THE NATIVE SEQUENCE DETERMINES SIDECHAIN PACKING IN A PROTEIN, BUT DOES OPTIMAL SIDECHAIN PACKING DETERMINE THE NATIVE SEQUENCE ?

P. KOEHL
UPR 9003 du CNRS, Boulevard Sebastien Brant, 67400 Illkirch Graffenstaden, France.

M. DELARUE
Laboratoire d'Immunologie Structurale, Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France.

Globular proteins have highly compact structures and the corresponding packing interactions are widely considered as the principal determinant of the native structure. It is therefore important that theoretical approaches to protein design explicitly take in account packing, which requires that a full atomic representation of the designed protein is maintained. As a first step towards this goal, we have developed in this report an inverse folding algorithm with the aim of specifically designing amino acid sequences which optimise sidechain packing for a given protein fold. The design is performed by a global Monte Carlo optimisation in sequence space, with constant amino acid composition and a full-atom representation of the various protein models. Packing is defined by a Lennard-Jones potential. The program was tested by designing stable sequence variants for the chymotrypsin inhibitor fold. The final protein models showed an increase in intramolecular atomic contacts and a decrease in the overall volume compared to the native structure. Starting from the backbone only of the target structure, the algorithm did gradually retrieve reliable though limited sequence information. Higher compatibility might be achieved by improving the potential, however our results suggest that packing interactions are an essential element of a yet-to-be-defined successful energy function for protein design.

1. Introduction

It is the ability of proteins to fold into unique three dimensional structures that allows them to exert their biological function. Hence, comprehending how the amino acid sequence is related to the 3D conformation in the native state is essential to an understanding of biological processes. The direct approach to this problem consists of finding the folded conformation of a protein based on its sequence. Although considerable theoretical as well as experimental efforts have been made in recent years, the attainment of this ultimate goal does not appear imminent. An alternative approach is the inverse protein folding problem, which consists in identifying which sequences are compatible with a given fold[1] (for recent reviews see [2-4]). Elaboration of this alternative view is not only theoretically interesting but is also important for protein design and engineering. This paper focuses on sequence design.

Several global protein sequence design methods have been developed either using lattice models[5-7] in which a systematic search of all possible sequences is tractable (see for example [8]), or only considering mainchain and Cß atoms[9-11]

using simplified pairwise potentials in order to compensate (to a certain extent) for the missing information, such as packing. However, none of these methods can capture detailed atomic interactions and need to be expanded such that they maintain a complete all-atom representation of the designed protein structure. In the early work on sequence design, Ponder and Richards[12] provided the first step towards this goal. Under the assumption that residues in the interior of a protein are the most important in determining its fold, they tested systematically combinations of sidechains fitting in the cores of small proteins, based on steric overlaps, hydrogen bonding and packing density criteria. The number of residues included in the combinatorial search was however limited for practical computing reasons. The same problem was encountered in more recent applications which limited the search in sequence space to residues in the core of the protein[13-16].

Because exposed residues also contribute to the structure and activity of native proteins, they ought to be included in the sequence design process. In this paper, we propose a global protein sequence design algorithm which maintains an all-atom representation. The optimisation procedure is based on a Monte Carlo procedure is sequence space, where random moves are either accepted or rejected using the Metropolis criterion[17]. At each step of the calculation, a chimeric protein based on the known backbone structure and the current designed sequence is built, using our recent fast method for sidechain placement[18]. The internal Van der Waals' energy of this model protein is used to evaluate the compatibility of the sequence with the desired protein fold. As such the strategy is set up to select sequences with optimised sidechain packing within a given 3D framework. It was tested on the compact structure of the chymotrypsin inhibitor of barley seed, PDB[19] code 2CI2. We compared the properties of the corresponding designed sequences with properties of the native sequence, in order to assess the extent to which protein sequences are determined by VdW packing interactions.

