The inverse protein folding problem: self consistent mean field optimisation of a structure specific mutation matrix.

M Delarue, P Koehl

THE INVERSE PROTEIN FOLDING PROBLEM: SELF CONSISTENT MEAN FIELD OPTIMISATION OF A STRUCTURE SPECIFIC MUTATION MATRIX.

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

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

The goal of the inverse folding problem is to supply a list of sequences compatible with a known protein structure. If two-body interactions are taken into account in energy calculations, an exhaustive exploration of the energy landscape in sequence space cannot be achieved because of the huge number of possible combinations. To circumvent this problem, we propose a method in which multiple copies corresponding to every possible side-chain type are attached to each Cα position in the protein. The weights of each copy (stored in the sequence matrix **SM**) are refined using mean field theory : each side-chain copy interacts with the mean field generated by all possible side-chain copies at neighbouring positions, weighted by their respective probabilities. The potential energy is simply taken to be amino acid pair potentials of mean force. The method converges in a few cycles to a self-consistent solution. The refined matrix does not depend on the starting point; therefore the method succeeds in removing memory effects. Starting solely from the backbone of the known structure, and without information from the initial sequence, the final sequence matrix **SM** is shown to be able to retrieve significant sequence information, as observed through a series of structure-recognizes-sequence(s) computer experiments. The issue of specificity is discussed in detail.

1. Introduction

The Protein Data Bank[1] which gathers most known protein structures to date, contains a number of structures exceeding several times the number of distinct folds[2]. This has led to the hypothesis that the total number of different folds is finite, and roughly of the order of one thousand[3, 4]. Should examples of every fold be known, protein structure prediction would reduce to the inverse protein folding problem, which consists in identifying which sequences are compatible with a given fold[5]. Elaboration of this alternative view is not only theoretically interesting but is also important for protein design and engineering, as well as structure prediction.

Ponder and Richards[6] provided the first step towards this goal. They systematically tested combinations of sidechains fitting in the core of small proteins, based on steric overlaps, hydrogen bonding and packing density criteria. The number of residues included in the combinatorial search is however limited for practical computing reasons. In fact, as soon as energy calculations contain two-body interaction terms, which are indeed ubiquituous in the classical treatment of proteins then a huge combinatorial problem arises. To alleviate this problem, current methods either refine a single sequence, or limit the search to available sequence databanks (for review, see [7]). Two major approaches have been derived.

In the first one, each known protein structure is represented as a contact map containing position dependent residue-residue contact preferences; aligning a sequence using these contact maps requires a double dynamic programming algorithm[8]. This approach has been successfully applied by Jones et al[9], and is now referred to as the "threading" algorithm. Its drawback however is that it is computationally slow.

Alternatively, the protein structure information can be reduced into one dimension, yielding the so-called 3D-1D profile[10]. In this approach, a position and structure dependent scoring table is built, which contains as many rows as residues in the structure, and a column for each of the twenty amino acids. Aligning a sequence on such profiles resorts to dynamic programming methods developed for pairwise sequence comparisons[11]. In previous applications, scores were usually calculated from the residue environment in the known protein structure (the so-called frozen approximation). One possible caveat is that this is based on the hypothesis that residue environments are conserved within proteins adopting similar folds. This hypothesis has recently been questioned by Russell and Barton[12].

In this paper, we present a new strategy for profile construction based on a self consistent mean field (SCMF) optimisation protocol (for review, see [13]), which does not resort to the frozen approximation. The basic idea of our method is to attach to each Cα of the known structure multiple copies of the side-chain, corresponding to all twenty common amino acids. Each type of side-chain is given a probability, or weight, stored in a sequence matrix, **SM**. This matrix is initialised such that each amino acid type has the same probability, for all residues in the protein (i.e. the system has no memory of the native sequence). Each residue is then considered in turn : the matrix row corresponding to the residue is updated, based on the mean field generated by the multiple side-chains at neighbouring residues, and the procedure is repeated till convergence (i.e. self consistency) is reached.

One of the main results of the paper is to show that the sequence matrix **SM** refined through this procedure, based only on the coordinates of the backbone of the protein, recovers significant sequence information. The discriminatory power of **SM** this matrix is evaluated both in the sequence-recognises-structure protocol[14], in which a given sequence is threaded through an ensemble of profiles derived from known protein folds, as well as in the structure-recognises-sequence assay, in which a large number of sequences is threaded through a given profile.

2. Materials and Methods

A complete description of the self consistent mean field (SCMF) approach is described in our previous works[13, 15]. Specific problems related to optimisation in sequence space are discussed below.

