The use of position specific rotamers in model building by homology.


* Center for Genetic Engineering and Biotechnology, PO BOX 6162, Havana, Cuba.
** EMBL-EBI, The EBI, Hinxton Hall, Cambridge, UK.
*** EMBL-Heidelberg, Meyerhofstrasse 1, Heidelberg, Germany.

This was an early draft of this article. The final version can be found in the November 95 volume of PROTEINS

Abstract

In this study we concentrate on replacing sidechains as a sub task of model building by homology. Two problems arise. How to select probable rotamers? And how to avoid the computational explosion that results from the combination of many residues for which multiple good rotamers are predicted?

We use position specific rotamers rather than standard rotamer libraries, and by sorting the residues that have to be modeled as a function of their entropy in rotamer space, the combinatorial problem can be overwon.

If the backbone of the homologous protein is sufficiently similar to the one to be modelled, equally good or sometimes even better models can be built than with most other methods.

The advantages of our method are the high quality of the models in case of high backbone similarity, the small amount of human intervention needed, and the fact that the method automatically estimates the reliability with which each method got modelled.

The major drawbacks are the poor modeling in case of backbone rearrangements between structure and model, and the fact that insertions can not be modelled at all.

Introduction

The process of model building by homology normally follows distinct steps:

- A sequence alignment between the protein to be modeled and a protein of known structure has to be established. Obviously, the higher the percentage identity observed in this alignment, and the lower the number of insertions and deletions, the easier the modeling task will be, and the better the final result.
- Based on this alignment a backbone has to be generated. This is normally the backbone of the most homologous structure, but a hybrid backbone can be chosen too.
- Sidechains have to be placed in the model. Normally it is decided not to alter the sidechains of the conserved residues. When placing sidechains in the model the major problem is that many residues have multiple favourable side chain conformations (rotamers). Often rotamers of more than one residue can occupy the same space. Complicated techniques like Monte Carlo procedures, tree searching algorithms etc., have been described to solve this combinatorial problem.
- If the protein to be modeled has insertions or deletions with respect to the known structure, loops have to be re-modeled, or modeled ab initio. Database searches for loops with similar anchoring points in the structure are often used to build these loops, but energy based ab initio modeling techniques have also been employed.
- The final model needs to be optimized. Energy minimizations, sometimes combined with molecular dynamics, are normally used for final structure optimization.
The quality needs to be estimated. Several techniques have recently been described that can be used to estimate the quality of protein models, but visual inspection is still required to verify that the structural aspects of the model are not contradicting what is known about the functional aspects of the molecule.

Correctly placing new sidechains in a known structure is one of the key steps in model building by homology. It is relatively simple to place one or a few new sidechains, but when sidechain conformations have to be predicted for many residues in parallel, the choice between just a few seemingly good possibilities at each position quickly leads to combinatorial problems.

In this article we will concentrate on the problem of crafting new side chains onto the backbone scaffold of the known structure. We introduce two new techniques.

- Usage of position specific rotamers rather than a standard rotamer library improves the quality of positioning side chains.
- Careful selection of the order in which the residues are added to the model can overcome the combinatorial problem.

The combination of these two techniques leads to a fast and robust modeling method. If the similarity between the backbone of the structure and the backbone of the model is sufficiently high, models can be produced which are equally good - as judged by all atom root mean square deviations (RMSD) - as models produced with other techniques.

**Position specific rotamers**

In model building by homology one needs to replace one residue by another. Suppose that there is adequate space, would the introduced residue be happy in its new environment, or would it only fit with an energetically very unfavorable sidechain conformation? More general, which are the preferred rotamers for a certain residue type at a certain position in the molecule? Several studies have been performed on this topic and our understanding of protein structure has been greatly enhanced by their results. Residues tend to cluster their chi-1 torsion angle around 60 (g-), 180 (t), and 300 (g+). Depending on the local secondary structure the population of these three states can differ enormously. For example, g- is hardly ever observed for any residue in an alpha helix, except for serine. Most residues in beta strands show the strongest preference for g+, except for valine which prefers t. Most of these rotamer preferences are easily understood from an analysis of local sterical effects. Rotamer libraries have been created with several levels of sophistication, and they have shown great value for modeling proteins. Especially reconstruction of a whole molecule from only alpha carbons, an exercise often met in protein structure prediction and protein design, is often done using a rotamer library.

