Pacific Symposium on Biocomputing 4:266-277 (1999)

MDL AND THE STATISTICAL MECHANICS OF PROTEIN POTENTIALS

T.G. DEWEY
Department of Chemistry and Biochemistry, University of Denver, Denver, CO 80208, USA

The combination of a wealth of structural data and impressive computational power provides detailed information pertaining to the structure and dynamics of biomacromolecules. A natural inclination is to incorporate this information into models to gain added predictive power on protein folding and stability. There has been considerable recent interest in developing "knowledge-based" potentials to describe internal interactions in proteins. In these approaches, probability distribution functions are inferred from existing knowledge. A common assumption has been the "quasi-chemical approximation" or "Boltzmann device". This method relates statistical mechanical probabilities to observed frequencies. The validity of this approach is discussed in detail from a statistical mechanics perspective. Because statistical mechanics is a form of statistical inference based on a lack of knowledge of the system, the "Boltzmann device" does not have a rigorous theoretical justification. In the present work, a statistical mechanics based on partial knowledge of the system is employed. This statistical mechanical scheme uses the minimum description length (MDL) of phase space as its main tool. With this approach, "knowledge-based" potentials can be derived in a rigorous fashion. In practical calculations, these potentials are best obtained using Bayesian inference methods similar to those used in image reconstruction.

1 Introduction: The Problem with the Boltzmann Device

One of the "holy grails" of biochemistry is to find the algorithm that allows a protein's three-dimensional structure to be predicted from its sequence. The solution of this form of the protein folding problem requires an accurate representation of protein potentials. These intermolecular potentials must not only account for the interactions between amino acids within a protein but also must consider solvation effects. To this end, a plethora of techniques have arisen. These techniques fall into two broad categories. The first tries to calculate the interaction potentials either directly from high level quantum calculations or indirectly from empirical force fields derived from direct calculations. An alternate approach is to use the large database of structural information to infer potentials. Potentials derived in this manner are often referred to as "statistical potentials" or "knowledge-based potentials". Essentially, one is using the structural knowledge in an existing database to deduce a potential.

The Boltzmann device is central to the computation of statistical poten-

tials. It has been observed that the frequency of occurrence of certain structural features found in the diverse database of proteins follows the following empirical law:

$$Frequency = \exp\{-E/RT\}. \tag{1}$$

where E is the energy associated with the structural feature, R is the gas constant and T is the "conformational temperature". This law is, of course, the familiar form of the Boltzmann distribution. This distribution has been observed for a number of protein substructures and motifs. These include frequencies of cis-trans isomerization of prolines, distributions of Φ-Ψ dihedral angles, of charged residues, and of sizes of empty cavities. Additionally, residue stabilization of secondary structure follows a Boltzmann law (see [1] and reference therein). Although this phenomenology suggests that conventional Boltzmann statistics is obeyed, early on this result was disputed [2]. Despite concerns regarding the origin of 1, it has been proposed that the Boltzmann law can serve as a rigorous starting point for connecting observed structural frequencies to potential functions [3].

As an example of how the Boltzmann device is used, the calculation of an intermolecular potential for alanine-alanine interactions in a protein is considered (cf. [4]). For knowledge-based methods, some protein database would be examined and the distance, r, between all alanine pairs in all different proteins is measured. The distance scale is discretized, so that the number of pairs within a given range or bin can be counted. It is then assumed that the frequency, $f_{AA}(r)$, of occurrence of alanine pairs at a fixed distance, r, follows the law:

$$f_{AA}(r) = \frac{1}{Z}\exp[-\frac{E_{AA}(r)}{RT}] \tag{2}$$

with the "partition function" defined as:

$$Z = \sum_{r}\exp[-\frac{E_{AA}(r)}{RT}] \tag{3}$$

where the sum is over all discretized distances. The goal is to back calculate E_{AA} from the frequencies.

There are a number of immediate problems with this calculation and these have been circumvented in various ways. The most obvious one is the value to be assigned to the temperature. Often this is taken as room temperature. However, the "temperature" observed when examining protein substructures has been quite variable and has ranged from 150-600 K^o. The physical significance of this remains obscure. In some applications the temperature is not of great concern because it is a constant multiplicative factor.

A second, related problem is what reference state should be assumed. Choice of a reference distance frequency, $f(r)$, yields a net pair potential given by:

$$\Delta E_{AA} = E_{AA} - E_{referencepair} \tag{4}$$

This net potential is than determined from:

$$\Delta E_{AA}(r) = -RT \ln\left(\frac{f_{AA}(r)}{f(r)}\right) + \ln\left(\frac{Z}{Z_{ref}}\right) \tag{5}$$

One possible choice of reference frequencies is to use distances from all amino acid pairs in the database. However, it is unclear how to make an optimal choice of reference states. With an ideal choice, the term $\ln(Z/Z_{ref})$ will be small and the net potential can be determined directly from frequency ratio, an observable quantity. The choice of the reference state remains one of the more difficult issues with this approach [5] and requires good physical intuition for the specific system under consideration.