2.1. The chimeric molecule: the multiple copy representation.

The mainchain atoms (N, Cα, C, O) of the known protein structure are fixed in all subsequent calculations. All twenty amino acids are attached to each Cα .

Let us denote as N the total number of residues in the protein. The chimeric molecule described above will be characterised by the sequence matrix **SM** of dimension Nx20, such that SM(i,j) is the probability that residue i is an amino acid of type j. **SM** is initialised such that each residue type is given the same probabiliy.

2.2 . Mean Field Theory : a tool for efficient energy minimisation.

2.2.1. The effective energy of the chimeric molecule The multiple copies of side-chains on a given Cα do not interact with each other; they do interact however with the mean field exerted by all multiple copies of interacting neighbouring positions in the protein. The energy function that we seek to minimize is taken to be:

$$E_{eff} = \sum_{i=1}^{N} \sum_{j=1}^{20} SM(i,j) f[E(i,j)] \qquad (1)$$

where:

$$E(i,j) = \sum_{\substack{k=1 \\ k \neq i}}^{N} \sum_{l=1}^{20} SM(k,l) W(\boldsymbol{x}_{ij}, \boldsymbol{x}_{kl}) \qquad (2)$$

here $\mathbf{x}_{ij}$ are the coordinates of sidechain type j at position i in the sequence, W the pair potential energy function, and

$$f[E(i,j)] = E(i,j) \qquad (3)$$

corresponding to the classical formulation of the effective energy function[13]; subsequent applications of equation (3) will be referred to as Method A.

We also studied an alternative form for f, which will be denoted Method B :

$$f[E(i,j)] = \frac{1}{2}\left[\frac{E(i,j) - U_{av}(j)}{U_{sig}(j)}\right]^2 \qquad (4)$$

where $U_{av}(j)$ is the average energy for an amino acid of type j as observed in native proteins, and $U_{sig}(j)$ the width of the distribution of U(j) over these states. It should be noted that replacing equation (4) in equation (1) leads to an E_{eff} which does not correspond anymore to a term having the dimension of a true energy.

2.2.2. The potential energy function W. W is taken to be the amino acid pair potentials of mean force defined by Sippl[16], which are of the form:

$$W(\boldsymbol{x}_{ij}, \boldsymbol{x}_{kl}) = -RT \ln\left(\frac{f_t^{jl}(r_{ik})}{f_t(r_{ik})}\right) \qquad (5)$$

where j and l correspond to the two amino acids at position i and k in the protein, respectively, t=j-i is the separation of these residues along the sequence, and r_{ik} is the spatial distance between the C_α atoms of i and k. The pair interactions are represented by several variants of potentials, depending on the separation t along the sequence. In the short range, t = 1,2,3,4,5, and 6 individual potentials are compiled for each value of t. For medium, $7 \leq t \leq 9$, and large separations, $10 \leq t$, the pair potentials are condensed to a single type of potential. A cutoff at r=20Å was imposed[17]. Potentials described by Equation (5) were shown to be adequate both for protein folding and inverse protein folding problems[18].

The required density distributions f (corrected for the problem of small dataset, as described by Sippl[16]) were compiled from a database of 83 structurally unrelated proteins (see [19] for a list), excluding all $(\beta/\alpha)_8$ folds and globin folds since triosephosphate isomerase and myoglobin will be used as the main test molecules.

U_{av} and U_{sig}, defined in equation (4) were calculated from the same database, with the restriction that in all protein energy calculations, the respective protein is substracted from the database and the potentials are recompiled (i.e. the jack-knife procedure).

2.2.3. Self-consistent mean field optimisation. Mean field theory consists in finding the minimum of the free energy with respect to all variables[20]. This leads to[13, 21]:

$$SM(i,j) = \exp\left[-\frac{V(i,j)}{RT}\right] \Big/ \sum_{k=1}^{20} \exp\left[-\frac{V(i,k)}{RT}\right] \qquad (6)$$

where

$$V(i,j) = \frac{\partial E_{eff}}{\partial SM(i,j)} \qquad (7)$$

V(i,j) can be seen as the local mean field experienced by amino acid type j at position i in the protein.

Given the initial uniform matrice **SM** and the backbone of the protein, a series of steps is taken to define new probabilities for all amino acids at all positions in the protein. First, all effective potentials are calculated based on Equations (1) and (7). The effective potentials are converted into probabilities using equation (6) and the procedure is iterated as many times as required to reach convergence (i.e. no further modification of **SM**), yielding the final self consistent sequence matrix. To avoid convergence problems such as oscillations, a "memory" was set to the system [15,22].