The main problem with rotamer prediction is that the body of solved protein structures is still too small to fully extract all rules needed to uniquely determine the conformation for each residue. We are therefore forced to neglect certain aspects of protein structure. In a previous study we analyzed wildtype and mutant protein structures and asked how much the local backbone conformation determines the side chain conformation. Position specific rotamer distributions were determined while residues distant in the sequence and all nearby side chains were neglected. Surprisingly, more than 80% of all mutant structures could be predicted correctly, and in the other 20% no prediction rather than a wrong prediction was made in the majority of all cases. This indicates that position specific rotamer distributions contain more information and less noise than averaged distributions as present in standard rotamer libraries.

A rotamer distribution for a certain residue type at a certain position, called a position specific rotamer distribution, is determined by extracting from a database of non-redundant protein structures (Hobohm et al., 1992) all suitable fragments of 5 or 7 residues (7 in helix and strand, 5 in case of irregular local backbone). Suitable fragments are those that have a local backbone conformation similar to the one around the evaluated position, and have the same residue type at the central
position. In these analyses, the RMS deviation of the alpha carbon positions between the structure and the database fragment was kept below 0.5 Angstrom. In this study rotamers were rejected if they would lead to severe Van der Waals' clashes when placed in the model. A severe Van der Waals' clash is defined as the distance between two atom being shorter than the sum of the Van der Waals radii minus 0.5 Angstrom.

The database fragment extraction technique has been described before. It is based on a distance geometry loop search algorithm. For the two N- and C-terminal residues other than the central residue of the database fragments have to be used.

In the modeling process, position specific rotamers are used twice. They are first used to determine the entropy in rotameric space of a residue, and thereafter to decide how to place the side chains.

**Rotameric entropy**

If many new side chains have to be placed in the model, multiple side chains can potentially occupy the same space. A big problem for most modeling methods is that they are based on an energy function that includes contacts between residues distant in the sequence, but close in space. This implies that the whole molecule needs to be build before any selected rotamer can be evaluated. This leads to a "chicken and the egg" problem. In order to place a residue correctly, all other residues have to be placed correctly first, so, where to start? Several techniques have been described to overcome this problem. Monte Carlo procedures seem the most appropriate for this purpose, but Hazes et al., already indicated that other solutions might exist.

![Figure 1](image.png)

Fig. 1. Example of two overlapping rotamer distributions.

Figure 1 shows several examples of position specific rotamer distributions. Figure 1A shows an example where the position specific rotamer distribution is extremely narrow. If such a case would occur in a modeling study, this residue should be modeled immediately, and never looked at again. Figure 1F shows an example of a very wide rotamer distribution. Such a residue should obviously be modeled late in the modeling procedure, because it can much more easily adapt to the space left to it after all other side chains have been placed.
We base our modeling strategy on simple probability principles. The narrower the rotamer distribution, the higher the probability that this is the rotamer needed in the structure to be modeled. To quantify rotamer distributions we define a rotameric entropy. In figure 2 some examples of rotamer distributions are shown and the derived rotameric entropies are given. The rotameric entropy is defined by $E = \frac{P \times P_{\text{tot}}}{F \times F_{\text{tot}}}$ in which $P$ is the population with chi1 within 45 degrees of the most populated of the three standard chi1 values (60, 180, 300 degrees), $P_{\text{tot}}$ is the sum of all rotamers that fall within 45 degrees of any of these three standard chi1 values, $F$ is the total number of rotamers in this distribution, and $F_{\text{tot}}$ is the maximal number of rotamers obtainable for any distribution (we chose $F_{\text{tot}} = 40$ in this study). The normalisation with respect to $F_{\text{tot}}$ avoids that residues for which only very few position specific rotamers are found get a large value.

In the modeling process first a sparse model was generated. In this sparse model all not conserved residues were left untouched, but other residues were mutated into alanine, unless they had to become glycine or proline. Glycines and prolines were placed immediately, without the use of position specific rotamers. No backbone adaptations were made at this stage. For all alanines that subsequently had to be mutated the rotameric entropy in this sparse model was determined. Side chains were then placed in order of decreasing rotameric entropy.

**Placing side chains**

For most residue positions 40 rotamers could be extracted from the database. Per position all observed rotamers were scored using the function:

$$S = w_1 \frac{P_{\text{loc}}}{P_{\text{tot}}} - w_2 R_f - w_3 f_p + w_4 Q + w_5 H - w_6 B$$
S score for one rotamer,
Ploc population in the chi1 bin of this rotamer (see above),
Rf RMS deviation of the fit of the database Cas on the corresponding Cas in the backbone of the model.
fp function of the difference between phi-psi (fp) in the structure and in the database hit. fp is zero if both phi and psi are < 30 degree wrong. fp is a quadratic function of the errors in phi and psi which becomes the dominant term when phi or psi is more than 90 degrees wrong.
Q quality of packing (DACA score)
H number of hydrogen bonds formed
B function of the Van der Waals' clashes of the side chain
w1-6 weights that are set such that all six terms on average contribute equally much to the spread in S.

