TOPICAL REVIEW — 8th IUPAP International Conference on Biological Physics

# Knowledge-based potentials in bioinformatics: From a physicist's viewpoint*

Zheng Wei-Mou(郑伟谋)†

*Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing* 100190, *China*

(Received 29 April 2015; revised manuscript received 14 June 2015; published online 20 October 2015)

Biological raw data are growing exponentially, providing a large amount of information on what life is. It is believed that potential functions and the rules governing protein behaviors can be revealed from analysis on known native structures of proteins. Many knowledge-based potentials for proteins have been proposed. Contrary to most existing review articles which mainly describe technical details and applications of various potential models, the main foci for the discussion here are ideas and concepts involving the construction of potentials, including the relation between free energy and energy, the additivity of potentials of mean force and some key issues in potential construction. Sequence analysis is briefly viewed from an energetic viewpoint.

**Keywords:** Knowledge-based potentials for proteins, protein native structures, sequence signals

**PACS:** 87.10.–e, 02.50.–r **DOI:** 10.1088/1674-1056/24/12/128701

## 1. Introduction

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. We are facing ever growing huge amounts of biological raw data including genomic sequences and protein structures. Bioinformatics has become an important part of many areas of biology, combining biology, computer science, mathematics, physics and chemistry.

Physics forms the theoretical basis of all fields of natural science. Many ideas in data mining are borrowed from physics. Ising models find widespread applications in various fields of machine learning, including linguistics, computer vision, and artificial intelligence. In the seminal publication[1] by John Hopfield, a neural network model for machine learning was connected to physics and statistical mechanics of spin glasses. Boltzmann machines, as the stochastic, generative counterpart of Hopfield networks, have imported a variety of concepts and methods from statistical mechanics.[2] The explicit analogy drawn with statistical mechanics in the Boltzmann machine formulation led to the use of terminology borrowed from physics. As one more example, simulated annealing, which is a generic tool for the global optimization problem for locating the global optimum of a given function in a large search space, mimicking heating and cooling to affect both the temperature and the thermodynamic free energy.

Structures of many proteins have been determined, but much more are proteins with known sequences whose structures are still unknown. Extracting knowledge from the raw data of structures in order to understand protein folding and guide protein design is a central task in the structural biology. Here I shall discuss knowledge-based potentials for proteins used in bioinformatics. There are many reviews on this topic, most of which are written for practicers in bioinformatics, but not for physicists who are not working on bioinformatics. Contrary to these reviews, I would concentrate on ideas and concepts in the construction of knowledge-based potentials rather than technical details.

## 2. Energy versus free energy

In bioinformatics one among the problems which are most closely related to physics is how to construct knowledge-based potential functions for protein structures. Proteins are molecular machines and building blocks of living cells. A knowledge of the three-dimensional structure of a protein is necessary for understanding how it functions.

Anfinsen's dogma (also known as the thermodynamic hypothesis) claims that, at least for small globular proteins, the native structure of a protein is determined only by its amino acid sequence.[3] The native structure is regarded as a unique, stable and kinetically accessible minimum of the free energy. Kinetical accessibility means that the folding of the chain happens spontaneously, and must not involve highly complex changes in the shape.

How a protein reaches its native structure is the subject of the field of protein folding. The number of possible conformations available to a given protein is astronomically large. It is impossible for a protein, even a small protein of 100 residues, to explore all possible conformations and choose the appropri-

*Project supported in part by the National Natural Science Foundation of China (Grant Nos. 11175224 and 11121403).

†Corresponding author. E-mail: zheng@itp.ac.cn

© 2015 Chinese Physical Society and IOP Publishing Ltd 

ate one as the native structure. This is known as the Levinthal paradox.[4] In fact, according to the argument of reduction to absurdity by Levinthal, the folding process should not be ergodic. Levinthal suggested that protein folding is sped up and guided by the rapid formation of local interactions which then determine the further folding of the peptide. Thus, a few 'route markers' of local amino acid segments carrying stable interactions would be embedded in protein sequences to guide the folding. This makes the real protein sequences rare and distinct from the huge number of non-protein sequences. That is, in opposition to the more or less homogenous polymers often considered in physics, the inhomogeneity plays an essential role for proteins.