2.3. Evaluation of Sequence-structure fitness.

The dynamic programming algorithm provides an efficient means of determining the best arrangement of a particular sequence in a particular structure, represented by its sequence matrix. We used a straightforward adaptation of the Smith and

Waterman alignment algorithm[11]. The dynamic algorithm requires additivity and independence. Independence is inherent to the use of the mean field approximation. In order to fulfill the additivity requirement, the sequence matrix is transformed into a profile matrix **PM**:

$$PM(i,j) = \log\left[\frac{SM(i,j)}{Prand}\right] \qquad (8)$$

where Prand is the probability to observe residue type j at position i by chance (Prand is set to 0.05). If SM(i,j) is zero, SM(i,j) is reset arbitrarily to 0.01. The profile matrix was further linearly modified so that its larger and smaller elements are respectively set to 1 and -0.5. This is similar to what is done in recent works[23,24]. For the results reported here, all alignments were performed with a gap opening penalty and a gap elongation penalty of 4.5 and 0.1, respectively, in regions involved in secondary structure, and 0.45 and 0.01 otherwise (typically, a gap can be compensated by 4 or more optimal residue fits in a secondary structure). Secondary structures were defined using the program DSSP[25].

If gaps are not allowed, the score of a particular sequence on a given profile matrix **PM** of the same length is obtained directly as the sum over all positions of the scores of the amino acid type in the sequence, as read in the corresponding row of **PM**.

2.4. The database of test proteins.

108 well refined proteins were included in the database. The corresponding entries in the PDB[1] are : 451c, 156b, 8abp, 8adh, 3adk, a8atc, b8atc, a2aza, 3blm, 1bp2, 2ca2, 1cc5, 1ccr, a2ccy, 3cd4, 2cdv, 3cla, 3cna, a4cpa, 5cpa, 2cpp, 1cpv, 1crn, 2cro, e1cse, i1cse, 1ctf, 2cy3, 2cyp, 8dfr, a4dfr, a1dhf, e2er7, l2fb4, 1fd2, 1fx1, 3fxc, 4fxn, a3gap, 2gbp, 1gcr, o1gd1, 1hip, a2hla, b2hla, 1hoe, 1i1b, 3icb, 7icd, 1l01, 2lbp, 6ldh, 31lrd, a2ltn,1lz1, a4mdh, 2mhr, 2mlt, 2ovo, a2pab, 9pap, 2paz, 1pcy, a1pfk, 3pgk, 3pgm, 1phh, 5pti, 4ptp, 1rhd, 2rhe, 2rnt, 7rsa, 5rxn, 2sga, 3sgb, 1sn3, 2sns, o2sod, 2ssi, 2stv, i1tgs, 6tmn, 4tnc, a1tnf, 1ubq, 1utg, a9wga, r2wrp, b1wsy, 5 globins : 1eca, a1hhb, 1lh1, 5mba, 1mbd and 13 $(\beta/\alpha)_8$ barrels : 2aaa, 4enl, 1fba, a1fcb, 1gox, a5rub, 2mnr, a1pii, b1pii, 1tim, a1wsy, a4xia, and a2ypi.

3. Results

3.1. The sequence matrix recovers information on the native sequence of the protein.

In previous works, self consistent mean field optimisations based on Method A (i.e. equations (1) and (3)) were used for side-chain conformation prediction[15] as well as for homology modelling[26], with reasonable success. Both applications concerned

optimisation in the conformational space : the sequence is known, and the conformation is changed untill a minimum is found. Direct extension of this approach to the problem of optimisation in sequence space leads however to poor results, as illustrated on figure 1 in the case of chicken triosephosphate isomerase (PDB code 1TIM; 245 residues). Calculation of the sequence matrix was carried out over 120 cycles and convergence is achieved in more than 15 cycles (figure 1A).