The rotamer with the highest S is selected, and put in the model. The first three terms are dominated by information from the local backbone, the terms 4-6 are dominated by information about contacts with residues further away in the sequence.

The combination of this scoring scheme with the sorting of residues as function of the rotameric entropy has an additional advantage. Early in this modeling process the residues are built that have a very narrow rotamer distribution, which indicates that the conformation is mainly determined by the local backbone, and the absence of many not yet modeled residues is not a disadvantage. Residues with wider rotamer distributions which therefore are more influenced by the rest of the molecule are built later when more residues are already completed.

We used the following modeling protocol:

- Create the sparse molecule;
- Determine the rotameric entropy for the residues to be modeled;
- Mutate as described above in order of decreasing entropy; After the completion of this model several possibilities exist to fine-tune the model. Energy minimisations are routinely used as the last step in modeling experiments. We tried several optimisation protocols but got the best result using the following steps:
  - Patch up the altered residues. One by one all residues that got altered are mutated into alanine, and directly afterwards mutated back into the correct residue, using the mutation protocol described above. This way the environment of the residues is more complete than in the previous step. Most times this results in only slightly different solutions for the rotamer selection problem, but sometimes it leads to rather different rotamer choices, especially for surface residues. This round uses the same order of residue positions as the first ala->X mutation round.
  - Do a brute force search in torsional space to get rid of severe Van der Waals' clashes. In this search all side chain torsion angles of all residues in the model are free to change up to maximally 5 degrees.
  - Do a second run in which all torsion angles of all residues that still show clashes are left completely free.
  - Do a short (100 steps) energy minimization in vacuum (using GROMOS)

Testing the method

The methods were tested in several different ways. The most realistic test was the Asilomar meeting, but in the months before we have been building models of known structures, using properly Jack-knifed databases. For this purpose a set of pairs of proteins was collected that is representative for present contents of the PDB. Roughly equally many all-alpha, all-beta and mixed type protein pairs were selected. We made sure that several molecules had co-factors, and that our test cases were spread evenly over the 35% - 95% sequence identity range. We realised that reliably modeling of loops is not yet possible without high homology, and therefore concentrated only on side chain modeling completely disregarding all insertions or deletions. The article on model quality gives a
good overview of the test set that we used.

As a second test we converted several proteins into poly-alanine, and modeled them back to the original molecule using the protocol described above. The subset of the WHAT IF database used for the reconstruction never held homologs to the molecule being modeled. In contrast to the previous test case, there are no conserved residues (except a few alanines, glycines and prolines), but the backbone is perfect. Models reconstructed this way had a lower all atom RMSd than models built by homology. This is not so surprising, because all backbone atoms are perfect. However, the sidechain atom RMSd was better too, (despite the larger number of side chains that had to be modeled) and the number of core residues with chi1 more than 45 degrees wrong was about the same. This clearly indicates the importance of the quality of the backbone coordinates for our modelling strategy.

We tried to optimize our modeling protocol by systematically changing the parameters and some of the algorithms. We tried to arrive at a single number that could be an index of quality of a model. The RMSd seems the natural candidate for this index. The following examples explain why this index is not very good:

- It is often not certain whether the experimentally determined 3D structure represents the truth. In many cases multiple side chain conformations exist, but only one is experimentally observed. After all, an energy difference of only 0.1 Kcal/mole between two conformations will already lead to a difference in population of 8:1. With such a population difference crystallography normally can only detect 1 conformation. If we model the lesser populated rotamer, this residue would be called modeled incorrectly, despite that the error is only 0.1 Kcal/mole. Filippis et al., showed an example were two likely rotamers were found in a modeling study. These two conformations were indeed observed in the crystal structure, and deposited as two separate PDB files. Also, the hydrogen bonding pattern of a molecule can only be determined correctly if all crystal waters are available. We cannot yet model water positions, and subsequently residues for which the sidechain conformation depends on the presence of water molecules are likely to be modeled incorrectly. An additional problem is that protein structures contain many errors. For example, more than 10% of all asparagines, glutamines and histidines in all files in the PDB need a 180 degree rotation in one of the side chain torsion angles to make proper hydrogen bonds (The HPR structure from which some examples are shown below contained only one such case: glutamine 88, and is in this respect one of the best files in the PDB).

- Figure 3 shows, as an example, residue 11 in HPR in the true structure and in the model. In the monomer, this residue sticks out into the solvent, and makes neither in the structure, nor in the model any intra molecular contacts. The large difference between the structure and the model is therefore irrelevant, but has a large impact on the RMSd.