Optimisation in sequence space such as that proposed here requires special care, since the only physical competition existing in the actual folding process is between different structures for one sequence. In the procedure described here, we try in fact to mimic natural evolution, which explores different sequences in the framework of one 3D structure, with the difference that this exploration is done at the DNA level, and not directly at the 3D structure level. This difference between folding and sequence design raises the problem of specificity, i.e. the incompatibility of the designed sequences with folds different from the specified one. This is discussed below.

2. Methods

A good strategy for protein design requires that the designed sequence shows a high compatibility with the desired structure while at the same time exhibiting low

compatibility with alternative structures, i.e. the "design in" and "design out" procedures proposed by Yue and Dill[5]. A thorough discussion on this issue was recently presented by Jones[3]. To alleviate this problem we use in this study the approach described by Shakhnovich and Gutin[20] in which the sequence composition is maintained during the sequence optimisation. This sequence constraint was also added to account for the fact that the amino acid composition of a protein is known to be highly dependent on its folding class[21].

Most methods which try to solve the protein folding problem or inverse protein folding problems have two components, i.e. an optimisation algorithm, and a procedure that measures sequence-structure fitness:

2.1 Monte Carlo optimisation in sequence space

Optimal sequences with a given amino acid composition can be obtained through a Monte Carlo simulation in sequence space, following the strategy described by Shakhnovich and Gutin[20]. Starting from a random sequence S0 with the required composition, whose energy is E0, two positions are chosen at random, and the corresponding amino acid types are exchanged. The energy E1 of the new sequence S1 is calculated, and the move is accepted or rejected according to the Metropolis acceptance probability given by[17] :

$$P\{S0 \rightarrow S1\} = \begin{cases} e^{-\frac{E1-E0}{Tmc}} & if \quad E1 - E0 > 0 \\ 1 & otherwise \end{cases} \qquad (1)$$

where Tmc is a parameter usually referred to as the temperature of the Monte Carlo simulation.

2.2 Sequence structure fitness

The coordinates of the backbone atoms (including Cß) of the desired target structure are given as input to our programs. For residue positions occupied by glycine in the target structure, a Cß is built using standard geometry when required. The current sequence during the Monte Carlo simulation is threaded on this backbone, and sidechains are positioned using our approach based on an iterative self consistent mean field theory[18]. Sidechain conformations are selected from a fixed set of rotamers[22], and the selection is based solely on a Lennard Jones potential for VdW interactions; no full energy minimisation is performed at the end to remove possible clashes between rotamers, for sake of computing efficiency. The Lennard-Jones potential of the final model is used to score the compatibility of the corresponding sequence with the target fold.

2.3 Implementation for computing efficiency

Our sidechain placement technique is based on a fixed set of sidechain positions. Since the backbone of the protein is maintained rigid, all possible sidechain-sidechain interactions can be calculated once, for all types of amino acids, yielding a large but efficient energy matrix which can be used at each cycle, both for the mean field optimisation (see [18] for details), as well as for estimating the energy of the model proteins. A complete Monte Carlo simulation over 50000 cycles for a 65 residue protein required 7 hours of CPU time on an IBM 43 P computer.

3. Results

3.1 Designing protein sequences by optimising sidechain packing

The sequence design procedure was tested on the chymotrypsin inhibitor of barley seed (PDB code 2CI2; 65 residues). The Monte Carlo simulation was run over 50000 cycles, with Tmc chosen such that the final acceptance ratio is 20% (Tmc = 1). The initial sequence was chosen to be a random reshuffling of the native sequence. The evolution of the VdW energies of the protein models built from the designed sequences is shown in Figure 1.

