross–modeling of the 434 repressor using the 434 cro template and the 434 cro using the 434 repressor template. Another comes from the mean for the modeling of P22 repressor based on the 434 repressor template and the modeling of the P22 repressor based on the 434 cro template. The last data point comes from the mean of all the self–modeling predictions for P22 repressor on itself, 434 cro on itself and 434 repressor on itself. As there are approximately 60–residues in each of these proteins, 20 residues were selected to define the most–buried third and 30–residues to define the most–buried half of each protein. The 8 modeling exercises, 1 self–modeling and 7 cross– modeling, conducted here on the globin family constitutes the second source of data points.

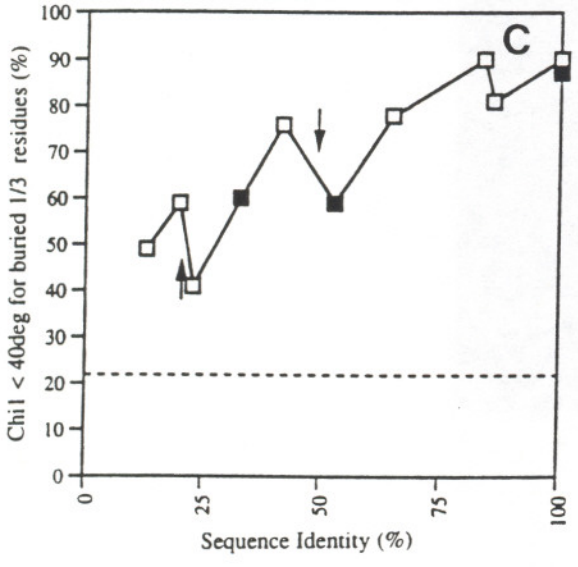
All plots in Fig. 1 display a smooth, monotonic decrease in prediction accuracy with decreasing similarity between the template and the target reference structures. Overall, the data points from the modeling of both the DNA–binding protein family and the globin family independently confirm the same general trend in all plots. That is, eliminating the data points from either family would leave essentially the same generally smooth, monotonic relationships. Nevertheless, in each plot, against the backdrop of the overall trend, the three DNA–binding protein data points display greater variation than the globin data points. Presumably, this reflects familial variation, particularly those relating to the inherent statistical noise associated with the modeling of the much smaller number of residues in the DNA–binding proteins (20 or 30 buried residues) as compared to the globins (50 or 80 buried residues). All other things being equal, with the exception of the side– chain r.m.s. measure of prediction accuracy (Fig. 1A, 1B, 1G, 1H), the 6 plots (Fig. 1A–1F) corresponding to the prediction of the smaller set of more–buried residues

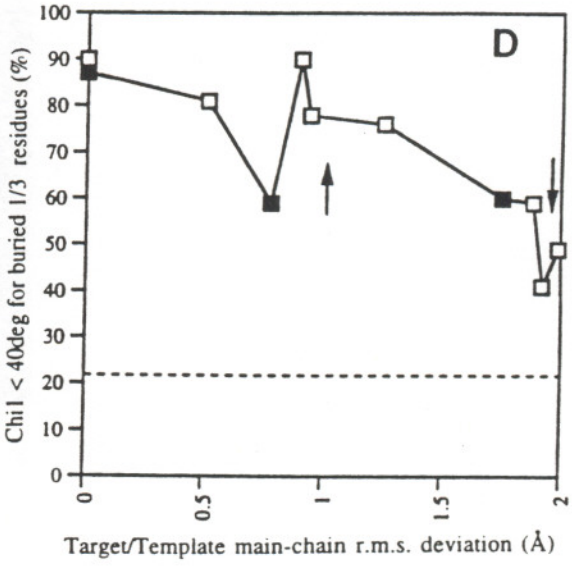
Fig. 1. Summary of sidechain r.m.s. errors and accuracy of torsional $\chi 1$ and $\chi 2$ angle predictions of the buried residues as function of sequence identity and backbone r.m.s. deviation for target/template protein pairs in modeling studies (Table 1).