- The reason why lysine 11 is sitting this way is crystal contacts. All atoms in this residue make contacts with a neighbour molecule in the crystal.

- A typical modeling experiment in which we model one known structure from the other would give the following results: RMSd for all core atoms: 0.8A; RMSd for residues involved in crystal contacts either in of the two structures: 1.8A; RMSd for the other surface residues: 1.5A.
Based on these facts we decided to use the RMSd on all atoms as model quality index, but to exclude atoms in the sidechains of residues that were hyper exposed (i.e. did not make intra molecular contacts) or that were involved in symmetry contacts.

One problem with our approach is the assumption that conserved residues have a conserved conformation. This assumption is not always correct. However, using this assumption leads to better models because the number of exceptions to this rule is a factor of two smaller than the number of residues that were modeled incorrectly in a testcase where we did not make this assumption. Once more and better data is available this assumption will probably be abandoned.

**Discussion**

What did we learn from Asilomar, what went right, and what went wrong? As you can see in Mike James' contribution in this same volume, we participated with three models. These three models were made fully automatically, only using standard procedures incorporated in the WHAT IF program. Every WHAT IF user would arrive at these same models. Manual intervention only took place in three cases in the CRABP model because the program requested this.

In the CRABP case the alignment was difficult, we got it wrong, and that is the end of the discussion. If a sidechain is placed in a wrong conformation, it can be rotated to a better position later during the modeling process, but the possibility to modify the alignment during the modelling has not yet been implemented.

In the HPR project we were lucky, no insertions or deletions, and a reasonably high backbone similarity. Consequently, WHAT IF did reasonably well. However, there is still a lot to be improved.
Figure 2 shows some of the advantages and disadvantages of our method. The loop around histidine 15 shows errors up to 2.0 Å in the model. We cannot predict such loop displacements. Here the changes are caused by intensive crystal contacts in the HPR structure that are not present in the structure of the 2HPR file on which we based our model, but often such loop displacements are an intrinsic property of the structure, and ought to be predictable. The aspartic acid at position 10 shows two preferred rotamers in the position specific rotamer distribution. The most populated one is the correct one, and thus this residue gets modeled correctly. There is only one preferred rotamer for valine 12, and this is also the conformation found in the real structure. However, this rotamer is also the preferred one in all existing standard rotamer libraries, and would have been modeled correctly with every procedure.

Leucine 86 sits 'the wrong way around' (chi 1 40 degrees off and chi 2 90 degrees off) in the model (see fig 5). This leads to a large RMSd in the twoCd atoms, however, the Cd atoms in the model and in the structure occupy the same space. Both Cd atoms make the same contacts in the model and the structure, and although every residue should ultimately be modeled correctly, fixing problems such as these does not get our highest priority.

NDK was modeled reasonably well, but we could not predict the reorganization of one surface loop. That meant that the total RMS got bad, despite that most of the model actually was done rather well by WHAT IF. The unfortunate thing is that one cannot know this in advance, and thus, with every model one should always be prepared for the worst....
The aforementioned modeling protocol is the best we could derive so far. Several alternative protocols were attempted, but they either did not improve the models, or the improvement was too small to warrant the extra CPU time needed.

Conserved residues do not always have the same rotamer in two molecules. The largest RMSd values for conserved residues were seen for methionines and lysines. We attempted to treat these residues as if they were not conserved, i.e. convert them to alanine like all residues that had to be changed, and mutate them back as if they were not conserved. However, this made the RMSd between the models and the real structures on average a little bit worse.

The rotameric entropy is based on rotamers that could be modeled given the sparse model. After putting several of the low entropy side chains into this sparse model, the rotameric entropy of the residues still left to be modeled can have changed. Periodic updating of the rotameric entropy (e.g. after modeling 25%, 50% and 75% of all residues to be modeled) did not significantly improve the RMS deviation between the models and the real structures, but costs much CPU time.

The best results are obtained by a very short energy minimization. Several energy minimization protocols were tried. Longer energy minimizations, or energy minimizations combined with short low temperature molecular dynamics runs (all in vacuum) normally increased the RMSd. We have not yet tried to perform molecular dynamics runs in water.

Where do we stand today, and where should we go from here? Where do the errors come from, and what needs most urgently to be improved?

There are several main areas for improvement. The present techniques probably still can be improved. Techniques need to be developed to predict alterations in the backbone, techniques need to be developed to model insertions, and we need more and better techniques to detect errors in models. But most importantly, we need a better understanding of the why and how of protein structure so that better force fields for energy minimization can be developed, and even better models can be
presented at the next Asilomar protein structure prediction meeting.

For some silly reason, the references got lost. If you really need them, all the important references are available in the review in this same article series.