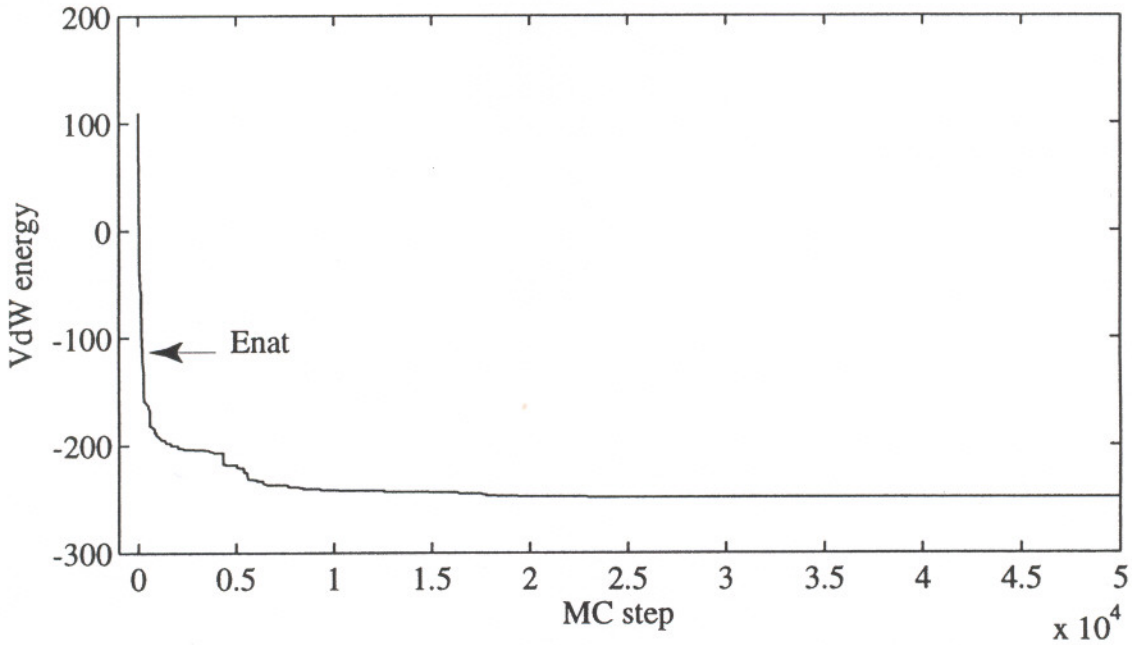

Figure 1 : Evolution of the VdW energies of the various sequences threaded on the backbone of chymotrypsin inhibitor 2CI2 in the course of the Monte Carlo simulation is sequence space. Enat shows the energy (-110 Kcal/Mol) with which the natural sequence of 2CI2 is fitted into its own tertiary structure.

Clearly, the procedure improves packing as defined by Lennard-Jones interactions, yielding proteins with VdW energies much lower than the native protein. It should be mentioned that in all these protein models, the position of the sidechains are approximate, and correspond to discrete conformations derived from a rotamer library. To test the influence of this approximation on our sequence design procedure, protein models were selected in the course of the simulation and energy minimised *in vacuo* using CHARMM[23] version 24b1, based on an all atom force field. A shifted cut-off for non bonded interaction of 13Å and a distance dependent dielectric factor (ε=4R) were used. Results are shown on Figure 2.

