Mean-field minimization methods for biological macromolecules

Patrice Koehl* and Marc Delarue†

Simulations of macromolecular structures involve the minimization of a potential-energy function that presents many local minima. Mean-field theory provides a tool that enables us to escape these minima, by enhancing sampling in conformational space. The number of applications of this technique has increased significantly over the past year, enabling problems with protein-homology modelling and inverted protein structure prediction to be solved.

Addresses
*UPR 9003 du CNRS, Ecole Supérieure de Biotechnologie de Strasbourg, Boulevard Sébastien Brant, 67400 Illkirch Graffenstaden, France; e-mail: koehl@bali.u-strasbg.fr
†Laboratoire d'Immunologie Structurale, Institut Pasteur 25, rue du Docteur Roux, 75015 Paris Cedex, France; e-mail: delarue@pasteur.fr

Current Opinion in Structural Biology 1996, 6:222–226

Abbreviations

CM	conformational matrix
LES	locally enhanced sampling
MFA	mean-field algorithm
MFT	mean-field theory
SCEO	self-consistent ensemble optimization
SCMF	self-consistent mean field

Introduction

A major challenge in computational biology is the prediction of the native structure (that with the lowest energy) adopted by a macromolecule (nucleic acid or protein) *in vivo*. The search for this global minimum is hindered by the existence of multiple minima on the conformational energy surface [1]. It is easy to find the local minimum associated with a given conformation, but very hard to escape from this minimum in the process of searching for the global minimum. For simplified systems, such as lattice models of small peptides, this problem can be solved through an enumeration and energetic evaluation of all possible conformations [2••]. This method, however, is not suitable for the study of a larger system, such as a complete protein, because of the exponentially growing configurational space arising from the many possible ways in which the backbone and side chains of the protein can pack. For example, for a 50-residue protein in which each residue can adopt no more than five different states, a total of 9×10^{34} conformations need to be tested. This would require years of calculation on even the fastest available supercomputers. Various procedures have been developed in an attempt to alleviate this problem. This review focuses on one of these procedures: namely, mean-field theory (MFT).

In the first application of MFT in protein simulation, presented by Elber and Karplus [3], approximate mean-field treatment of protein–ligand dynamics enabled detailed studies to be made of the diffusion pathways of carbon monoxide through myoglobin. Finkelstein and Reva [4] published another application of MFT: testing stable protein folds. Other applications followed: in particular, variants of MFT were used to find minimum-energy conformations for the side-chain modelling problem [5,6••–9••], for protein conformational optimization [10–12], for protein structure prediction on a lattice [13,14], for loop construction in protein homology modelling [15••,16••] and for protein sequence design [17,18••]. Here, we will focus on the application of MFT to side-chain prediction and sequence design.

The mean-field theory

We define the coordinate vector of all atoms in a molecular system as $\mathbf{X}$. The probability of finding the molecular coordinates between $\mathbf{X}$ and $\mathbf{X}+d\mathbf{X}$ is denoted by $\rho(\mathbf{X})d\mathbf{X}$, where $\rho(\mathbf{X})$ is the probability density of coordinates, normalized to 1. The total energy (E) of the system is given by

$$E=\int U(\boldsymbol{X})\rho(\boldsymbol{X})d\boldsymbol{X} \quad (1)$$

where U is the potential energy function. If the system contains one molecule with a unique conformation whose coordinates are $\mathbf{X}^0$ [i.e. $\rho(\mathbf{X})=\delta(\mathbf{X}-\mathbf{X}^0)$], Equation 1 becomes the following equation.

$$E=U(\boldsymbol{X}^0) \quad (2)$$

The native conformation of the molecular system is derived from the global minimum of E. The search for this global minimum is hindered by the presence of numerous local minima. One way to alleviate this problem is to consider an effective, larger system as a computational tool to enhance the sampling protocol of the minimization algorithm. MFT is one possible approach for efficiently studying this effective system. It is always based on the approximation that the probability density ρ is replaced by a product of independent probability densities of different subsystems, using a Hartree product [19]:

$$\rho(\boldsymbol{X})=\prod_{j=1}^{J}\rho_j(\boldsymbol{X}_j) \quad (3)$$

As an example of such a partition into subsystems, a ligand and a protein can be considered separately [3,20]. Another example is the partition of a protein into backbone and side chains [5,6••–9••]. The effective system is then built by considering multiple copies of each subsystem j; each of the ρj is expanded in delta functions:

$$\rho_j(\boldsymbol{X}_j) = \sum_{k_j=1}^{K_j} V(j,k_j)\delta(\boldsymbol{X}_j - \boldsymbol{X}_{k_j}^o) \quad (4)$$

where k_j runs over all K_j copies of the subsystem j. $V(j,k_j)$ are normalization factors verifying the following.

$$\sum_{k_j=1}^{K_j} V(j,k_j) = 1 \quad (5)$$

For practical reasons, K_j is always finite.