Energy or Hamiltonian is a microscopic quantity of a system. In contrast, free energy is a macroscopic quantity referring to an ensemble or a distribution of microscopic states. What should we mean then by 'minimum of the free energy' for protein native structure? While the term energy for a protein refers to the conformational energy of an individual chain evaluated as a function of variables of the conformation and its surroundings, the free energy is evaluated with respect to a set of configurations or a distribution of configurations. The kinetical accessibility of the native state of a protein implies the non-ergodicity of protein folding. This makes the configurational distribution for the free energy calculation, even the supporting set of the distribution, hardly determined.

When calculating the energy of a conformation of a protein, the chain connectivity of the protein is explicitly considered. However, when constructing knowledge-based potentials for protein structures, the free energy is only referred to various simplified models of proteins. Except for theoretical discussions of the cooperative formation of alpha-helices or beta-sheets, in most models of proteins the chain connectivity is almost ignored. (Sometimes the sequence distance of two residues is included in potential models as a parameter.) Proteins are then regarded as a system consisting of mixtures of multi-component molecular species. A given native structure is regarded as a configuration sample taken from a 'liquid-like' thermodynamical phase of the mixture; the energy of the configuration, instead of free energy, is then calculated using some effective potentials.

## 3. Potentials of mean force

A long-standing goal in protein structure theory is the development of energy functions for the protein-solvent system to determine native folds solely from the information contained in amino acid sequences. The force fields or energy functions obtained from quantum mechanical calculations and then extrapolated to the macromolecular level are too complicated to achieve the goal. Furthermore, such physics-based potential functions need not fully capture all of the important physical interactions. The interactions encountered in proteins are very complicated.

The known three-dimensional structures of proteins contain a large amount of information on the forces stabilizing proteins. It is believed that potential functions and the rules governing protein stability can be revealed from statistical analysis on known structures. Thus, an alternative way to extract the effective potentials is to take the known native structures as the only reliable source of information.

Assume that we are dealing with the classical Hamiltonian of a system consisting of $N$ particles:

$$\mathcal{H}(\boldsymbol{r}^N,\boldsymbol{p}^N)=K(\boldsymbol{p}^N)+U(\boldsymbol{r}^N), \tag{1}$$

where $(\boldsymbol{r}^N,\boldsymbol{p}^N)$ denotes the microscopic state, $K(p^N)$ is the kinetic energy, and $U(r^N)$ is the potential energy. The distribution among the microscopic states in the thermodynamical equilibrium is governed by Boltzmann's principle:

$$P(\boldsymbol{r}^N,\boldsymbol{p}^N)\propto\exp\{-\beta[K(\boldsymbol{p}^N)+U(\boldsymbol{r}^N)]\}. \tag{2}$$

After an integration over the momentum variables, the probability density for conformation $\boldsymbol{r}^N$ is given by

$$P(\boldsymbol{r}^N)=Q_N^{-1}\exp\{-\beta U(\boldsymbol{r}^N)\}, \tag{3}$$

where $Q_N$ is the conformational integral:

$$\begin{aligned}Q_N&=\frac{1}{N!}\int \mathrm{d}\boldsymbol{r}_1\cdots\mathrm{d}\boldsymbol{r}_N\exp\{-\beta U(\boldsymbol{r}^N)\}\\&\equiv\frac{1}{N!}\int \mathrm{d}^N\boldsymbol{r}\,\mathrm{e}^{-\beta U}.\end{aligned} \tag{4}$$

We shall concentrate only on this non-trivial conformational part coming from $U(\boldsymbol{r}^N)$.