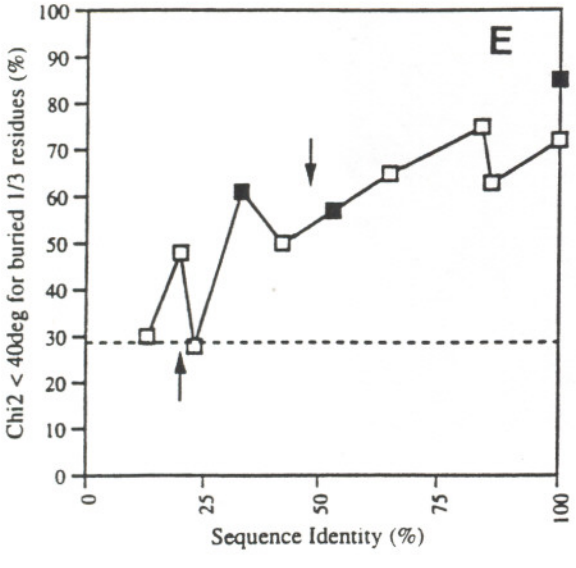
(A–F) is for the more buried residue set (1/3) and (G–L) is for the less buried residue set (1/2). The open boxes represent globin modeling data points and the filled boxes represent the average values of the repressor modeling data points. The horizontal dotted line in each panel indicates the corresponding r.m.s. error or χ angle accuracy that can be expected for completely random prediction (i.e., 3.1 Å, for side–chain r.m.s. error, 22% for $\chi 1$ accuracy, and 29% for $\chi 2$ accuracy). In all plots, the two arrows indicates the two regions of particular interests discussed in the text. The first arrow, at either 22% sequence identity or at 2 Å target/template main–chain r.m.s. error, corresponds to the 'twilight zone' of protein sequence homology. The second arrow, at about 50% sequence identity or at 1 Å target/template r.m.s. error, corresponds roughly to the intermediate transitional zone where the side–chain packing methods give either reliable or only moderately reliable predictions.

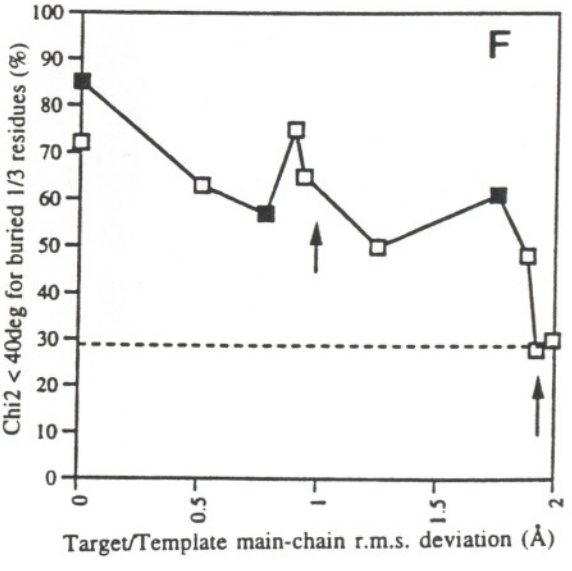

A
Side-chain r.m.s. error for buried 1/3 residues (Å)
Sequence Identity (%)

