porphyrin complexes. Model compound I has been shown by X-ray crystallography to have two thioether ligands coordinated to Fe(III)[18]. The similarity between the *g*-values of these model compounds and bacterioferritin is striking.

The evidence from the NIR-MCD spectrum and the comparison between the *g*-values of the protein and bis-thioether model compounds leaves little doubt that the haem group in *Ps. aeruginosa* bacterioferritin is ligated by the thioether side chains of two methionine residues. Table 1 indicates that this seems to be a common feature amongst bacterioferritins from a variety of sources. In no bacterioferritin has a haem content of greater than 0.5 haem per protein subunit yet been observed. For example, neither *Azotobacter vinelandii*[4] nor *E.coli.*[20] bacterioferritin apparently contains more than 0.5 haem per subunit and the protein from *Ps.aeruginosa* is reported to have only one haem per five subunits[5]. If 0.5 haem per subunit is the maximum number of haem groups which can bind to bacterioferritin, either the haem group is bound at the interface between a pair of subunits or only half the subunits bind haem. The identification of the haem ligands as methionine does not, at this stage, rule out the latter possibility, but we consider it most likely that the haem group is bound by individual subunits, as is found in other multimeric haemoproteins.

Andrews *et al.*[19] have reported the amino-acid sequence of *E. coli* bacterioferritin and a secondary structure analysis which indicates a high α-helical content consistent with a structure composed of bundles of four α-helices. Such a tertiary structure for bacterioferritin is consistent with its similarity to mammalian ferritins[1], shown by X-ray crystallography to consist of bundles of four α-helices[1,21]. *E. coli* cytochrome b_{562} (ref. 22) and *Rhodospirillum molischianum* cytochrome c' (ref. 23) also have a four-α-helix bundle structure. In these two proteins the haem lies between two pairs of helices so that it is located at one end of the bundle with the haem normal perpendicular to the bundle axis. On the basis of the analysis of Andrews *et al.*[19], a similar location for the haem of bacterioferritin is possible, involving ligation in the *E. coli* protein by Met 1/Met 144 near the N and C termini of the polypeptide chain. Ligation by Met 31/Met 119 or Met 120 may be sterically possible. A partial sequence of the related protein from *Nitrobacter winogradskyi*[24] shows that Met 31 is not conserved, however, and hence cannot be a ligand, suggesting that the Met 1/Met 144 site is likely. □

Received 10 April; accepted 21 June 1990.