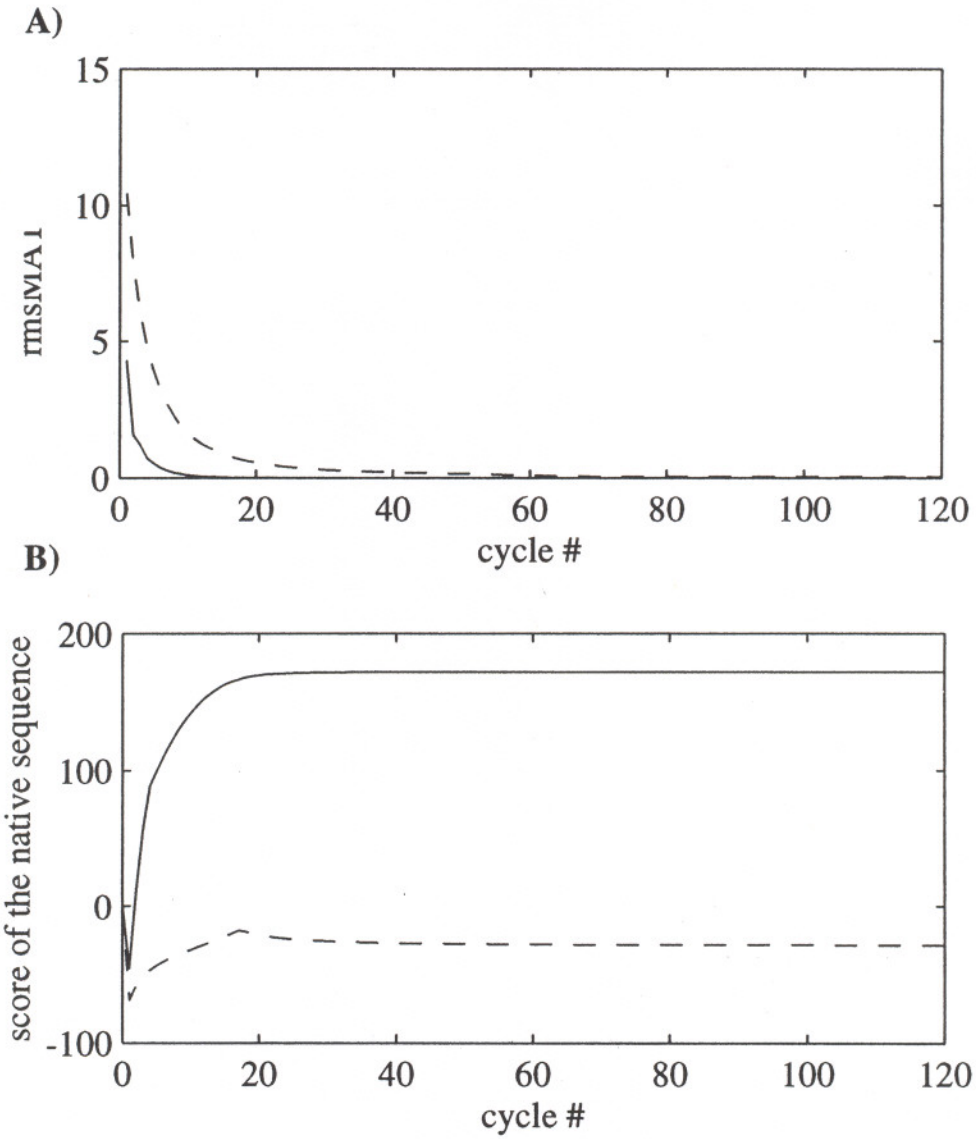

Figure 1 : Illustration of the convergence of the SCMF approach in sequence space in the case of TIM (245 residues) for method A (solid lines) and method B (dashed lines): (A) Starting with a uniform matrix, **SM** is computed iteratively using equations 1, 6 and 7; the r.m.s. difference (rmsMAT) between the successive sequence matrices **SM** are given versus the iteration or cycle number (B) At each cycle of the SCMF refinement, a profile matrix is computed from **SM** using equation (8); the score of the native sequence threaded without gap on this profile matrix is shown versus the cycle number.

The final **SM** shows no specificity towards the native sequence of TIM (figure 1B). In fact, **SM** ends up strongly biased towards sequences of unrealistic amino acid composition. In searching for a minimum of E_{eff}, the procedure selects residue types which provide the lowest contact energies; this happens to be observed for cysteine-cysteine contacts (cysteine has a probability higher than 0.9 in 62 out of the 245 rows of the converged sequence matrix), hence a drift in sequence space. This phenomenon is similar to the design of 'superstable' HP sequences for a given fold just by making all buried residues non polar and surface residues polar[27]. Preliminary experiments (unpublished results) showed that, even at constant amino acid composition, a simple Monte Carlo minimisation in sequence space is able to decrease the total potential energy proteins of known structure to a value lower than the native sequence (for details on the Monte Carlo procedure, see [27]). This means that native structures are not even local minima for the potential energy function we have used; therefore, this energy should not be minimized *per se*, but, rather, should be constrained to take values that are commonly observed for native proteins. We have consequently chosen a new form of total effective energy, based on equation (4) to take this into account. This new formulation, i.e. method B, bears similarity with a procedure adopted in applications of spin glass theory to protein folding[28], as well as in the modified Monte Carlo method of Shakhnovich et al[29, 30].