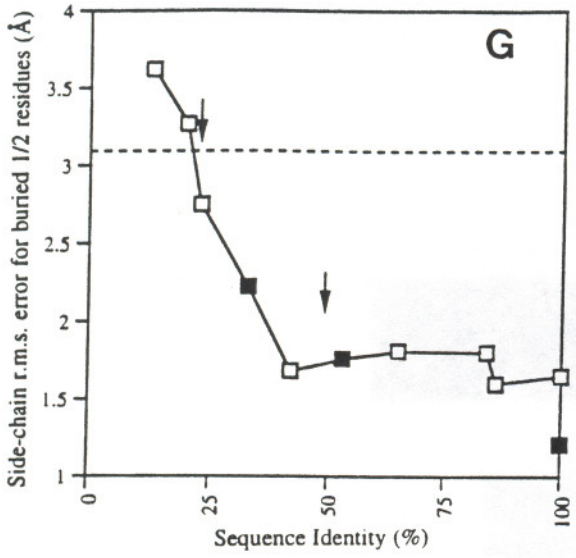

B
Side-chain r.m.s. error for buried 1/3 residues (Å)
Target/Template main-chain r.m.s. deviation (Å)

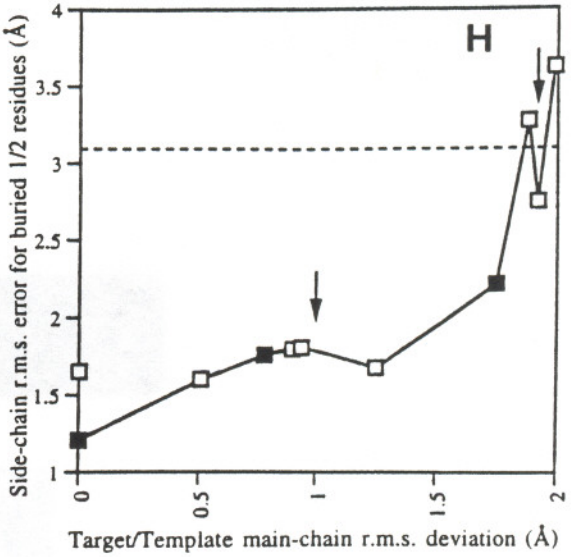

C
Chi1 < 40deg for buried 1/3 residues (%)
Sequence Identity (%)

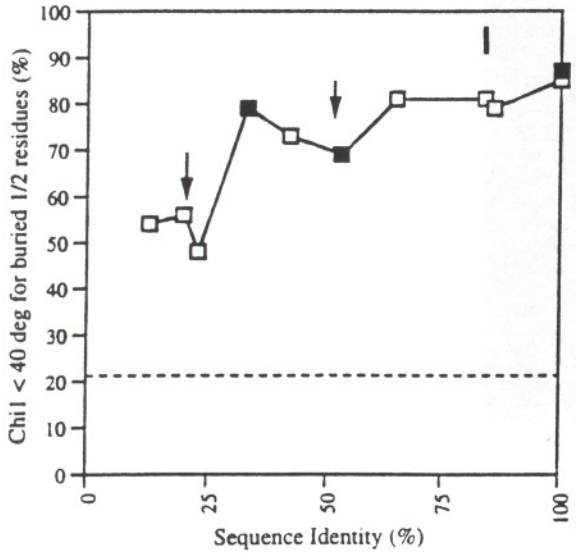

D
Chi1 < 40deg for buried 1/3 residues (%)
Target/Template main-chain r.m.s. deviation (Å)

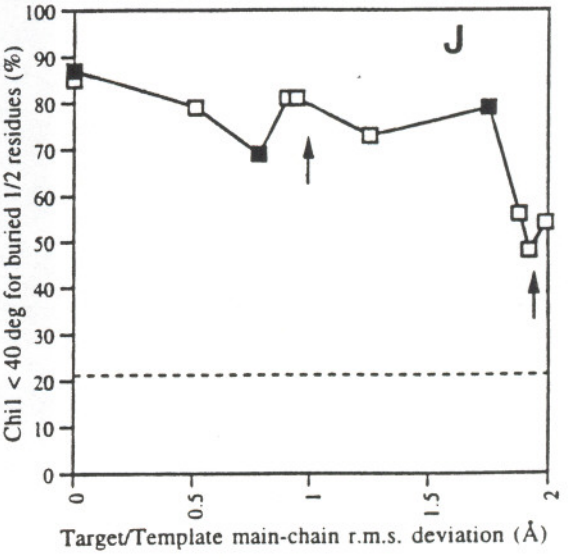

E
Chi2 < 40deg for buried 1/3 residues (%)
Sequence Identity (%)