1. Ford, G. C. *et al. Phil. Trans. R. Soc.* B**304,** 551–565 (1984).
2. Laulhere, J. P., Lescurie, A. M. & Briat, J. F. *J. biol. Chem.* **263,** 10289–10294 (1988).
3. Yariv, J. *et al. Biochem. J.* **197,** 171–175 (1981).
4. Stiefel, E. I. & Watt, G. D. *Nature* **279,** 81–83 (1979).
5. Moore, G. R., Mann, S. & Bannister, J. V. *J. Inorg. Biochem.* **28,** 329–336 (1986).
6. Smith, J. M. A., Ford, G. C. & Harrison, P. M. *Biochem. Soc. Trans.* **16,** 836–838 (1988).
7. Smith, J. M. A., Quirk, A. V., Plank, R. W. H., Diffin, F. M., Ford, G. C. & Harrison, P. M. *Biochem. J.* **255,** 737–740 (1988).
8. Moore, G. R. *Biochem. J.* **227,** 341–342 (1985).
9. Moore, G. R. & Pettigrew, G. W. *Cytochromes c: Evolutionary, Structural and Physico-chemical Aspects* 370–372 (Springer, New York, Heidelberg, 1990).
10. Mathews, F. S. *Prog. Biophys. molec. Biol.* **45,** 1–56 (1985).
11. Dickerson, R. E. & Timkovich, R. *The Enzymes* 3rd edn, Vol. 11 (ed. Boyer, P. D.) 397–547 (Academic. London, 1975).
12. Rigby, S. E. J. *et al. Biochem. J.* **256,** 571–577 (1988).
13. Gadsby, P. M. A. & Thomson, A. J. *J. Amer. chem. Soc.* **112,** 5003-5011 (1990).
14. Smith, D. W. & Williams, R. J. P. *Struct. Bonding* **7,** 1–45 (1970).
15. Cheng, J. C., Osborne, G. A., Stephens, P. J. & Eaton, W. A. *Nature* **241,** 193–194 (1973).
16. Sievers, G., Gadsby, P. M. A., Peterson, J. & Thomson, A. J., *Biochim. biophys. Acta* **742,** 637–647 (1983).
17. Moore, G. R., Williams, R. J. P., Peterson, J., Thomson, A. J. & Mathews, F. S. *Biochim. biophys. Acta* **829,** 83–96 (1985).
18. Mashiko, T., Reed, C. A., Haller, K. J., Kastner, M. E. & Scheidt, W. R. *J. Am. chem. Soc.* **103,** 5758–5767 (1981).
19. Andrews, S. C., Smith, J. M. A., Guest, J. R. & Harrison, P. M. *Biochem. biophys. Res. Commun.* **158,** 489–496 (1989).
20. Smith, J. M. A., Ford, G. C., Harrison, P. M., Yariv, J. & Kalb A. J. *J. molec. Biol.* **205,** 465–467 (1989).
21. Harrison, P. M. *et al.* in *Proteins of Iron Storage and Transport* (eds Spik, G. *et al.*) 67–79 (Elsevier, Amsterdam, 1985).
22. Lederer, F., Glatigany, A., Bethge, P. H., Bellamy, H. D. & Mathews, F. S. *J. molec. Biol.* **148,** 427–448 (1981).
23. Weber, P. C., Howard, A., Xuong, N. G. H. & Salemme, F. R. *J. molec. Biol.* **153,** 399–424 (1981).
24. Kurokawa, T., Fukumori, Y. & Yamamoka, T., *Biochem. biophys. Acta* **976,** 135–139 (1989).

ACKNOWLEDGEMENTS. We thank SERC and the Wellcome Trust for support of this work and Professor P. M. Harrison for valuable discussions on the structure of bacterioferritin. The gift of a sample of *E. coli* bacterioferritin from S. C. Andrews, J. R. Guest and P. M. Harrison is gratefully acknowledged.

Implications of thermodynamics of protein folding for evolution of primary sequences

E. I. Shakhnovich* & A. M. Gutin

Institute of Protein Research, Academy of Sciences of the USSR, 142292 Puschino, Moscow Region, USSR

NATURAL proteins exhibit essentially two-state thermodynamics, with one stable fold that dominates thermodynamically over a vast number of possible folds, a number that increases exponentially with the size of the protein. Here we address the question of whether this feature of proteins is a rare property selected by evolution or whether it is in fact true of a significant proportion of all possible protein sequences. Using statistical procedures developed to study spin glasses, we show that, given certain assumptions, the probability that a randomly synthesized protein chain will have a dominant fold (which is the global minimum of free energy) is a function of temperature, and that below a critical temperature the probability rapidly increases as the temperature decreases. Our results suggest that a significant proportion of all possible protein sequences could have a thermodynamically dominant fold.

We investigate a model of a protein chain in which protein folds are characterized by sets of coordinates of C_α atoms $\{r_i^m\}$ where index i denotes the position of an amino-acid residue in the sequence and m denotes the fold. Monomers are positioned in sites of a three-dimensional lattice to account for steric interactions. Energy of a fold m, say, is represented in a simple form:

$$E_m = \sum_{i,j}^{N} B_{ij}\Delta(r_i^m - r_j^m) \qquad (1)$$

where N is the total number of monomeric units, B_{ij} is the interaction energy between monomers i and j when they are adjacent in space; this energy depends on the types of these monomers. $\Delta(r_i^m - r_j^m) = 1$ if monomers i and j are lattice neighbours and 0 otherwise.