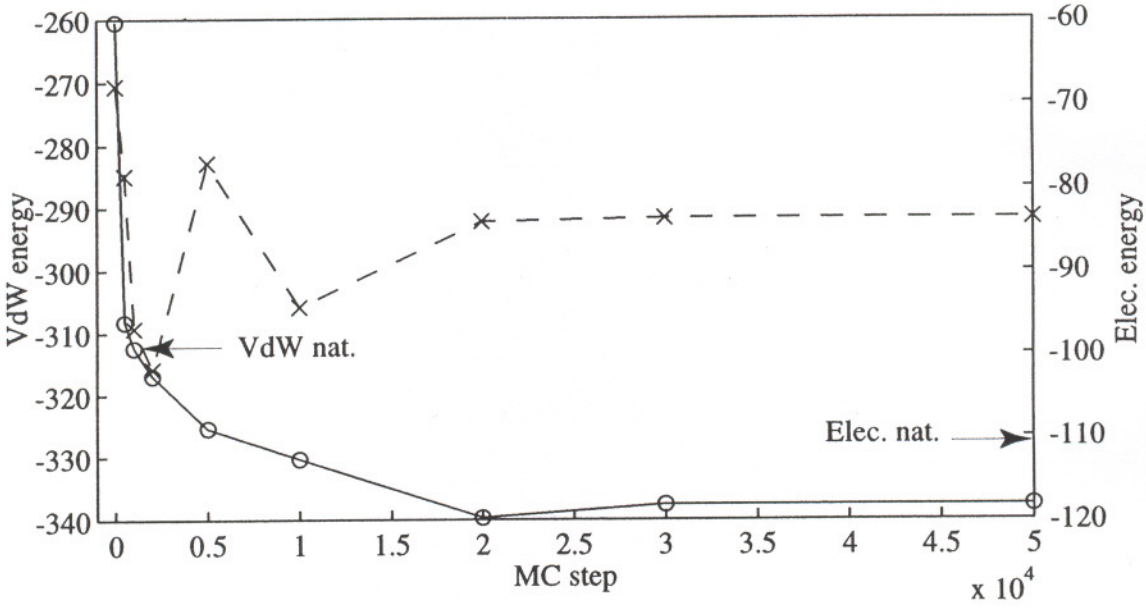

Figure 2 : CHARMM VdW energies (o) and electrostatic energies (x) of protein sequence variants of 2CI2 selected from the full set of structures derived from the Monte Carlo simulation shown in Figure 1 and energy minimized, versus the corresponding Monte Carlo step number. Points were joined by full lines for VdW data and dashed lines for electrostatics data, for sake of clarity. The VdW energy (-313 Kcal/Mol) and electrostatics energy (-112 Kcal/Mol) of the native 2CI2 are shown as arrows.

The total energy of the final protein model (+437 Kcal/Mol) is higher than the total energy of the native protein (+420 Kcal/Mol). However we confirm that our sequence design procedure improved packing as measured by VdW interactions, and even designed proteins with slightly better VdW energies than the native protein (Figure 2). The electrostatic potential, which was not included for sidechain placement or in the scoring function for the Monte Carlo procedure, is not refined but remains negative and reasonable (Figure 2).

Improving sidechain packing yields highly compact protein models. This was monitored by measuring the decrease of the accessible surface areas (ASA) and of the

volumes of some of the model proteins generated during the Monte Carlo sequence design. Results are shown on Figure 3.