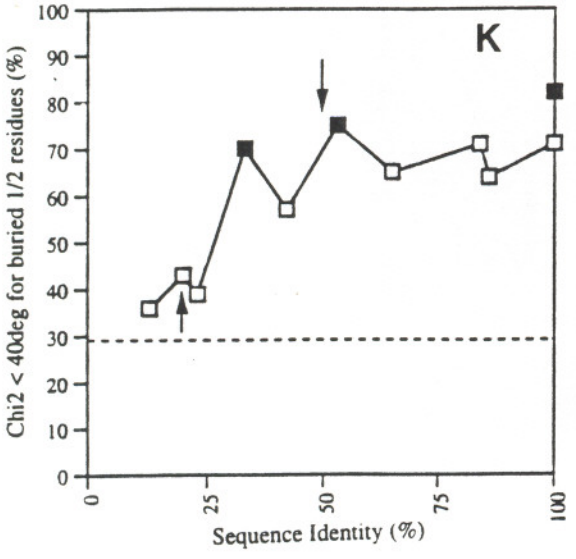

F
Chi2 < 40deg for buried 1/3 residues (%)
Target/Template main-chain r.m.s. deviation (Å)

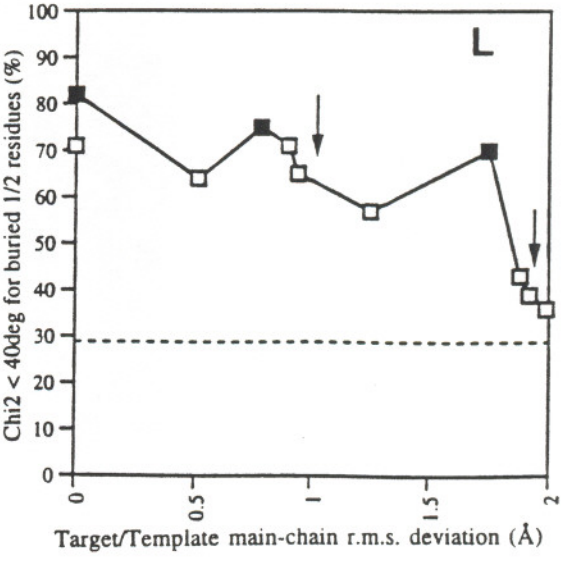

G
Side-chain r.m.s. error for buried 1/2 residues (Å)
Sequence Identity (%)


H
Side-chain r.m.s. error for buried 1/2 residues (Å)
Target/Template main-chain r.m.s. deviation (Å)


I
Chi1 < 40 deg for buried 1/2 residues (%)
Sequence Identity (%)


J
Chi1 < 40 deg for buried 1/2 residues (%)
Target/Template main-chain r.m.s. deviation (Å)


K
Chi2 < 40deg for buried 1/2 residues (%)
Sequence Identity (%)


L
Chi2 < 40deg for buried 1/2 residues (%)
Target/Template main-chain r.m.s. deviation (Å)