The Boltzmann's principle relates the energy $U$ of a conformation to its equilibrium probability. The 'inverse Boltzmann's principle' associates a potential to a reduced or marginal distribution function. This potential is the so-called potential of mean force (PMF).[?] For example, the pair distribution function is obtained by integrating over the coordinates of $(N-2)$ particles:

$$\begin{aligned}n_2(\boldsymbol{x}_1,\boldsymbol{x}_2)&=\int \mathrm{d}\boldsymbol{r}_1\cdots\mathrm{d}\boldsymbol{r}_N P(\boldsymbol{r}^N)\sum_{i\neq j}\delta(\boldsymbol{x}_1-\boldsymbol{r}_i)\delta(\boldsymbol{x}_2-\boldsymbol{r}_j)\\&=\left\langle\sum_{i\neq j}\delta(\boldsymbol{x}_1-\boldsymbol{r}_i)\delta(\boldsymbol{x}_2-\boldsymbol{r}_j)\right\rangle.\end{aligned} \tag{5}$$

The potential associated with $n_2$ is

$$w_2(\boldsymbol{x}_1,\boldsymbol{x}_2)\equiv-\beta^{-1}\log n_2(\boldsymbol{x}_1,\boldsymbol{x}_2), \tag{6}$$

which describes the averaged effective interaction between two particles in the background of the remaining other particles.

Often we have to include the background, e.g. water molecules surrounding a protein, in our model, but we have

no intrinsic interest in knowing its details. The degrees of freedom of the background can be removed by an integration over them. Thus, the effective potential of a protein in water solvent is

$$W_p\left(\boldsymbol{r}^N\right) = -\beta \log\left\{ Q^{-1} \int \mathrm{d}^M \boldsymbol{q} \exp[U_{\mathrm{p}}(\boldsymbol{r}^N) + U_{\mathrm{w}}(\boldsymbol{q}^M) + U_{\mathrm{pw}}(\boldsymbol{r}^N, \boldsymbol{q}^M)] \right\}, \quad (7)$$

where $Q$ is the configurational integral of the joint protein-solvent system, $U_{\mathrm{p}}$, $U_{\mathrm{w}}$, and $U_{\mathrm{pw}}$ are potentials of protein, that of water and their interaction, respectively. By keeping only averaged effects an approximation like the mean-field theory can be developed.

## 4. Some key issues in the construction of potentials for proteins

The interpretation of probabilities by taking their negative logarithms as potentials is convenient. Correlations among variables are converted to interactions in potentials. To include as many as possible important correlations is essential for estimates of probabilities. However, it should be emphasized that the correlations we can infer are dependent on the sample size. For a small sample size we cannot distinguish a sampling fluctuation from a true correlation.

As mentioned above, a protein is modeled as a mixture of multi-component molecular species. We usually use coarse-grained representations of polypeptide chains. We reduce the protein structure into units or centers of residues, or other types of simplified structural units, depending on how the specific models are constructed. Such centers are regarded as various molecular species, and their spatial distances are grouped into few bins.[6,7]

In the simplest model a residue is represented by just one center, and there are only two distance bins, i.e. 'contacted' and 'non-contacted' with respect to a cut-off distance or some other criteria.[8] The Miyazawa-Jernigan (MJ) contact potential for globular proteins is a widely used knowledge-based potential built on such a model.[9,10] The MJ potential has only 210 energy values $e_{ab}$ to characterize the tendency of a residue of identity $a$ in contact with another residue of identity $b$ in a native structure. There is an additional parameter of the cut-off distance for the contact definition. The probability for a conformation $C$ of a chain with sequence $S = a_1 a_2 \ldots a_n$ to be native is estimated as

$$P(C|S) \propto \exp\left\{ -\sum_{i=1}^{n} \sum_{j>i}^{n} e_{a_i a_j} \chi_{ij} \right\} \equiv \exp\{-U(S,C)\}, \quad (8)$$

where $\chi_{ij}$ is the characteristic function of contact, i.e. $\chi_{ij} = 1$ if residues $i$ and $j$ are in contact, and $\chi_{ij} = 0$ otherwise. In terms of distance $r_{ij}$, cut-off $r_{\mathrm{c}}$, and the step function $\Theta$, we have $\chi_{ij} = \Theta(r_{\mathrm{c}} - r_{ij})$.