Substituting Equations 3 and 4 into Equation 1 and integrating over the spatial variables leads to the equation below.

$$E_{eff} = \sum_{\substack{k_j=1 \\ j=1,\ldots,J}}^{K_j} \left(\prod_{l=1}^{J} V(l,k_j)\right) U(\boldsymbol{X}_{k_1}^o, \boldsymbol{X}_{k_2}^o, \ldots, \boldsymbol{X}_{k_j}^o, \ldots, \boldsymbol{X}_{k_J}^o) \quad (6)$$

For simplicity, we also assume that the potential function U can be written in pairwise form:

$$U(\boldsymbol{X}) = \sum_{j=1}^{J} U_j(\boldsymbol{X}_j) + \sum_{j=1}^{J} \sum_{i>j} U_{ij}(\boldsymbol{X}_i, \boldsymbol{X}_j) \quad (7)$$

in which case Equation 6 reduces to:

$$E_{eff} = \sum_{j=1}^{J} \sum_{k_j=1}^{K_j} V(j,k_j) U_j(\boldsymbol{X}_{k_j}^o) + \sum_{j=1}^{J} \sum_{k_j=1}^{K_j} \sum_{i>j}^{J} \sum_{l_i=1}^{K_i} V(j,k_j)V(i,l_i)U_{ij}(\boldsymbol{X}_{k_j}^o, \boldsymbol{X}_{l_i}^o) \quad (8)$$

which is the energy used in most MFT applications discussed here. In a sense, the problem of finding the global minimum energy in the total configurational space is mapped to the problem of finding the minimum of this 'effective' potential energy, which is the sum of real potential energies calculated at different points (obtained from the delta function expansions in Eqn 4) multiplied by normalization factors. They correspond to all possible combinations of the coordinates for the various subsystems j. For example, there are $K_i \times K_j$ alternative configurations of subsystems i and j that can be examined using a *single configuration* of the effective system. Roitberg and Elber [5] have demonstrated an important feature of E_{eff}: namely, that the global minimum of the effective system described by the energy function E_{eff} coincides with that of the real system, which is described by a single configuration for each subsystem j. It also seems that the energy landscape corresponding to E_{eff} is much simpler than that of the original system.

The minimization of E_{eff} can be obtained by two different routes. Firstly, the normalization factors v are kept constant (usually taken as $V(j,k_j) = 1/J$, where J is the number of copies for subsystem j); the positions of each copy for each subsystem are then optimized by solving Newton-like equations of motion. This procedure corresponds to the LES (locally enhanced sampling) protocol [3,21], as well as to the multicopy sampling proposed by Rosenfield *et al.* [22] and Zheng *et al.* [23]. Secondly the position of the various copies of the subsystems are supposed to be known and fixed in space (they correspond to the various possible rotamer states for a given side chain). The effective system is then described by an array v, whose current element V(j,k) is the probability that subsystem j is described by its possible state k. In this case, the free energy function ($F = E_{eff} - TS$, where S is the entropy of the system and T the temperature) is minimized with respect to the variables V(j,k) [4,24] and the probabilities V(j,k) for each copy k of subsystem j are obtained as:

$$V(j,k) = \frac{\exp\left(-\frac{W(j,k)}{kT}\right)}{\sum_{l=1}^{K_j} \exp\left(-\frac{W(j,l)}{kT}\right)} \quad (9)$$

where W(j,k) corresponds to the molecular field potentials felt by the copy k in subsystem j as shown in the equation below.

$$W(j,k) = \frac{\partial E_{eff}}{\partial V(j,k)} \quad (10)$$

The system of Equations 8–10 is then iterated until convergence, that is, until self-consistency is achieved.