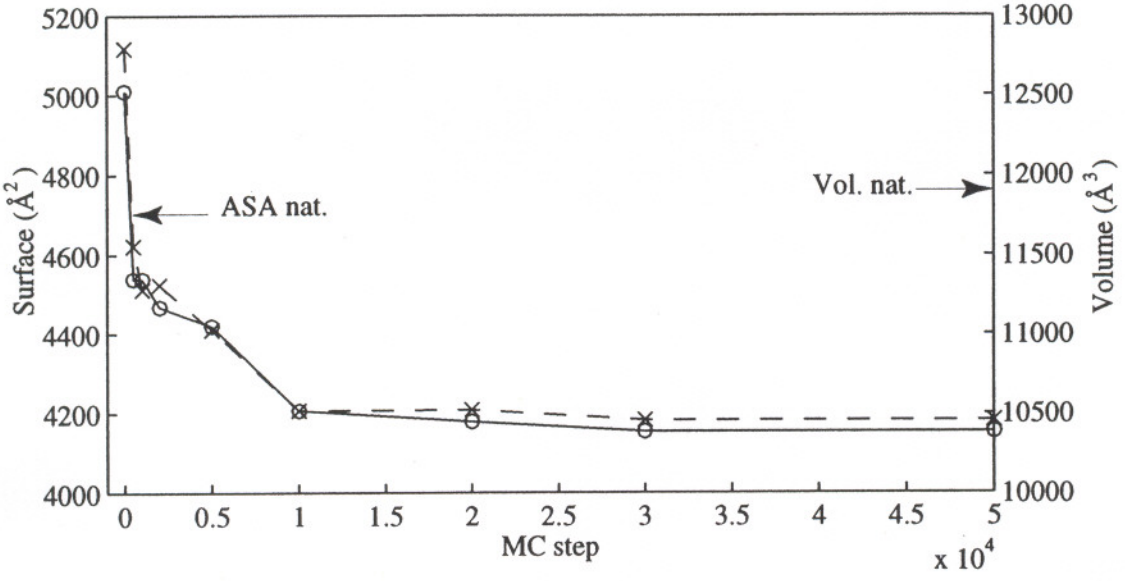

Figure 3 : Evolution of the accessible surface areas (o joined by full lines) and volumes (x joined by dashed lines) of selected protein variants of 2CI2 derived from the Monte Carlo simulation shown in Figure 1, versus the corresponding Monte Carlo step number. ASAnat (4707 $Å^2$) and Vnat (11878 $Å^3$) are the ASA and volume of the native protein, respectively. Surface areas were calculated using the method of Shrake and Rupley[24] implemented in our program ENVIRON[25], and an approximate protein volume was derived from the method proposed by Hao *et al*[26]. The ASA and the volume of the final model protein are 12% lower than the corresponding parameters of the native protein.

3.2 Comparing designed and natural sequences.

An interesting feature of our sequence design procedure is that it succeeds in gradually retrieving information concerning the native sequence (Figure 4). An alignment of the final model sequence (bottom) with the native sequence (top) is given below :

```
NLKTEWPELVGKSVEEAKKVILQDKPEAQIIVLPVGTIVTMEYRIDRVRLFVDKLDNIAEVPRVG
      |   ||           |  |   |        |  |           |
AGRLIWVIEVGETDLIDLLEIRQFEPVTVDKRNPEGKANVPVVLPVEVIEKKVATMLSKIKDYRQ
```

9 residues are identical between the two sequences : W6, V10, G11, I21, Q23, P26, P34, G36 and V48, all located within secondary structures of the protein except for the first tryptophan. Surprisingly, both core and exposed residues are present : W6, I21 and V48 are buried, while all six others have a high fraction of their surface area accessible to solvent. Most of the exposed conserved residues are small. Both

G11 and G36 have positive phi values. Interestingly, L2, which has a positive phi value of 155° in the native structure of 2CI2, has been replaced by a glycine in the optimised model protein.

If the Monte Carlo simulation is repeated at infinite temperature (i.e. all mutations are accepted), the generated sequences are on average 8% identical to the native sequence, which corresponds to the information contained in the constraint of constant amino acid composition.