In the MJ potential all correlations higher than two-body ones have been ignored. Furthermore, the additivity of the energy terms implies that each residue pair contributes $P(C|S)$ independently. This independency assumption is one of the most important issues in the potential construction. A more general form of potential could be

$$U = \sum_i h_{a,\sigma} \chi_{a,\sigma}(a_i, \sigma_i) + \sum_{i,j>i} e_{ab,\tau} \chi_{ab,\tau}(a_i, a_j; \tau_{ij}), \quad (9)$$

where one-body term $h_{a,\sigma}$ may be interpreted as an external field exerting on a residue (or one of its representative centers) with feature or descriptor $\sigma$, the characteristic function $\chi_{a,\sigma}(a_i, \sigma_i) = \delta_{aa_i}\delta_{\sigma\sigma_i}$, and the meaning of $\tau$ is analogous to $\sigma$. An example of $\sigma$ is the binary partition of residues as 'exposed' and 'buried' according to the accessible surface area or number of surrounding neighbors inside a cut-off sphere, and an example of $\tau$ could be the pair spatial distance, hydrogen-bonding status, or the secondary structure categories of the pair.

Generally, one-body terms contain the least number of parameters, and hence can be estimated most reliably. By an eigenvalue decomposition analysis of the $20 \times 20$ symmetric matrix representing the MJ potential, it is found that contact energies $e_{ab}$ are, in fact, dominant by the 'one-body' desolvation terms.[11] This can also be verified by a parameter fitting.[12] Thus, it is advisable to first extract one-body terms and then include two-body terms as further corrections in constructing potentials.

After taking $e_{ab} = u_a + u_b$ as a good approximation, the probability for structure $C$ of sequence $S$ to be native is estimated as

$$P(C|S) \propto \exp\left\{ -\sum_{i=1}^{n} \sum_{j>i}^{n} e_{a_i a_j} \right\} = \frac{1}{2} \exp\left\{ -\sum_{i=1}^{n} u_{a_i} \nu_i \right\}, \quad (10)$$

where $\nu_i$ is the number of residues in contact with $a_i$. A direct derivation of the one-body $u_a$ in the framework of the MJ potential model is given in Ref. [13].

(From the probabilistic viewpoint, an approximation of the joint probability $P(C,S)$ under an independency assumption may be written as

$$P(C,S) = \prod_i P(a_i, \sigma_i), \quad (11)$$

which may be interpreted as a mean field approximation. To include correction from two-body correlation, we may write

$$P(C,S) = \prod_i P(a_i, \sigma_i) \left[ \prod_j \frac{P(a_j, \tau_{ij} | a_i, \sigma_i)}{P(a_j, \sigma_j)} \right]^{1/\nu_i}, \quad (12)$$

where we have assumed that the combination $(\sigma_i, \sigma_j)$ is consistent with $\tau_{ij}$. If the one-body independency assumption is valid, factors of the product inside the square bracket are of

value 1, the one-body form (11) is then recovered. Here, exponent $1/\nu_i$ is added to remove multi-counting of pair effects, but need not be optimal. The pair corrections here will be converted to pair interaction in potentials. It should be mentioned that to avoid multiple counting of one and the same effect is rather tricky.

In real cases, independency assumptions are usually not valid. However, conditional independency can still be found as good approximations by designing features or descriptors carefully according to the nature of problems. For example, after a structure is referred to as a sequence $\{\nu_i\}$ of coordinate numbers, usually we do not have the independency for the joint probability $P(C,S)$, i.e., $P(C,S) \neq \prod_i P(a_i,\nu_i)$, since the relation among $\nu_i$'s is extremely intricate. However, $P(S|C) = \prod_i P(a_i|\nu_i)$, which well catches the main effects of hydration, could still be a good approximation.