Interestingly, Equations 9 and 10 have been independently derived by direct evaluation of the partition function Z using the saddle-point approximation for MFT applications of neural network minimization [13,25,26]. In these cases, E_{eff} and W are generalized energies and kT is a parameter. Neural networks based on MFT have been applied to the problem of protein folding on a lattice [13]. The original application of MFT to neural networks is based on

$$V(j,k) = \frac{\exp\left(-\frac{W(j,k)}{kT}\right)}{\exp\left(\frac{W(j,k)}{kT}\right) + \exp\left(-\frac{W(j,k)}{kT}\right)} \quad (11)$$

instead of Equation 10 [27]. It should be mentioned that Equation 10 maintains the row normalization of the variables V(j,k), which is defined by Equation 5, whereas the use of Equation 11 requires that extra terms are added to the pseudopotential functions U_j in order to impose Equation 5.

MFT as a tool for side-chain modelling in proteins

The major problem that hinders the prediction of side-chain conformations for a protein with a known backbone scaffold is of a combinatorial nature: a systematic search of all possible side-chain packings for even a small protein is computationally unrealistic. Different strategies have been proposed to solve this problem ([9••]; Vásquez, this issue, pp 217–221).

The MFT protocol described above is well adapted to this problem. A protein of M residues is separated into the framework (containing the backbone and P fixed side chains) and the N = M–P side chains to be modelled, yielding N+1 'independent' subsystems. For each side-chain subsystem, multiple side-chain copies are connected to the same Cα. An application of the LES protocol to refine the positions of these side-chain copies was described by Roitberg and Elber [5]. Recent applications focus on the self-consistent optimization protocol.

SCMF

The SCMF (self-consistent mean field) optimization method [6••] is a direct application of the MFT described above. It is based on the rotamer library of Tuffery *et al.* [28] and iteratively refines a conformational matrix (**CM**) of the side chains of a protein using Equations 9 and 10 (**CM** is identical to the array v defined above). During the refinement, the temperature is kept fixed at 300 K. The potential energy specified by U in Equation 8 includes 12–6 van der Waals interactions as well as geometric constraints for disulphide bridge formation. After convergence, the rotamers with the highest probability in the self-consistent **CM** are used to define the conformation of the side chains, yielding levels of accuracy similar to other automated side-chain modelling techniques. This accuracy is obtained at a much lower computational cost, which is found to vary only linearly with the size of the protein. The optimized **CM** can also be used to provide crude estimates of the conformational entropy of the side chain in the folded state of the protein. Limitations of the SCMF procedure are related to the definition of the potential function (which does not include electrostatics, for example), the inherent problem related to the use of a discrete rotamer library (the ideal solution, in which all side chains have been replaced by their closest root mean square [rms] rotamer, does not always correspond to the minimum of the potential function), and the choice of the side-chain conformations based on the self-consistent **CM**.

SCEO

SCEO (self-consistent ensemble optimization), another application of MFT to the problem of packing side chains in proteins, was independently described by Lee [7••] at the same time as the description of SCMF. SCEO is similar to SCMF, in that it also uses Equations 8–10, but it differs in that the temperature T is taken as an adjustable parameter. An equivalent to the **CM** is refined, starting at $T=\infty$ (in which case **CM** is uniform) and gradually cooling down to 0 K, using the self-consistent refinement at each temperature step. At 0 K, only one copy of each subsystem has a non-zero probability. In practice, it was found that the cooling process could be stopped at 298 K. Another way in which SCEO differs from SCMF is that it does not rely on a rotamer library. Instead, torsional angles vary in the full range of rotations from 0–360°, in 33 discrete steps of about 12°; a side chain defined by $\chi 1$ and $\chi 2$ is then defined by 1089 copies (this should be compared to a maximum of 16 based on the rotamer library of Tuffery *et al.* [28]). The free rotation of the torsional angles imposes the addition of a torsional potential to the potential energy U. Because of the large numbers of copies for each side chain, explicit integration of Equation 8 over all possible residue–residue interactions would require intensive computation. Instead, in SCEO, the mean field is approximated using an elegant Monte Carlo sampling biased by the residues' conformation probabilities included in **CM**. Computations, however, are still limited to a small number (N) of side chains to be modelled (typically six or eight). Predicted energies based on SCEO refinement of a series of nine mutants of the hydrophobic core of the λ repressor protein correlated well with experimental free energies of unfolding.

Mean-field algorithm

Recently, Vásquez [9••] derived a third method, referred to as MFA (mean-field algorithm) for side-chain modelling using mean-field theory that starts in exactly the same way as the SCMF procedure, but follows a different mean-field Monte Carlo procedure to select the correct rotamer on the basis of the self-consistent conformation matrix **CM** for each residue to be modelled. Using this MFA protocol, a larger rotamer library and a 9–6 van der Waals potential (instead of the classic 12–6 potential), Vásquez found significant improvements in protein side-chain modelling accuracy using this method compared to SCMF.