An example of the convergence of the SCMF approach based on Method B is also given in Fig. 1 in the case of TIM. Calculation were carried out over 120 cycles, but convergence is achieved in less than 20 cycles. The matrix obtained is independent of the starting point. If the initial matrix is set to reflect the initial sequence (probability of 1. for the wild type amino acid type and zero for all others), the refined matrix is identical to the one obtained with an initial random matrix (data not shown). As expected, the total energy of the system (Eq. 1) is minimised during the optimisation. An interesting feature of the specific function given in Eq. 4 is that it succeeds in gradually retrieving information concerning the native sequence (Fig. 1B). In all subsequent calculations only method B is used.

To test if the sequence information achieved through this procedure is significant, a computer experiment was performed. 10,000 random sequences obtained by shuffling the native sequence of TIM are threaded either directly (i.e. without gaps) or by dynamic programming (allowing gaps) on the profile matrix derived from the refined sequence matrix. In both cases, it is seen that the native sequence is well discriminated, with significant z-scores (calculated as $(Snat-Sav)/\sigma$, where Snat is the score of the native sequence, Sav the average score over all sequences and σ the corresponding standard deviation) of 4.66 when no gaps are considered, and 4.11 when gaps are allowed. Both values can be seen as measures of the fitness of the native sequence on the native structure. Fig. 2 summarises the z-scores (allowing gaps) obtained on the database of 108 protein structures, representing a wide range of tertiary folds (α, β, $\alpha\beta$ and $\alpha+\beta$) whose sizes vary from

26 (melittin, 2mlt) to 476 residues (α-amylase, 2aaa). The lowest z-scores are found for small proteins in which only few non local contacts are present.

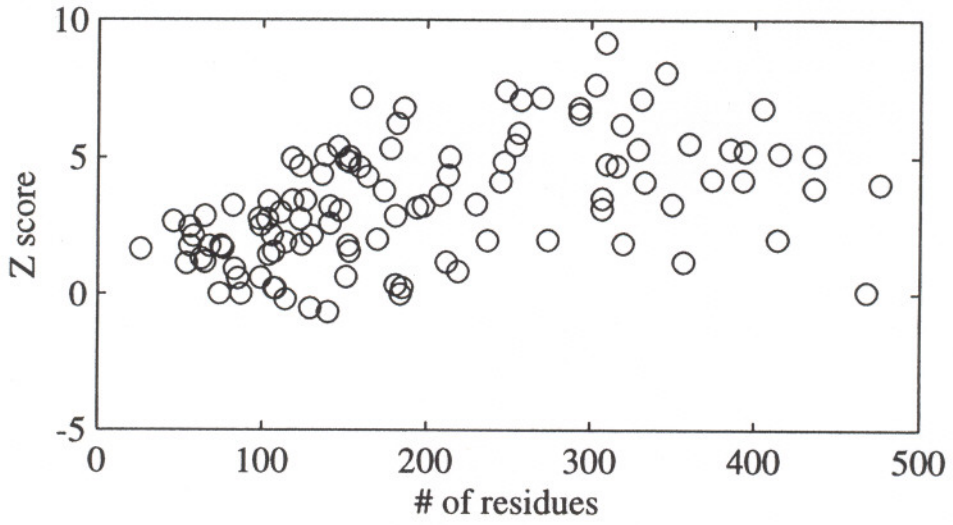

Figure 2 : Sequence profile fitness (expressed as z-scores) as a function of protein size. All z-scores are calculated with respect to an alignment allowing gaps of 5,000 shuffled sequences. Results are shown for the 108 protein structures in our database (see Materials and Methods).

3.2. Sequence-recognises-structure.

In this test, a given protein sequence is aligned to the database of 108 protein structures, one at a time, presented as SCMF profile matrices. The fitness of each sequence-structure alignment is evaluated by means of a z-score calculated from an ensemble of 1000 shuffled sequences. The native sequence scored highest for 63 out of the 87 sequences containing more than 100 residues (smaller sequences were not considered since their fitness values on their native profiles were found to be small; see Fig. 2). Of course, the ultimate practical aim of this kind of experiment is the prediction of the 3D structure of a protein, given its sequence. It is expected to work at least for homologous structures. This was observed for the globin fold, using the SCMF profile matrix (Fig. 3A). The same procedure was applied on TIM, in which case structurally related β/α barrels other than TIM sequences were not detected to a significant level (figure 3B); for a successful application of 3D-1D profile in this difficult test case, see Wilmanns and Eisenberg[31]. In order to improve our method it might be advisable not to give the same weight to all positions in the profile; the question then arises of how to identify those positions that are to be considered with a large weight.