(buried 1/3 residues) are generally very similar to their 6 counterparts for the less–buried set (buried 1/2 residues, Fig. 1G–1L). This is not surprising, since the other measures of prediction accuracy, that of $\chi 1$ and $\chi 2$, are designed on the basis of a single, somewhat arbitrarily selected cutoff of 40 degrees for judging similarity, while the side–chain r.m.s. deviation measure is a more continuous one. Thus, on the basis of a single cutoff value, both $\chi 1$ and $\chi 2$ are predicted about as well for the more–buried set as the less–buried one (Fig. 1C–1F, 1I–1L). This is true over all ranges of both sequence identityand target–template r.m.s. deviation similarity. In contrast, the side–chain r.m.s. error, particularly when the target–template similarity is relatively high, clearly suggests that the more–buried residues are better predicted (Fig. 1A, 1B, 1G, 1H). For the more–buried residue set (buried 1/3 residues), side– chain r.m.s. error increases from about 1.2 Å when the available template is perfect, to about 1.5 Å when the target–template sequence identity decreases to 50% (Fig. 1A). The same range of side–chain r.m.s. error, 1.2 Å to 1.5 Å, is traversed when the target–template similarity measured in terms of the target–template backbone r.m.s. deviation increases from 0 Å to about 1 Å (Fig. 1B). For the less–buried residue set (buried 1/2 residues), side–chain r.m.s. error is generally higher, and rises from about 1.4 Å to about 1.75 Å for the same decrease in sequence identity from a 100% to 50% (Fig. 1G). This relatively poorer range in side–chain r.m.s. prediction accuracy for the less– buried residue set is, as before, a result of the template models becoming progressively poorer from 0 Å to 1 Å (Fig. 1H). However, when the similarity between template and target decreases further, the side–chain r.m.s. error for both the more–buried and the less–buried residue sets steadily approaches the same random expectation of 3.1 Å, and accordingly, the difference in prediction quality between the two sets vanishes. This value of 3.1 Å is exceeded when the sequence identity reaches about 22%. It is worth noting that the 20–25% sequence identity range is the so–called 'twilight zone' of sequence alignment below which the level of sequence identity is insufficient to unambiguously reflect an evolutionary relationship between a pair of sequences (31, 32). While there are some well–known cases where significant structural homology (backbone r.m.s. deviation < 1.5 Å) is seen for a sequence identity as low as 1%, in the main and particularly below 10%, these do not reflect divergent evolution but rather convergent evolution (5, 33). Thus, the twilight zone is an empirically observed region where the case for divergent evolution is unclear and in the absence of other biologically relevant information, the case for convergent evolution is plausible. So it is intriguing that the inability for the template backbone to constrain the side–chains and so result in a random prediction should occur precisely in this twilight zone of sequence homology. As can be seen from the plot of sequence identity against backbone r.m.s. deviation for our protein families (Fig. 2), this twilightzone value corresponds to a little less than 2 Å r.m.s. deviation in the backbone coordinates. Interestingly, an intermediate sequence identity value of 50% corresponds to an intermediate backbone r.m.s. of about 1 Å. Both these specific sequence identity/backbone r.m.s. deviation correspondences, as well as the overall curve for the repressor/globin families, are very similar to the same plot using a much larger number of protein families (3–4).