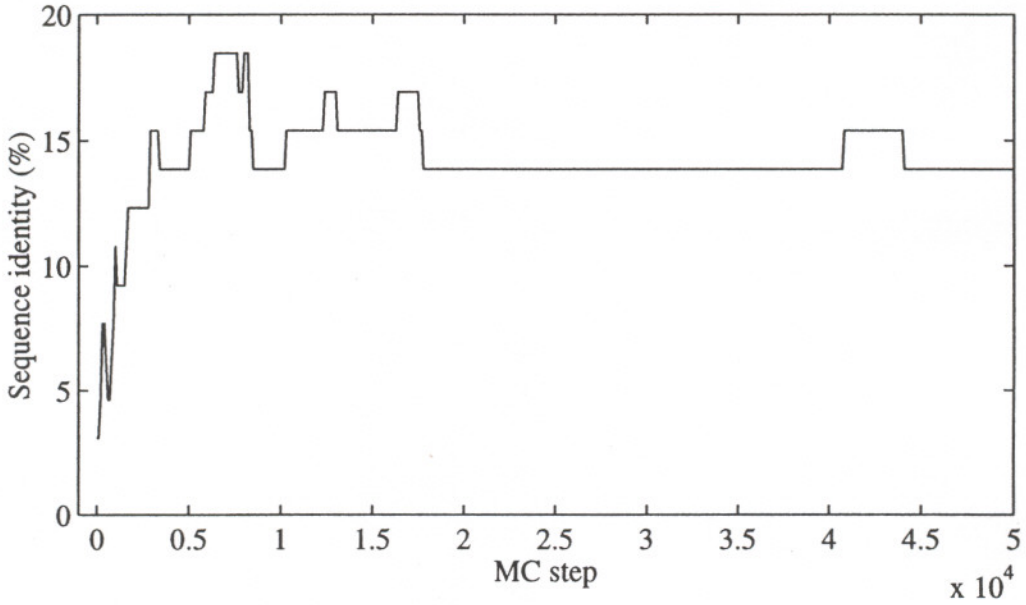

Figure 4 : Sequence identity between the native sequence of 2CI2 and the sequences optimized in the course of the Monte Carlo simulation shown in Figure 1. This information was calculated every 100 cycles, for computing efficiency.

3.3 Specificity of the designed sequence

With our definition of the energy, the sequences we designed are the most stable for the given chymotrypsin inhibitor fold. But will these sequences fold to this structure, i.e. are they incompatible with other protein folds ? To address this question"hide-and-seek" computer experiments were performed[27,28]. In this procedure, the target structure X for a given sequence S is hidden among a large number of non-native folds C, and the task is to retrieve X using an energy criteria. Success is achieved if the energy of S threaded on X, E(S,X), is lower than any E(S,C). A measure of this success is provided by a z-score[29, 30] :

$$z = \frac{E(S,X) - \langle E(S,C) \rangle}{\sigma} \qquad (2)$$

where <> stands for the average over all conformations C, and σ is the corresponding standard deviation.

optimised sequence (-3.12) is even lower than the corresponding z-score for the native sequence (-2.12).

VdW interactions however only describe one aspect of protein structures. Using a more general energy function such as the potentials of mean force of PROSAII, we observe that the sequences designed to increase packing density within the framework of 2CI2 are not specific to this structure.

4. Discussion

Considerable experimental and theoretical efforts have been directed toward de novo protein design. Though most experimentally designed proteins contain significant amounts of secondary structures and appear to fold into an approximately correct topology, they generally lack a well-defined and uniquely structured folded state. A common suspicion is that these designed proteins lack the specific dense packing interactions observed in structures of natural proteins. Hence it is important that theoretical approaches to protein design explicitly take in account packing, which requires that a full atomic representation of the designed protein is maintained. As a first step towards this goal, we have developed in this study an all-atom inverse folding algorithm with the aim of specifically designing amino acid sequences which optimise sidechain packing for a given protein fold. The design is performed by a Monte Carlo optimisation in sequence space, with constant amino acid composition. Packing is defined by a Lennard-Jones potential. Using this program, we have designed highly compact variants of the chymotrypsin inhibitor 2CI2. The final protein model had better van der Waals packing interactions than the native structure. Its density was higher, as observed by smaller total accessible surface areas, as well as by a smaller overall volume, when compared with the native fold. Similar results (not shown) were obtained with chicken triosephosphate isomerase (PDB code 1TIM; 245 residues).

At this stage, our algorithm succeeded in optimising sidechain packing. Can this be of some use for protein design ? Starting from the backbone only of the target structure, the algorithm gradually retrieved sequence information. In the case of 2CI2, the final optimised sequence was only ≈14% identical to the native sequence. Interestingly, the algorithm imposed glycines in position where they are favoured due to steric hindrance. No constraints were imposed on the phi angle of prolines, hence distorted prolines were certainly constructed. This could easily be corrected by adding an energy term to constrain prolines. These two remarks illustrate the positive aspects of working with an all-atom representation of the protein. Though it was not expected that the algorithm would predict the native sequence exactly (proteins are known to be tolerant to mutations[34], we did expect to