How can we examine the validity of independency of a given model? Suppose that a good set of representative proteins is given, which are extracted from various types of structures, and any two samples from the set are not too similar either in sequence or in structure. With such a set we can easily estimate the probability of an amino acid, say alanine ($A$):

$$P(A) = N(A) \Big/ \sum_x N(x)\,, \tag{13}$$

where $N(x)$ denotes the total number of residues of amino acid type $x$ in the set. The normalization constraint $\sum_x P(x) = 1$ makes only 19 of the 20 parameters being free. With a subset obtained by randomly choosing samples from the full set the estimated amino acid frequencies, which are now denoted by $Q(x)$, will change little. A convenient measure of distance between two distributions $P(x)$ and $Q(x)$ is the Kullback-Leibler divergence defined by

$$D(P,Q) = \sum_x P(x)\log[P(x)/Q(x)], \tag{14}$$

which is known also as relative entropy, and is non-negative (and asymmetric between $P$ and $Q$).[14] Under the assumption that both distributions $P$ and $Q$ are close, the leading term of $D(P,Q)$ is a $\chi^2$ variable.[15] (A $\chi^2$ variable with $k$ degrees of freedom is a sum of the squares of $k$ independent standard normal random variables.) Considering the above binary partition of residues into 'exposed' ($e$) and 'buried' ($b$), we may estimate two amino acid distributions $P_e(x)$ and $P_b(x)$ conditional on the two categories (with descriptors $\sigma \in \{e,b\}$). The difference between $P_e(x)$ and $P_b(x)$ or $D(\{P_e(x)\},\{P_b(x)\})$ is large: hydrophobic residues have high values of $P_b(x)$ while values of $P_e(x)$ are high for polar ones, showing the hydration effect. This indicates the strong correlation between the hydrophobicity and environmental class of residues. There are many standard tests on statistical independency. The log-linear analysis is a useful for constructing potentials in this respect.[16,17]

The standard statistical mechanics of a simple gas (of single component without any internal degree of freedom) in equilibrium starts with a pairwise additive potential

$$U(\boldsymbol{r}^N) = \frac{1}{2}\sum_{i,j} u(\boldsymbol{r}_i - \boldsymbol{r}_j), \tag{15}$$

where $u(\boldsymbol{r}_i - \boldsymbol{r}_j) \equiv u_{ij}$ is the 'naked' interaction between particle pair $i$ and $j$. The Boltzmann distribution is $P(\boldsymbol{r}^N) \propto \exp\{-\beta U(\boldsymbol{r}^N)\}$. The potential $w_2(\boldsymbol{r}_1,\boldsymbol{r}_2)$ associated with the reduced pair distribution $n_2$ in Eq. (6), as a potential of mean force, contains the effect of all other particles. Usually, $w_2$ does not possess the additivity. Pair distribution $n_2$ (and hence $w_2$) can be directly extracted from a sample set, but it is usually rather complex to infer $u_2$. A well-developed theory relating $u_2$ to $w_2$ is the Ornstein–Zernike equation together with an appropriate closure relation (which looks like the Dyson equation $G = G_0 + G_0\Sigma G$ for Green's function).[18] This idea has been used to compute the effective potentials between amino acid residues.[19] A simple example to illustrate the difference between $u_2$ and $w_2$ is the multi-variate normal distribution where the latter corresponds to the variance, and the former to correlation parameters of the distribution. They are related as inverse of matrices.[20,21]

A main goal of constructing potentials for proteins is to determine the native structure of a protein from its sequence. In order to understand how a protein folds to its native structure, we require that potentials should dynamically work, so that they can be used in a simulation like molecular dynamics to mimic the folding process. We are still in a position somewhat far from this intriguing destination. In most cases, potentials are used 'statically' to discriminate between native and non-native conformations. For such a purpose, potentials are essentially the effective 'free energies' of native structures with respect to certain reference conformations, or energy differences between native and reference structures. One direct way is to construct potentials separately for native and reference structures and then take their difference. However, there are alternatives based on optimization under some criteria, e.g., maximal energy gap between known native conformations and decoy ones. Such potentials depend not only on native structures, but also on decoy structures in given training sets. It should be emphasized that to create a set of reference conformations for a general purpose is by no means a simple task.