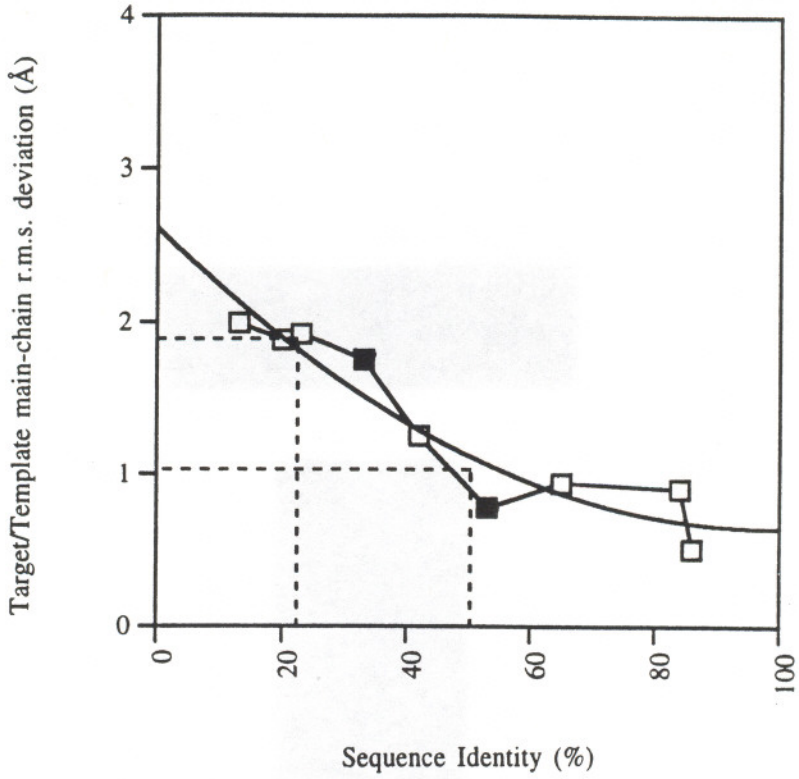


Fig. 2. Relationship between sequence identity and backbone r.m.s. deviation for the target/template protein pairs in modeling studies (Table 1).

The open boxes represent globin modeling data points and the filled boxes represent the average values of the repressor modeling data points. The two vertical dotted lines indicates the two regions of particular interests discussed in the text. The 'twilight zone' of protein sequence homology is shown by the first vertical dotted line at 22% sequence identity, corresponding to about 2 Å target/template main–chain r.m.s. error. The second vertical line, at about 50% sequence identity, corresponding to about 1 Å target/template r.m.s. error, reflects roughly to the transitional region where the sidechain packing methods give either reliable or only moderately reliable predictions.

Turning to the χ angle measures, as stated before, there is no significant difference in results between the two sets of buried residues considered. Therefore, all further discussion of χ 1 angle prediction accuracy will implicitly refer to values that have been averaged over the two burial sets. Prediction accuracy for the $\chi 1$ angles range from 85% when the template is perfect to about 50% when the sequence identity is at the twilight zone of 22% or when the target–template r.m.s. is about 2 Å (Fig. 1C, 1D, 1I, 1J). At the intermediate value of 50% sequence identity, the prediction accuracy is about 70%; this similar to the 75% accuracy that can be obtained with the intermediate target–template r.m.s. of about 1 Å.

Similarly, the $\chi 2$ angles can be predicted to some 80% accuracy with perfect templates, but when the sequence identity approaches the twilight zone, it can only be predicted to about 30%, which is virtually identical to the 29% that can be expected for a random prediction (Fig. 1E, 1F, 1K, 1L). When the target–template r.m.s. deviation is 2 Å, the prediction accuracy of 35% is not quite at this random level, suggesting that the backbone is only truly insufficient to constrain the side–chains when it deviates a little more than 2 Å. At the intermediate sequence identity of 50% the $\chi 2$ prediction accuracy is 60%. At the intermediate level of target–template similarity of 1 Å, the prediction accuracy is a shade better at 65%. Overall, and perhaps not surprisingly, the $\chi 1$ angles are better predicted than the $\chi 2$ angles for similar levels of dissimilarity between the target and template. Interestingly, while $\chi 1$ is better predicted by 5% for perfect templates, it is better predicted by some 10% at the intermediate levels of similarity between target and template (i.e. at 50% sequence identity or 1 Å target–template r.m.s.). By the time the twilight zone is reached or target–template deviation exceeds 2 Å r.m.s., this discrepancy widens to some 20%. This trend suggests that relative to $\chi 2$ the $\chi 1$ prediction accuracy is more resilient to progressive mutation and increasing distortion of the backbone. This would imply that the template backbone places stronger constraints on the $\chi 1$ angle compared to the $\chi 2$ angle. Obviously, this could have been expected, since it is well known that the close proximity of the backbone restricts the range of side–chain rotamer choice of $\chi 1$ more than $\chi 2$. This observation explains the relative difference in behavior of $\chi 1$ and $\chi 2$ at the twilight zone in which $\chi 1$, unlike $\chi 2$ or the side–chain r.m.s. error, is significantly better predicted than random. Having accounted for the $\chi 1$ accuracy at the twilight zone in terms of a local constraining effect of the backbone, it is highly noteworthy that $\chi 2$, like the side–chain r.m.s. error, converges to its random expectation at precisely the twilight zone of sequence identity. This is particularly curious, since any constraining effects of the backbone on both $\chi 2$ and the side–chain r.m.s. deviation would involve interactions that are more tertiary in nature than that between $\chi 1$ and its proximal backbone.