find sequences bearing significant resemblance to the native sequence, and this was not found to be the case. It is not clear however that we could expect much better :
-first, the potential energy function which measures the fitness of a sequence with a given structure is based on VdW interactions only and is clearly incomplete for a reliable sequence design algorithm. It does not include for example electrostatics, nor a potential to estimate hydrophobic interactions. These extra terms might not be required if the core only of the protein was considered, but are compulsory for a global sequence design procedure, if we want to maintain patterns of hydrophobic residues in the core, and hydrophilic residues at the surface of the protein.
- second, native structures do not always privilege sidechain packing : it was shown recently that local packing in the vicinity of ß-sheets are not optimal in order to preserve the backbone-to-backbone hydrogen bond patterns within the secondary structures[35]. It should be mentioned that even if we had used the (yet unknown) true complete potential, we might not have achieved the native sequence since nature designs sequences not only for stability but also for function. Natural proteins are usually not optimal : their stability can be improved when for example buried polar groups are replaced by hydrophobic residues or when exposed hydrophobic residues are made polar[34]. These considerations will be difficult to include in any protein design scheme.
- finally our procedure may not infer specificity to our sequences. A sequence is optimal for a given fold if, in addition to being fit to the target fold, it is incompatible with all other possible folds. A commonly used approach consists in maximising the so-called 'energy gap', defined as the difference in energy between the native state and either the best non native conformation[36, 37], or the mean energy of all non native states[38, 39]. The energy gap condition was found to be necessary and sufficient for a 27 mer HP model[40]. It can be included explicitly[39], or implicitly by imposing the amino acid composition[20]. Our approach is based on the latter, and we have shown that by considering VdW interactions only, it did improve specificity (figure 5). The fact that the sequences we designed reached levels of specificity better than the native sequence within the framework of VdW interactions, but were shown to bear no specificity when using another more complete potential may be related to the problem of the definition of the energy function mentioned above. This is in agreement with the conclusion reached by Behe et al[41], i.e. that packing is not the principal cause of conformational specificity. The procedure itself might be questioned : it was shown to fail in the process of designing large HP sequences[42]. Imposing the amino acid composition introduces another limitation : a given protein may not include all twenty amino acids, and the missing ones will never be considered in the sequence design procedure.

The method proposed here can be seen as a framework for a global sequence design procedure based on a full atom representation of proteins. There is room for improvement, including the testing of various potential energy functions that would

provide better estimates of structure-sequence fitness than VdW interaction alone.. Other limitations of our method are related to the constraint of using an overly precise structural template. Crystallographic studies have shown that both backbone and sidechain adjustments occur when residues within protein cores are mutated (for review, see [34]). We are currently working on including backbone flexibility, using multiple copies of the backbone, as described in our earlier work[43]. The self consistent mean field approach itself can be improved to include dynamics[44].

5. Acknowledgements

We thank M. Sippl for making his program PROSAII available to us.