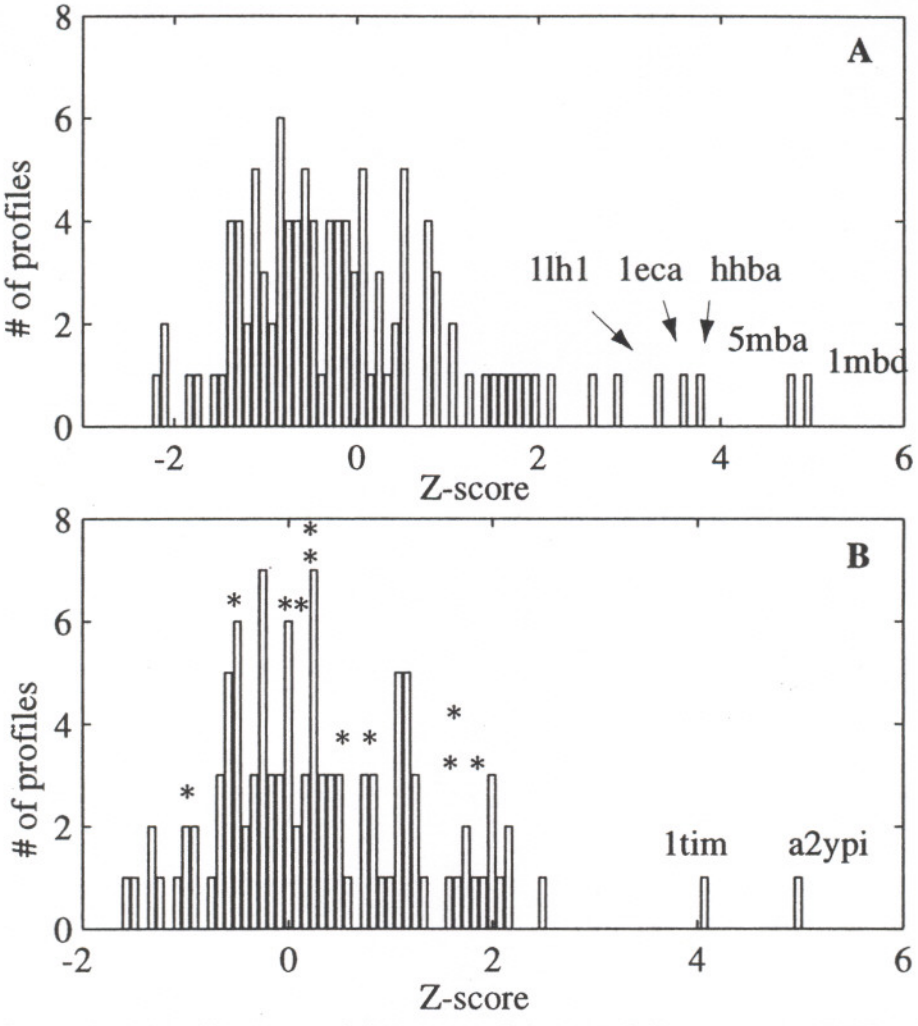

Figure 3 : A) threading histogram of the sperm whale myoglobin sequence (PDB code 1mbd, referred to as MBD) on the SCMF matrices derived from the 108 proteins in our database. All proteins with a globin fold score first. The sequence identity between MBD and A1HHB, 5MBA, 1LH1, 1ECA is 29%, 24%, 17% and 18%, respectively.
B) threading histogram for TIM on the same database of SCMF matrices. The profile matrix derived from the native structure for TIM, i.e. 1tim, scores second after 2ypia, which corresponds to a triosephosphate isomerase from yeast (the 2 proteins have 53% sequence identity). Other related α/β fold are shown as *.

4. Discussion

When searching for all sequences, or just the optimal sequence, that fit to a given protein fold, minimisation of the free energy will usually yield sequences that do not fold into unique conformations. This was observed in the HP model, in which case superstable sequences were designed with all H inside and all P at the surface of the protein[27]; these sequences are not specific in that they can fold into many native conformations; this was demonstrated for 2D lattice simulations[32, 33]. Successful sequence design requires then additional constraints. A sequence is optimal for a