The Boltzmann probability of a fold p_m is given by

$$p_m = \exp(-E_m/k_BT)/Z$$

where $Z = \sum_m^M \exp(-E_m/k_BT)$ is the partition function of the chain. M is the total number of folds; it grows exponentially with N so that $M = \gamma^N$ and γ is the number of conformations per monomer.

The existence of a unique structure of a protein implies that one fold m_0 corresponding to this structure dominates thermodynamically over all other folds so that:

$$p_{m_0} = 1 - \varepsilon \qquad (2)$$

where $\varepsilon \ll 1$.

To determine the probability that a random protein folds into a unique structure, we assume the interaction energies B_{ij} to have random values. There are 20 types of amino-acid residues and, hence, $210 \gg 1$ types of interactions between them. Therefore we assume the probability distribution of B_{ij} to be

* Present address: Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA.

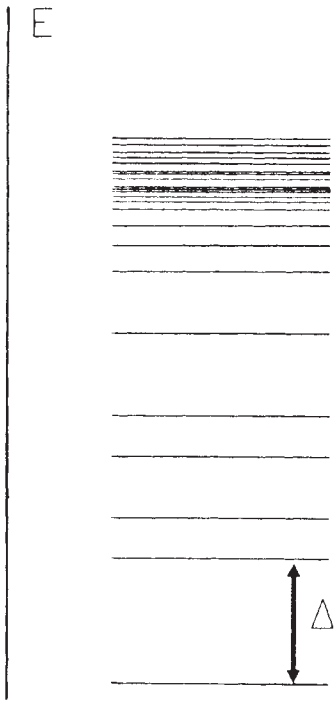

FIG. 1 Typical energy spectrum of a protein. Each line corresponds to a fold. The total number of energy levels is the same as the number of folds, γ^N. The energy of each fold has a random value as we consider a random realization of a sequence. The vast majority of folds belong to the top quasicontinuous part of the spectrum, which corresponds to disordered conformations with energies lying in the range $(\bar{E}-B\sqrt{2\rho N}, \bar{E}+B\sqrt{2\rho N})$. Few levels ($\sim N$) lie in the bottom part of the spectrum; these levels correspond to folds with best-fit contacts and have energies $\sim\bar{E}-Ne$ where $e=B\sqrt{2\rho \ln \gamma}$. At high temperature the energy of a chain corresponds to the top part of the spectrum which is entropically favourable; decrease of temperature leads to the movement downwards of the spectrum and at some definite temperature, T_c, a transition takes place to the bottom discrete part of the spectrum which is entropically unfavourable but favourable energetically. The density of levels in the discrete part of the spectrum is so low that the protein cannot jump from level to level (from fold to fold) by thermal fluctuations. This is the reason for the freezing transition at $T=T_c$ when only a few lowest-lying states become available to the protein. The transition temperature T_c is universal for all sequences but the actual number of thermodynamically relevant folds at $T<T_c$ depends on specific features of the bottom part of the spectrum for a given sequence and especially on the energy gap Δ between the ground fold and the first 'excited' fold.

continuous and gaussian; that is:

$$P(B_{ij})=\frac{1}{\sqrt{2\pi B^2}}\exp\left[-\frac{(B_{ij}-B_0)^2}{2B^2}\right] \qquad (3)$$

where B_0 is the mean which determines overall compactness and B is the variance which corresponds to the heterogeneity of the chain.