In summary, our modeling studies conducted over a set of proteins that vary in function, size, and sequence. Other structural variations included the presence and absence of large prosthetic groups and relative deletions and insertions of secondary structure. Despite these variations, our results clearly establishes an approximately smooth, monotonic decrease in side–chain prediction accuracy with decreasing similarity between the template structure and the target that is to be modeled. The quality, consistency and general predictability of this relationship allow us to generalize some guidelines for the homology–modeler hoping to use side–chain packing methods to attain reliable models. Given that one wants to obtain a structural model of a target sequence using the sequence and structure of a homologous template protein these rules are:

1) When sequence identity is at least 50% or when there is reason to believe the backbones will not differ by more than 1 Å r.m.s., the buried side– chains can be predicted relatively accurately. The most buried third of residues can be predicted to better than 1.5 Å r.m.s. and the most–buried half to better than 1.75 Å. For both levels of burial at least 70–75% of $\chi 1$ angles and 60–65% of $\chi 2$ angles can be predicted correctly.

2) When sequence identity is below 50% or when there is reason to believe the backbones will differ by greater than 1 Å r.m.s., the side–chain predictions are still better than random and only reach randomness (i.e. 3.1 Å side–chain r.m.s. error, 22% accuracy on $\chi 1$ and 29% accuracy on $\chi 2$) when sequence identity drops to 22% (the twilight zone) or the backbones differ by more than 2 Å r.m.s. Thus, even if somewhat limited, homology modeling in this range could prove useful to the experimentalist.

3) In general, the relationship is not linear. When the target–template similarity is relatively high (i.e. greater than 50 % sequence identity or less than 1 Å for the target/template r.m.s. deviation) decreasing similarity does not hurt modeling accuracy as much as when the target–template similarity is relatively low (i.e. less than 50 % sequence identity or greater than 1 Å target/template r.m.s. deviation).

4) When sequence identity is less than 22% or if the expected backbone difference is greater than 2 Å, there is little point in using side–chain packing methods that enforce a fixed template backbone to produce homology models since such backbone templates are insufficient to constrain the packing orientations of the buried side–chains. Other homology modeling approaches could be used (34–36) or new packing methods that allow the backbone template to move could be developed.

Acknowledgments

We thank Enoch Huang, Jerry Tsai, and Michael Levitt for helpful discussions. We thanks professor Michael Levitt for providing computer programs

and computing facilities. This work was supported by NIH grant GM41455 to ML and USUHS grant R071CX to SYC.

The opinions or assertions contained herein are the private ones of the authors and are not to be construed as official or reflecting the views of the Department of Defense or the Uniformed Services University of the Health Sciences.