MFT applications to protein sequence design

In inverted protein design, one seeks protein sequences that are compatible with a known three-dimensional structure. Two main issues have to be addressed in this procedure. Firstly, the combinatorial problem of testing all possible sequences on the structure cannot be solved systematically, except for small parts of the protein [29]. Secondly, a measure of the compatibility of a sequence with a given structure is needed. A possible approach to the first problem is based on MFT and has been explored by Kono and Doi [18••]. In their method, the protein is again divided into a framework and the N residues to be modelled, yielding N+1 subsystems. For each residue subsystem, multiple side chains are attached to the Cα, corresponding to all rotamers of all amino acids that could fit. Each copy is assigned a probability, and these probabilities are iteratively refined using Equations 9 and 11 until self-consistency is reached. The potential function U in Equation 8 contains a 12–6 van der Waals potential, as well as terms to maintain row normalization (Eqn 5) and to favour strongly stable states, in which only one copy per residue has a non-zero probability. The initial probability matrix is chosen randomly, so as to generate a family of possible sequences. Though the method is restricted to the interior core region, and the list of possible amino acids at each site is limited to six hydrophobic residues, it produces encouraging results. The native sequence is always recovered among the best refined sequences, which are all well packed.

Extension of this procedure to the full protein, where all twenty amino acids are allowed at each site, is computationally easy but conceptually not straightforward. Without any sequence composition constraint and with only van der Waals energetics, unrealistic sequences will be obtained, as observed when using other sequence-design techniques [30••–33••]. Present theoretical efforts in sequence design focus more on the definition of the potential energy functions than on the minimization problem [34,35] (see also the two related reviews in this issue by Jernigan and Bahar [pp 195–209] and Thornton and Jones [pp 210–216]). In particular, it is not clear whether or not residue–residue statistical potentials derived from the Protein Structure Data Base can be used as real two-body potentials in Equation 8.

Conclusions

So far, MFT applications have mainly been focused on modelling theoretical problems: side-chain modelling [5,6••–9••], loop design in protein modelling [15••,16••], predicting protein mutant energetics [7••], fitting a native structure to a lattice [13], and the protein folding [13,14] and inverse folding problems [17,18••]. In fact, MFT can be applied to any optimization problem in molecular structures. Protein structure calculations based on experimental NMR and/or X-ray crystallography data should also benefit from it; an application of multicopy sampling for X-ray crystallography has recently been described [36••]).

Protein design will probably become a great challenge for MFT. The problem in this area is to identify amino acid sequences with no obvious similarity to known sequences that would adopt a given three-dimensional protein fold. The challenge arises from the seemingly limitless number of possible sequences from which to choose and MFT can provide one solution to overcome this problem.

Acknowledgements

We are grateful to H Orland, T Garel, B Reva, A Finkelstein, A Godzik, J-F Lefèvre, D Moras and T Simonson for fruitful discussions on protein structure simulations. We also thank G Hoffmann for editorial assistance.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Gibson KD, Scheraga HA: **The multiple minima problem in protein folding.** In *Structure and Expression: from Proteins to Ribosomes.* Edited by Sarma RH, Sarma MH. Guilderland, NY: Adenine Press; 1988:67–94.

2. •• Dill KA, Bromberg S, Yue KZ, Fiebig KM, Yee DP, Thomas PD, Chan HS: **Principles of protein folding – a perspective from simple exact models.** *Protein Sci* 1995, **4**:561–602.
A thorough review on exact hydrophobic–polar (HP) lattice models of proteins, and their use in studying the thermodynamics and kinetics of protein folding processes.

3. Elber R, Karplus M: **Enhanced sampling in molecular dynamics: use of the time-dependent Hartree approximation for a simulation of carbon monoxide diffusion through myoglobin.** *J Am Chem Soc* 1990, **112**:9161–9175.

4. Finkelstein AV, Reva BA: **A search for the most stable folds of protein chains.** *Nature* 1991, **351**:497–499.

5. Roitberg A, Elber R: **Modelling side-chains in peptides and proteins: application of the locally enhanced sampling and the simulated annealing method to find minimum energy conformations.** *J Chem Phys* 1991, **95**:9277–9287.