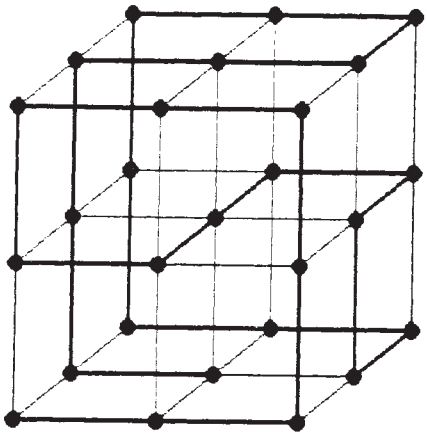

FIG. 2 The lattice and example of a compact self-avoiding walk of a chain of 27 monomers (thick line). Exhaustive enumeration by a first-depth algorithm yields all possible folds unrelated by symmetry. The total number of folds is 103,346. Two hundred realizations of sequences were generated with a distribution of contact energies between monomers B_{ij} given by equation (3). The mean B_0 is unimportant as we consider compact conformations with fixed number of contacts and this parameter gives a constant contribution to energy of each conformation. Therefore we assumed $B_0=0$. Standard variance B determines T_c. We estimated it according to data on inter-residue contact energies in proteins given in ref. 8: $B=0.6$ kcal m^{-1} which yields $k_BT_c\approx 0.9$ kcal m^{-1}. But the particular value of T_c is also not crucial here since we use dimensionless parameter T/T_c. Energy of each fold for each sequence was calculated according to equation (1) in which only nearest neighbours in space were assumed to interact. Boltzmann probabilities p_m ($m=1,\ldots 103{,}346$) of all folds for 200 realizations of sequences (sets of B_{ij}) were determined and parameter X (equation (5)) was evaluated for each sequence at different temperatures.

The thermodynamics of the chain is completely determined by the set of energies of all folds E_m, that is, by the energy spectrum (see Fig. 1). The statistical properties of this energy spectrum were obtained in ref. 1 where the microscopic model defined by equations (1) and (3) (with polymeric bonds taken into account explicitely) was investigated analytically by a replica approach. The result reads that energies of different folds can be treated as statistically independent random values with a gaussian distribution:

$$P(E_m)=\frac{1}{\sqrt{2\pi\rho NB^2}}\exp\left(-\frac{(E_m-\bar{E})^2}{2\rho NB^2}\right) \qquad (4)$$

(ρ is the average number of contacts between residues) and different low-energy folds (bottom lines of the spectrum) of the same sequence are structurally different (this is the reason why their energies are independent random values).

These results demonstrate the equivalence between disordered heteropolymers and the random energy model (REM) introduced by Derrida[2] in the theory of spin glasses. (This equivalence had been postulated *a priori* in ref. 3).

The parameter which gives the effective number of thermo-

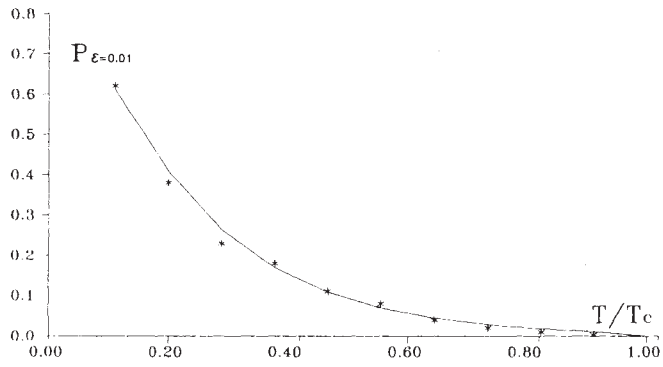

FIG. 3 Fraction of sequences that are able to fold into unique structure, that is, having Boltzmann probability of ground state $p_{m_0}>0.99$ (or $X<0.02$) plotted against T/T_c. The solid line denotes the analytical result of equation (6), asterisks denote the numerical results obtained from exhaustive enumeration of all conformations for 200 sequences.

dynamically relevant folds is

$$X = 1 - \sum_{m=1}^{M} p_m^2 \qquad (5)$$

(note that $\sum_{m=1}^{M} p_m = 1$). When one state m_0 dominates, $p_{m_0} \approx 1$ and $X \approx 0$. In the opposite case when all microstates have equal Boltzmann probabilities, $p_m = \gamma^{-N}$ and $X = 1 - \gamma^{-N} \approx 1$.