given fold if, in addition, it is incompatible with all other possible folds. Yue and Dill[33] have designed a series of rules to 'design out' incorrect conformations, for HP models on 2D lattice. Dill and coworkers[34] have recently proposed another promising approach for HP model. It is based on a modified energy function containing two terms, the first being designed to favour hydrophobic-hydrophobic contacts, while the second tends to avoid contacts between the solvent (not included explicitly) and hydrophobic monomers. The sequence generated by this method on a given protein structure agree reasonably well with the native sequence, and is shown to fold uniquely on a lattice model. Applications of these strategies [33, 34] to real proteins (i.e. including the 20 amino acids) have not yet been developed; this would be important because there are indications that 2 letter code proteins do not adequately describe real polypeptides[35]. The amino acid composition of a protein is known to be highly dependent on its folding class[36]; this can be imposed as one *global* constraint (see for example [37, 38]). However, even if this condition is necessary, it is certainly not sufficient[38] Another 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[39], or the mean energy of all non native states[40-42]. The energy gap condition was found to be necessary and sufficient for a 27 mer model fitted on a 3x3 lattice[43]. It should be mentioned that a sequence design experiment based on this criteria was shown to fail for larger HP sequences[44], but was found successful for long lattice chains designed in 20-letter code[45-47].

We confine ourselves to the profile formalism. The major novelty of our approach is to provide a rigourous treatment of two-body interactions without having to resort to the frozen approximation. A sequence matrix **SM** is calculated from the backbone only of the protein of interest using a self consistent mean field minimisation of the difference in energy of each residue with average energies found in native structures. The sequence matrix designed here recovers sequence information, yielding a significant sequence-structure fitness value (Fig. 2). It was also found to recognize structurally related proteins with low sequence identity, in cases in which the fold is well preserved, such as the globin fold (Fig. 3A). This success however was not found to be general (Fig. 3B). Limitations of **SM** may be related to the constraint of using a too precise structural template and/or to insufficently discriminative potentials. The profile formalism itself may be questioned: it is based on the independence of each position in the protein, which are then given the same weights. Positions however are not truly equivalent, depending on the presence of secondary structure elements.

The procedure described here is based on pair potentials of mean force derived from known protein structures[16]. Similar potentials have already successfully been used for protein sequence design[37], with a different cutoff. Words of caution are however in order here. Statistical residue-residue pair potentials

derived from preferences observed in the Protein Data Base (potentials of mean force) are by no means equal to the true two body interaction potentials; therefore, their use should always be considered with caution[48]. Improvements including different potential energies as well as optimisation of these potentials are currently being investigated.

5. Acknowledgements

We are grateful to H. Orland, T. Garel, B. Reva, A. Finkelstein and A. Godzik for fruitful discussions. Also, thanks are due to A. Godzik for communicating manuscripts prior to publication. We also acknowledge encouragement and support of D. Moras and J.F. Lefèvre, in whose laboratories part of this work was done.

6. References

1. Bernstein, FC, Koetzle, TF, Williams, G, Meyer, DJ, Brice, MD, Rodgers, JR, Kennard, O, Shimanouchi, T & Tasumi, M (1977). J. Mol. Biol. 112:535-542.
2. Orengo, CA, Flores, TP, Jones, DT, Taylor, WR & Thornton, JM (1993). Current Biology 3:131-139.
3. Orengo, CA, Jones, DT & Thornton, JM (1994). Nature (London) 372:631-634.
4. Chothia, C (1992). Nature (London) 357:543.
5. Drexler, KE (1981). Proc. Natl. Acad. Sci. (USA) 78:5275-5278.
6. Ponder, JW & Richards, FM (1987). J. Mol. Biol. 193:775-791.
7. Bowie, JU & Eisenberg, D (1993). Curr. Opin. Struct. Biol. 3:437-444.
8. Taylor, WR & Orengo, CA (1989). J. Mol. Biol. 208:1-22.
9. Jones, DT, Taylor, WR & Thornton, JM (1992). Nature (London) 358:86-89.
10. Bowie, JU; Lüthy, R & Eisenberg, D (1991). Science 253:164-170.
11. Smith, TF & Waterman, MS (1981). J. Mol. Biol. 147:195-197.
12. Russel, RB & Barton, GJ (1994). J. Mol. Biol. 244:332-350.
13. Koehl, P & Delarue, M (1996). Curr. Opinion Struct. Biol. 6:222-226.
14. Hendlich, M, Lackner, P, Weitckus, S, Floeckner, H, Froschauer, R, Gottsbacher, K, Casari, G & Sippl, MJ (1990). J. Mol. Biol. 216:167-180.
15. Koehl, P & Delarue, M (1994). J. Mol. Biol. 239:249-275.
16. Sippl, MJ (1990). J. Mol. Biol. 1990:859-883.
17. Sippl, MJ & Jaritz, M (1994). In *Predictive power of mean force pair potentials*. Edited by H. Bohr and S. Brunak. IOS Press, Amsterdam. 113-134.
18. Rooman, MJ & Wodak, SJ (1995). Prot. Eng. 8:849-858.
19. Koehl, P & Delarue, M (1994). Proteins: Struct. Funct. Genet. 20:264-278.
20. Kubo, R (1965). *Statistical Physics*. North Holland Publishing Co, Amsterdam.
21. Rabow, AA & Scheraga, HA (1993). J. Mol. Biol. 232:1157-1168.
22. Finkelstein, AV & Reva, BA (1991). Nature (London) 351:497-499.