The advantage of knowledge-based energy functions is that they can model any behaviors seen in known protein structures, even if a good physical understanding of the behavior does not exist. However, a good physical understanding is always helpful in designing as simple as possible models of conditional independency to include important correlations. Recently, a simple and efficient statistical potential was reported, which has only a limited number of parameters but performs

with an unprecedented reliability.[22] When constructing the potential, the one-body feature is also the above $\sigma \in \{e,b\}$. The energy term, say $u_e(a)$ for amino acid $a$ being 'exposed', is estimated as

$$\begin{aligned} u_e(a) &= -\log[P(a|e)/P(a)] \\ &= \log P(a) + \log P(e) - \log P(a,e), \end{aligned} \tag{16}$$

where the joint probability $P(a,e)$ of amino acid $a$ with feature $e$, and marginal probabilities $P(a)$ and $P(e)$ are inferred from the total number $m_x^\sigma$ of residues of type $x$ found in the environment class $\sigma$ within the training dataset as

$$\begin{aligned} &P(a,e) = m_a^e \big/ \textstyle\sum_{x,\sigma} m_x^\sigma \,, \quad P(e) = \textstyle\sum_x P(x,e), \\ &P(a) = \textstyle\sum_\sigma P(a,\sigma). \end{aligned} \tag{17}$$

(Note that the independency $P(a,e) = P(a)P(e)$ would imply $u_e(a) = 0$.) The feature $\tau$ used for constructing pair energies is five classes of hydrogen-bonding statuses. Similarly, the pair energies are given by $u_\tau(a,b) = -\log[P(a,b|\tau)\big/P(a,b)]$. The disadvantage of knowledge-based potentials is that they are phenomenological and then hardly predict new behaviors absent from the training set.

Another problem related to the protein folding is the protein design.[23] Rational protein design approaches make protein-sequence predictions that will fold to specific structures. It is then also referred to as inverse folding: an unknown sequence is found to fold to a known conformation. The number of possible protein sequences is enormous, growing exponentially with the length of the protein chain, only a subset of them will fold reliably and quickly to a single native state. With a potential function given, a sequence 'mounted' on the target structure has a value of energy. In an ideal case, such energy values will form a normal-like distribution. Candidate sequences are expected to fall in the lower tail of the distribution. The space spanned by random sequences is not physical. Thus, we do not have the Boltzmann distribution in a rigorous sense. The tail of extremely low energies of the candidate sequences usually has an exponential form, which is described by the theory of large deviations.[24] A parameter of the so-called rate function appears, playing a role of temperature in the Boltzmann distribution, but having nothing to do with a real temperature.[25]

## 5. Discovery of sequence signals: From an energetic viewpoint

The discovery of patterns or signals shared by several sequences that differ greatly is a basic task in biological sequence analysis, and still a challenge. For example, a class of gene regulation proteins contains the helix-turn-helix (HTH) motif for DNA binding as a signal with a width of about 18 residues. A simple model describing a signal or a portion of local multiple sequence alignment is the position-specific weight matrix (PSWM) model, in which a signal is characterized with position-specific probabilities. Thus, the HTH signal may be represented with a matrix $\boldsymbol{M} = (P_{\alpha i})_{20\times 18}$, where entry $P_{\alpha i}$ indicates the probability for amino acid of type $\alpha$ to occur at position $i$ of the motif. Generally, a signal is recognized with respect to a noise or background. The simplest model for the background is a position independent distribution $\{Q_\alpha\}$. According to the inverse Boltzmann principle, we introduce two sets of energies: $\{w_i(\alpha) = -\log P_{\alpha i}\}$ and $\{b(\alpha) = -\log Q_\alpha\}$. We imagine that a residue may be painted either 'white' or 'blue', and then it acquires an energy $w$ or $b$. Assume that a sequence set sharing a signal is given, and the two sets of energies are known. For a 'coloring' of a sequence from the set, we may compute its total energy. By finding the coloring whose total energy is minimal, the signal location on that sequence is determined. On the other hand, if the signal position of each sequence is known, i.e. all the sequences have been aligned according to the signal region, we may infer the two sets of energies. However, at the beginning we know neither the energies (Hamiltonian of the system) nor the signal positions on each sequence. We have to solve the problem self-consistently. A common way to break the loop is to make a random initialization. However, we make some guesses based on our previous knowledge.[26,27] The signal discovery is then to minimize the total energy of all sequences in the set. The above discussion may be generalized to various cases. For example, sequences may share multiple signals, and a signal may be either present or absent or repeated in certain sequence. From the energetic viewpoint, pairwise sequence alignment may be thought of as a chain of one sequence in an external potential generated by the other sequence.