The equivalence of random heteropolymers to the REM gives the main clue to the estimation of the probability that a random chain will fold. The probability distribution of X had been calculated for spin glasses[4] and explicitly for the REM[5]. (For the general case of a disordered system that possesses an energy spectrum as shown in Fig. 1, with quasicontinuous top part and discrete bottom part, the same probability distribution has been derived[6].)

Using these results, it is easy to derive the probability P_ε that for a random chain a ground fold m_0 will dominate, that is, $p_{m_0} > 1 - \varepsilon$ and, hence, to the same order in ε, $X < 2\varepsilon$:

$$P_\varepsilon = \frac{\sin(\pi X_0)}{\pi X_0} \varepsilon^{X_0} \qquad (6)$$

where X_0 is X averaged over all sequences[4,5].

X_0 was found[1] as a function of temperature, T:

$$X_0 = \begin{cases} T/T_c & \text{if } T/T_c < 1 \\ 1 & \text{if } T/T_c \geq 1 \end{cases} \qquad (7)$$

where

$$T_c = \frac{B\sqrt{\rho}}{2k_B\sqrt{\ln \gamma}} \qquad (8)$$

T_c is the temperature when the chain backbone obtains a unique conformation. It may not coincide with T_d, the temperature of denaturational transition when fixation of rotational isomers of side chains and their tight packing occurs[7]. When $T_c > T_d$, a 'molten globule' with a unique backbone conformation would exist in the temperature range $T_c > T > T_d$.

Equations (6)–(8) give the main result of this work: at high temperatures, $T > T_c$, $X_0 \equiv 1$ therefore (see equation (6)) $P_\varepsilon \equiv 0$ and there are no sequences which can fold. When the temperature decreases below T_c the fraction of sequences that are able to fold grows drastically. For example, taking $\varepsilon = 0.01$, which corresponds to 99% ground fold dominance, we obtain $P_\varepsilon \approx 0.1$ at $T = T_c/2$. This means that under this condition every tenth sequence will have one fold with 99% dominance.

This result is universal: all microscopic characteristics of folded chains such as heterogeneity B, entropy per bond $\ln \gamma$, coordination of monomers ρ, are introduced via the single parameter T_c so that the main result, equation (6) depends only on the dimensionless parameter $X_0(T/T_c)$. It should be emphasized also that the 'folding capacity' of a random chain does not depend on its length, N. The above investigation implies also that folding of a random chain could lead to the global minimum of free energy.

We tested the analytical result equation (6) by exhaustive enumeration of all self-avoiding compact conformations of a 27-monomer chain in a 3*3*3 fragment of a simple cubic lattice (Fig. 2). The results (Fig. 3) match the analytical expression in equation (6).

We conclude that the requirement for folding may not be very restrictive for evolutionary choice of protein sequences. □

Received 30 November 1989; accepted 11 June 1990.

1. Shakhnovich, E. I. & Gutin, A. M. *Biophys. Chem.* **34,** 187–199 (1989).
2. Derrida, B. *Phys. Rev.* **B24,** 2613–2624 (1981).
3. Bryngelson, J. & Wolynes, P. G. *Proc. natn. Acad. Sci. U.S.A.* **84,** 7524–7528 (1987).
4. Mezard, M., Parisi, G., Sourlas, N., Toulouse, G. & Virasoro, M. *J. Phys.* (France) **45,** 843–854 (1984).
5. Derrida, B. & Toulouse, G. *J. Phys. (France) Lett.* **46,** L223-L225 (1985).
6. Mezard, M., Parisi, G. & Virasoro, M. *J. Phys. (France) Lett.* **46,** L217–L220 (1985).
7. Shakhnovich, E. & Finkelstein, A. *Biopolymers* **28,** 1667–1694 (1989).
8. Myazawa, S. & Jernigan, R. *Macromolecules* **18,** 534–552 (1985).

ACKNOWLEDGEMENTS. We thank O. B. Ptitsyn for fruitful discussions. E.I.S. thanks the staff of Service de Physique Theorique C.E.N. Saclay (Paris) for hospitality during his stay in Saclay where part of this work was done.