6. •• Koehl P, Delarue M: **Application of a self consistent mean field theory to predict protein side-chain conformations and estimate their conformational entropy.** *J Mol Biol* 1994, **239**:249–275.
A method for modelling protein side chains on a given framework is described. The method is based on a rotamer library and iteratively refines a conformational matrix of the side chains of a protein, using a self-consistent mean-field approach. The energy function only includes 12–6 Lennard Jones potentials. The rotamers with the highest probability in the optimized conformation matrix are used for final prediction. The conformational matrix also provides estimates of the configurational entropy of the side chains in folded proteins (see also [7••,9••])

7. •• Lee C: **Predicting protein mutant energetics by self consistent ensemble optimisation.** *J Mol Biol* 1994, **236**:918–939.
A self-consistent ensemble optimization is applied to predict side-chain conformations in the core of a protein and the effect of mutations on protein stability. It differs from the method presented in [6••] in that it does not rely on a rotamer library. Its application, however, is limited to small proteins, or to a limited number of side chains in a larger system (see also [9••]).

8. •• Zheng Q, Kyle DJ: **Multiple copy sampling: rigid versus flexible protein.** *Proteins* 1994, **19**:324–329.
The effects of protein flexibility on multiple copy sampling are systematically evaluated by studying a single side chain placement. The technique is shown to be significantly more efficient when a flexible, but harmonically constrained, protein is used instead of a rigid protein.

9. •• Vásquez M: **An evaluation of discrete and continuum search techniques for conformational analysis of side chains in proteins.** *Biopolymers* 1995, **36**:53–70.
Methodologies for calculation of side-chain conformations in proteins are evaluated, including heat bath algorithm and MFT techniques. It is shown that using non-ideal rotamers, as well as modification of the Lennard Jones energy function to a 9–6 potential instead of a 12–6 potential, leads to more accurate predictions.

10. Olszewski KA, Piela L, Scheraga HA: **Mean field theory as a tool for intramolecular conformational optimisation. 1. Tests on terminally-blocked alanine and met-enkephalin.** *J Phys Chem* 1992, **96**:4672–4676.

11. Olszewski KA, Piela L, Scheraga HA: **Mean field theory as a tool for intramolecular conformational optimisation. 2. Tests on the homopolypeptides decaglycine and icosalanine.** *J Phys Chem* 1993, **97**:260–266.

12. Olszewski KA, Piela L, Scheraga HA: **Mean field theory as a tool for intramolecular conformational optimisation. 3. Test on melittin.** *J Phys Chem* 1993, **97**:267–270.

13. Rabow AA, Scheraga HA: **Lattice neural network minimization. Application of neural network optimization for locating the global-minimum conformations of proteins.** *J Mol Biol* 1993, **232**:1157–1168.

14. Finkelstein AV, Reva BA: **Search for the stable state of a short chain in a molecular field.** *Protein Eng* 1992, **5**:617–624.

15. •• Zheng Q, Rosenfeld R, DeLisi C, Kyle JD: **Multiple copy sampling in protein loop modeling: computational efficiency and sensitivity to dihedral angle perturbations.** *Protein Sci* 1994, **3**:493–506.
The bond-scaling relaxation technique proposed earlier by some of the same authors is combined with multiple copy sampling to enable them to model loop conformations in proteins, leading to an increased efficiency of up to a factor of five. The variability in the converged loop conformations is used to estimate the accuracy of the models.

16. •• Koehl P, Delarue M: **A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling.** *Nature Struct Biol* 1995, **2**:163–170.
The SCMF approach presented in [6••] is generalized to apply to the problem of the simultaneous modelling of loops and the positioning of side chains in proteins. It relies on a database search to generate possible fragments for modelling gaps, and on a rotamer library to define side-chain conformations. Each loop fragment and each side-chain rotamer is characterized with a probability, which is refined using MFT. We show that, in the presence of multiple loops, simultaneous modelling yields more accurate predictions (see [14]).

17. Reva BA, Finkelstein AV: **A new approach to the design of a sequence with the highest affinity for a molecular surface.** *Protein Eng* 1992, **5**:625–628.

18. •• Kono H, Doi J: **Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry.** *Proteins* 1994, **19**:244–255.
This paper presents the use of a method of energy minimization, using an automata network, to predict a set of amino acid sequences and their side-chain conformations compatible with a given backbone geometry. It is essentially similar to the lattice neural network minimization procedure of Rabow and Scheraga [13] and bears some resemblance to the MFT techniques. Applications are restricted to the interior core region, with limited choice for amino acid types (six hydrophobic residues). The optimized sequences were found to be as well packed as the native sequence.