## 6. Concluding remarks

Relevant timescales for protein folding span over many orders of magnitude: from femtoseconds for bond-vibration, picoseconds for isomeration, nanoseconds for water dynamics and helix formation, to milliseconds for typical folders. This is one complexity of the protein folding. Thus, it is reasonable to build several models, each of which catches a certain nature of protein at a specific range of timescales. Associated with each model are certain levels of coarse-graining and corresponding potential functions. For example, to determine the packing of secondary structure elements of a protein (or its structural topology) is essential for structure prediction. So far, no potentials are particularly built and work for this purpose. There are fields for physics to display its prowess.

## References

[1] Hopfield J J 1982 *Proc. Natl. Acad. Sci. USA* **79** 2554

[2] Ackley D H, Hinton G E and Sejnowski T J 1985 *Cognitive Science* **9** 147
[3] Anfinsen C B 1973 *Science* **181** 223
[4] Levinthal Cyrus 1969 *Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House*, Monticello, Illinois USA, pp. 22–24
[5] Chandler D 1987 *Introduction to Modern Statistical Mechanics*, Sect. 7.3. (Oxford: Oxford University Press)
[6] Sippl M J 1995 *Curr. Opin. Struct. Biol.* **5** 229
[7] Lazaridis T and Karplus M 2000 *Curr. Opin. Struct. Biol.* **10** 139
[8] da Silveira C H, Pires D E V, Minardi R C, *et al.* 2009 *Proteins: Struct. Funct. Bioinf.* **74** 727
[9] Miyazawa S and Jernigan R L 1985 *Macromolecules* **18** 534
[10] Miyazawa S and Jernigan R L 1996 *J. Mol. Biol.* **256** 623
[11] Li H and Tang C 1997 *Phys. Rev. Lett.* **79** 765
[12] Pokarowski P, Kloczkowski A, Jernigan R L, Kothari N S, Pokarowska M and Kolinski A 2005 *Proteins: Struct. Func. Bioinf.* **59** 49
[13] Zeng H, Liu K S and Zheng W M 2012 *The Open Bioinf. J.* **6** 1
[14] Kullback S and Leibler R A 1951 *Ann. Math. Statist.* **22** 79
[15] Greenwood P E and Nikulin M S 1996 *A guide to chi-square testing* (New York: Wiley)
[16] Cootes A P, Curmi P M G, Cunningham R, Donnelly C and Torda A E 1998 *Proteins: Struct. Funct. Genet.* **32** 175
[17] Payne R W 1989 *Genstat 5 Reference Manual* (New York: Oxford University Press)
[18] McQuarrie D A 1976 *Statistical Mechanics* (Harper Collins Publishers)
[19] Pliego-Pastrna P and Carbajal-Tinoco M D 2003 *Phys. Rev. E* **68** 011903
[20] Gut A 2009 *An Intermediate Course in Probability* (New York: Springer)
[21] Marks D S, Colwell L J, Sheridan R, Hopf T A, Pagnani A, Zecchina R and Sander C 2011 *PLoS One* **6** e28766
[22] Cossio P, Granata D, Laio A, Seno F and Trovato A 2012 *Sci. Rep.* **2** 351
[23] Dahiyat B I and Mayo S L 1997 *Science* **278** 82
[24] Touchette H 2009 *Phys. Rep.* **478** 1
[25] Finkelstein A V and Ptitsyn O 2002 *Protein Physics: A Course of Lectures* p. 197 (Academic Press)
[26] Lawrence C E, Altshul S, Boguski M, Liu J, Neuwald A and Wootton J 1993 *Science* **262** 208
[27] Zheng W M 2005 *2005 Bioinformatics* **21** 938