23. Ouzounis, C, Sander, C, Sharf, M & Schneider, R (1993). J. Mol. Biol. 232:805-825.
24. Abagyan, R, Frishman, D & Argos, P (1994). Proteins: Struct. Funct. Genet. 19:132-140.
25. Kabsch, W & Sander, C (1983). Biopolymers 22:2577-2637.
26. Koehl, P & Delarue, M (1995). Nature Struct. Biol. 2:163-170.
27. Shakhnovich, EI & Gutin, AM (1993). Protein Eng. 6:793-800.
28. Goldstein, RA, Luthey-Schulten, ZA & Wolynes, PG (1992). Proc. Natl. Acad. Sci. (USA) 89:4918-4922.
29. Abkevich, V, Gutin, A & Shakhnovich, EI (1995). J. Mol. Biol. 252:460-471.
30. Mirny, L, Abkevich, V & Shakhnovich, EI (1996). Folding & Design 1:103-116.
31. Wilmanns, M & Eisenberg, D (1993). Proc. Natl. Acad. Sci. (USA) 90:1379-1383.
32. Chan, HS & Dill, KA (1991). J. Chem. Phys. 95:3775-3787.
33. Yue, K & Dill, KA (1992). Proc. Natl. Acad. Sci. (USA) 89:4163-4167.
34. Sun, S, Brem, R, Chan, HS & Dill, KA (1995). Prot. Eng. 8:1205-1213.
35. Onuchic, JN, Wolynes, PG, Luthey-Schulten, Z & Socci, ND (1995). Proc. Natl. Acad. Sci. (USA) 92:3626-3630.
36. Nakashima, H, Nishikawa, K & Ooi, T (1986). J. Biochem. 99:153-162.
37. Jones, DT (1994). Prot. Sci. 3:567-574.
38. Hinds, DA & Levitt, M (1996). J. Mol. Biol. 258:201-209.
39. Shakhnovich, EI & Gutin, AM (1990). Nature (London) 346:773-775.
40. Finkelstein, AV & Reva, BA (1992). Prot. Eng. 7:617-624.
41. Godzik, A (1995). Prot. Eng. 8:409-416.
42. Deutsch, JM & Kurosky, T (1996). Phys. Rev. Lett. 76:323-326.
43. Sali, A, Shakhnovich, EI & Karplus, M (1994). J. Mol. Biol. 235:1614-1636.
44. Yue, K, Fiebig, KM, Thomas, PD, Chan, HS & Shakhnovich, EI (1995). Proc. Natl. Acad. Sci. (USA) 92:325-329.
45. Shakhnovich, EI (1994). Phys. Rev. Lett. 72:3907-3910.
46. Abkevich, V, Gutin, AM & Shakhnovich, EI (1995). Prot. Sci. 4:1167-1177.
47. Shakhnovich, EI, Abkevich, V & Ptitsyn, O (1996). Nature (London) 379:96-98.
48. Thomas, PD & Dill, KA (1996). J. Mol. Biol. 257:457-469.

7. Appendix

7.1 Method A : the total effective energy of the chimeric molecule is given by :

$$E_{eff} = \sum_{i=1}^{N} \sum_{j=1}^{20} SM(i,j)E(i,j) \tag{A.1}$$

where E(i,j) is defined by equation (2).
The local mean field exerted on amino acid type b for residue a is (see equation 7) :

$$V(a,b) = \frac{\partial E_{eff}}{\partial SM(a,b)} = 2E(a,b) \tag{A.2}$$

7.2 Method B : the effective energy of the chimeric molecule is defined as :

$$E_{eff} = \sum_{i=1}^{N} \sum_{j=1}^{20} SM(i,j) \left(\frac{E(i,j) - U_{av}(j)}{U_{sig}(j)} \right)^2 \tag{A.3}$$

in which case :

$$V(a,b) = \left(\frac{E(a,b) - U_{av}(b)}{U_{sig}(b)} \right)^2 + 2 \sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{j=1}^{20} SM(i,j) \frac{W(x_{ab}, x_{ij})}{U_{sig}(j)} \frac{E(i,j) - U_{av}(j)}{U_{sig}(j)} \tag{A.4}$$