19. Landau LD, Lifshitz EM: *Quantum Mechanics.* New York: Pergamon Press; 1958.

20. Czerminski R, Elber R: **Computational studies of ligand diffusion in globins: 1. Leghemoglobin.** *Proteins* 1991, **10**:70–80.

21. Straub JE, Karplus M: **Energy equipartitioning in the classical time-dependent Hartree approximation.** *J Chem Phys* 1991, **94**:6737–6739.

22. Rosenfeld R, Zheng Q, Vajda S, DeLisi C: **Computing the structure of bound peptides: application to antigen recognition by class I MHCs.** *J Mol Biol* 1993, **234**:515–520.

23. Zheng Q, Rosenfeld R, Kyle DJ: **Theoretical analysis of the multi-copy sampling method in molecular modeling.** *J Chem Phys* 1993, **99**:8892–8896.

24. Kubo R: *Statistical Physics.* Amsterdam: North Holland Publishing Co; 1965.

25. Van den Bout DE, Miller TKI: **Improving the performance of the Hopfield-Tank neural network through normalization and annealing.** *Biol Cybern* 1989, **62**:129–139.

26. Peterson C, Söderberg B: **A new method for mapping optimization problems onto neural networks.** *Internat J Neural Systems* 1989, **1**:3–22.

27. Hopfield JJ, Tank DW: **"Neural" computation of decisions in optimization problems.** *Biol Cybern* 1985, **52**:141–152.

28. Tuffery P, Etchebest C, Hazout S, Lavery R: **A new approach to the rapid determination of protein side-chain conformations.** *J Biomol Struct Dyn* 1991, **8**:1267–1289.

29. Ponder JW, Richards FM: **Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.** *J Mol Biol* 1987, **193**:775–791.

30. •• Jones DT: ***De novo* protein design using pairwise potentials and a genetic algorithm.** *Protein Sci* 1994, **3**:567–574.
This paper describes a method that is based on a genetic algorithm for designing protein sequences which fit a given fold. The search is based on a potential energy function which contains statistically derived pairwise contact potentials (i.e. potentials of mean force) and solvation energies. Meaningful sequences are only obtained if the amino acid composition is restrained.

31. •• Sasai M: **Conformation, energy, and folding ability of selected amino acid sequences.** *Proc Natl Acad Sci USA* 1995, **92**:8438–8442.
This paper describes a simulated annealing procedure that designs protein sequences compatible with a given fold. It shows that a naive energy minimization leads to unstable and non-folding sequences. Better sequences are obtained when the energy landscape over a wide region of the conformational space is imposed. The hydrophobicity pattern and the glycine locations are conserved among the foldable sequences.

32. •• Hellinga HW, Richards FM: **Optimal sequence selection in proteins of known structure by simulated evolution.** *Proc Natl Acad Sci USA* 1994, **91**:5803–5807.
This paper describes a computational method that designs sequences that fit in a given protein core, using a simulated-annealing approach in which random walks over sequence and configurational space are performed. The potential energy function driving the search includes terms for van der Waals non-bonded interactions, solvation energies and entropy. Applications to the core of the phage λ c1 repressor are presented.

33. •• Godzik A: **In search of the ideal protein sequence.** *Protein Eng* 1995, **8**:409–416.
This paper describes a Monte Carlo procedure that designs protein sequences that fit a given fold. The native amino acid composition is imposed. To derive foldable sequences, the procedure optimizes the 'energy gap', that is, the difference in energy for a given sequence between its native and the best non-native conformations.

34. Wodak SJ, Rooman MJ: **Generating and testing protein folds.** *Curr Opin Struct Biol* 1993, **3**:247–259.

35. Wang Y, Zhang H, Li W, Scott RA: **Discriminating compact nonnative structures from the native structure of globular proteins.** *Proc Natl Acad Sci USA* 1995, **92**:709–713.

36. •• Burling FT, Brünger AT: **Thermal motion and conformational disorder in protein crystal structures: comparison of multiconformer and time-averaging models.** *Isr J Chem* 1994, **34**:165–175.
To account for thermal motion during the refinement of a protein crystal structure, three methods were tested: conventional refinement, multicopy sampling and time-averaging refinement using molecular dynamics. In the case of penicillopepsin at 1.8 Å, the best results were obtained with simultaneous refinement of between four and eight conformers.