References

1. Ponder, J. W. & Richards, F.M.(1987). J. Mol. Biol. **193**:775–791.
2. Chung, S. Y. & Subbiah, S.(1995). Protein Sci. In press.
3. Chothia, C. & Lesk, A.M. (1986). EMBO J. **5**:823–826.
4. Flores, T.P. Orengo, C.A., Moss, D.S., & Thornton, J.M.(1993). Protein Sci. **3**:2358–2365.
5. Subbiah, S., Laurents, D., & Levitt, M.(1993). Current Biology **3**:141–148.
6. Lee, C. & Subbiah, S.(1991). J. Mol. Biol. **217**:373–388.
7. Desmet, J., De Maeyer, M., Hazes, B., & Lasters, I. (1992). Nature **356**:539–542.
8. Dunbrack, Jr., R.L. & Karplus, M.(1993). J. Mol. Biol. **230**:543–574.
9. Eisenmenger, F., Argos, P., & Abagyan, R. (1993). J. Mol. Biol. **231**:849–860.
10. Holm, L. & Sander, C.(1991). J. Mol. Biol. **218**:183–194.
11. Koehl, P. & Delarue, M.(1994). J. Mol. Biol. **239**:249–275.
12. Lee, C.(1994). J. Mol. Biol. **236**:918–939.
13. Tuffery, P., Etchebest, C., Hazout, S., & Lavery, R.(1991). J. Biomolec. Struct. & Dyn. **8**:1267–1289.
14. Wilson, C., Gregoret, L.M., & Agard, D.A.(1993). J. Mol. Biol. **229**:996–1006.
15. Tanimura, R., Kidera, A., & Nakamura, H.(1994). Protein Sci.**3**:2358–2365.
16. Bashford, D. Chothia, C. & Lesk, A.M.(1987). J. Mol. Biol. **196**:199–216.
17. Gerstein, M., Sonnhammer, E.L.L., & Chothia, C.(1994). J. Mol. Biol. **236**:1067–1078.
18. Bernstein, et. al. (1977). J. Mol. Biol. **112**:535–542.
19. Kuriyan, J. & Petsko, G.A. (1986). J. Mol. Biol. **192**:133–154.
20. Hubbard, S.R., Hendrickson, W.A., Lambright, D.G., & Boxer, S.G. (1990). J. Mol.Biol. **213**:215–218.
21. Evans, S.V. & Brayer, G.D. (1990). J. Mol. Biol. **213**:885–897.
22. Petruzzelli, R, Aureli, E., Casale, M., Nardini, M., Rizzi, M., Ascenzi, P., Coletta, M., De Santis, G., & Bolognesi, M. (1993). Biochem. Mol. Biol. Int. **31**:19.
23. Birnbaum, G.I., Evans, S.V., Przybylska, M., & Rose, D.R. (1994). Acta Cryst. **D50**:283–289.

24. Bolognesi, M., Onesti, S., Gatti, G., Coda, A., Ascenzi, P., & Brunori, M. (1989). J. Mol. Biol. **205**:529–544.
25. Royer, Jr., W.E. (1994). J. Mol. Biol. **235**:657–681.
26. Hackert, M., Kolatkar, P., Ernst, S.R., Hackert, M.L., Ogata, C.M., Hendrickson, W.A., Merritt, E.A., & Phizackerley, R.P. (1992). Acta Crystallogr., sect. B. **48**:191–202.
27. Mondragon, A., Wolberger, C., & Harrison, S.C.(1989). J. Mol. Biol. **205**:179–188.
28. Mondragon, A., Subbiah, S., Almo, S.C., Drottar, M., & Harrison, S.C.(1989). J. Mol. Biol. **205**:189–200.
29. Sevilla–Sierra, P., Otting, & Wuthrich, K.(1994). J. Mol. Biol. **235**:1003–1020.
30. Miller, S., Janin, J., Lesk, A.M., & Chothia, C.(1987). J. Mol. Biol. **196**:614–656.
31. Doolittle, R.F.(1986). Of Urfs and Orfs: Primer on how to analyze derived amino acid sequences, University Science Books, Mill Valley, CA.
32. Sander, C. & Schneider, R. (1991). Proteins: Struct. Func. & Genetics. **9**:56–58.
33. Russel, R.B. & Barton, G.J. (1994). J. Mol. Biol. **244**:332–350.
34. Levitt, M. (1991). J. Mol. Biol. **226**:507–533.
35. Greer, J. (1991). Methods Enzymol. **202**:239–252.
36. Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. & Thornton, J.M. (1987). Nature **326**:347–352.