6. References

1. Drexler, K. E. (1981). Proc. Natl. Acad. Sci. (USA) 78:5275-5278.
2. Bowie, J. U. & Eisenberg, D. (1993). Curr. Opin. Struct. Biol. 3:437-444.
3. Jones, D. T. (1995). Curr. Op. Biotechnology 6:452-459.
4. Desjarlais, J. & Handel, T. (1995). Curr Opin. in Biotechnology 6:460-466.
5. Yue, K. & Dill, K. A. (1992). Proc. Natl. Acad. Sci. (USA) 89:4163-4167.
6. Shakhnovich, E. I. & Gutin, A. M. (1993). Proc. Natl. Acad. Sci. (USA) 90:7195-7199.
7. Deutsch, J. M. & Kurosky, T. (1996). Phys. Rev. Lett. 76:323-326.
8. Chan, H. S. & Dill, K. A. (1991). J. Chem. Phys. 95:3775-3787.
9. Jones, D. T. (1994). Prot. Sci. 3:567-574.
10. Sun, S., Brem, R., Chan, H. S. & Dill, K. A. (1995). Prot. Eng. 8:1205-1213.
11. Hinds, D. A. & Levitt, M. (1996). J. Mol. Biol. 258:201-209.
12. Ponder, J. W. & Richards, F. M. (1987). J. Mol. Biol. 193:775-791.
13. Hellinga, H. W. & Richards, F. M. (1994). Proc. Natl. Acad. Sci. (USA) 91:5803-5807.
14. Kono, H. & Doi, J. (1994). Proteins: Struct. Funct. Genet. 19:244-255.
15. Desjarlais, J. & Handel, T. (1995). Protein Sci. 4:2006-2018.
16. Dahiyat, B. & Mayo, S. (1996). Protein Sci. 5:895-903.
17. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). J. Chem. Phys. 21:1087-1092.
18. Koehl, P. & Delarue, M. (1994). J. Mol. Biol. 239:249-275.
19. Bernstein, F. C., Koetzle, T. F., Williams, G., Meyer, D. J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). J. Mol. Biol. 112:535-542.
20. Shakhnovich, E. I. & Gutin, A. M. (1993). Protein Eng. 6:793-800.

21. Nakashima, H., Nishikawa, K. & Ooi, T. (1986). J. Biochem. 99:153-162.
22. Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). J. Biomol. Struct. Dynam. 8:1267-1289.
23. Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. & Karplus, M. (1983). J. Comput. Chem. 4:187-217.
24. Shrake, A. & Rupley, J. (1973). J. Mol. Biol. 79:351-371.
25. Koehl, P. & Delarue, M. (1994). Proteins: Struct. Funct. Genet. 20:264-278.
26. Hao, M., Rackovsky, S., Liwo, A., Pincus, M. & Scheraga, H. (1992). Proc. Natl. Acad. Sci. (USA) 89:6614-6618.
27. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). J. Mol. Biol. 216:167-180.
28. Sippl, M. & Weitckus, S. (1992). Proteins: Struct. Funct. Genet. 13:258-271.
29. Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). Science 253:164-170.
30. Sippl, M. (1993). Proteins: Struct. Funct. Genet. 17:355-362.
31. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). Nature (London) 358:86-89.
32. Sippl, M. (1993). J. Comput. Aided. Mol. Des. 7:473-501.
33. Sippl, M. J. & Jaritz, M. (1994) in *Distance based approaches to Protein Structure Prediction*, H. Bohr and S. Brunak eds, 113-134.
34. Baldwin, E. P. & Matthews, B. W. (1994). Curr. Opin. Biotech. 5:396-402.
35. Schultz Beardsley, D. & Kauzmann, W. (1996). Proc. Natl. Acad. Sci. (USA) 93:4448-4453.
36. Bryngelson, J. D. & Wolynes, P. G. (1987). Proc. Natl. Acad. Sci. (USA) 84:7524-7528.
37. Shakhnovich, E. I. & Gutin, A. M. (1990). Nature (London) 346:773-775.
38. Reva, B. A. & Finkelstein, A. V. (1992). Prot. Eng. 5:625-628.
39. Godzik, A. (1995). Prot. Eng. 8:409-416.
40. Sali, A., Shakhnovich, E. I. & Karplus, M. (1994). J. Mol. Biol. 235:1614-1636.
41. Behe, M. J., Lattman, E. E. & Rose, G. D. (1991). Proc. Natl. Acad. Sci. (USA) 88:4195-4199.
42. Yue, K., Fiebig, K. M., Thomas, P. D., Chan, H. S. & Shakhnovich, E. I. (1995). Proc. Natl. Acad. Sci. (USA) 92:325-329.
43. Koehl, P. & Delarue, M. (1995). Nature Struct. Biol. 2:163-170.
44. Huber, T., Torda, A. & Van Gunsteren, W. (1996). Biopolymers 39:103-114.