Despite these criticisms and concerns, knowledge-based potentials have met with modest success (cf. [4 6]). Although knowledge-based potentials have improved considerably over the years, there still is the underlying problem of the validity of the Boltzmann device. Thomas and Dill have examined the residue-residue distance dependence for a protein lattice model [7]. They found that systematic errors arose in the derived potentials as a result of excluded volume effects and that a Boltzmann dependence is not followed. It is uncertain whether a correct treatment of excluded volume effects will re-instate the Boltzmann device. At this stage the entire approach must be viewed as empirical and, regardless of the level of success, is without a rigorous theoretical underpinning.

In the present work, the validity of the Boltzmann device is examined and an alternative approach is proposed. Section II discusses the "ensemble" used to define protein potentials. It is seen that a databank of different protein structures cannot be described by any of the traditional statistical mechanical ensembles. Consequently, one cannot *a priori* assume any specific form for the distribution law. In Section III, the connection between MDL and statistical mechanics is discussed. Not only can statistical mechanics be recast as a Bayesian inference model based on MDL, but this formulation can also be used to extend the applicability of statistical mechanics to microstates and to knowledge-based systems. In Section IV, a derivation of knowledge-based protein potentials is presented that is based on this statistical mechanical model.

2 The Statistical Mechanics of Protein Potentials

To recognize the formal problems associated with using the Boltzmann device, it is important to focus on the nature of the "ensemble" from which the statistics are derived. Protein structures determined from NMR and from X-ray crystallography are canonical ensemble-averaged structures that follow a Boltzmann law. In knowledge-based methods, one is not observing members of this ensemble. Rather a collection of many different ensemble-averaged species is considered. Unlike the ensembles of statistical mechanics, one now has different species in each partition. These database structures are independent and do not interact with each other. Thus, they represent an isolated system similar to a microcanonical ensemble. Some authors have used the term, a protein "zoo" and this is entirely appropriate. Each partition has a different species that in a sense is fenced off (no interactions) with all other species. The zoo is not entirely like a microcanonical ensemble because the microstate in one partition could never be duplicated in the next one. Also, even though they are isolated, each partition will not be at constant energy. So the question is: Where lies the statistical mechanics of the protein zoo?

The protein zoo is clearly not a canonical ensemble. Its members are not at constant temperature and are not capable of exchanging energy. It is also not a microcanonical ensemble because the partitions are not at constant energy. One can, however, come close to approximating it as a microcanonical ensemble. If the number of proteins in each partition is appropriately adjusted, a constant number of amino acids in a partition can be achieved. To a first approximation the system's energy will be dominated by covalent interactions and these interactions will be fairly sequence independent. Consequently, each partition with a constant number of amino acids will have a fairly constant energy. One can view this as a microcanonical ensemble in which the spread in the energy distributions is dictated by structural differences rather than by fluctuations as found in the traditional ensemble.

Indeed, the lack of fluctuations in the ensemble has strong implications for the proper statistical mechanical treatment of the zoo. If, for the sake of argument, we accept the idea of a protein zoo being a microcanonical ensemble, then it is apparent that the Boltzmann law, 1, cannot hold. Rather one has the microcanonical law for the probability of a given configurational state [9]:

$$P = \frac{1}{q} \tag{6}$$

for $E + \delta E \geq H \geq E$ and

$$P = 0 \tag{7}$$

otherwise. The microcanonical partition function is defined as:

$$q(E) = \int_{\Omega} \delta(E - H(z))\, dz \tag{8}$$

where z is defined as the phase space coordinates of an M particle system, $z \equiv (q_1, q_2, \ldots q_M, p_1, p_2, \ldots p_M)$ and Ω is the region of phase space that occupies the energy from E to $E + \delta E$ and H is the Hamiltonian for the system. The microcanonical probability has the advantage that temperature does not appear explicitly in 6. However, it is very cumbersome and would be difficult to implement in practical calculations.

While the microcanonical partition function meets many of the requirements of the protein zoo, there are still fundamental problems. These problems are associated with the very basis of statistical mechanics. It is probabilistic in nature and assumes minimal knowledge of phase space. In our protein zoo, there is detailed knowledge of the coordinates. The basic problem with using traditional statistical mechanics for the protein zoo is that the standard averaging process presumes that the energy spacing between microstates is much smaller than the measured uncertainty in the thermodynamic energy. We cannot make this assumption for the protein zoo.

To see why standard statistical mechanics is inappropriate for a system where there is detailed knowledge, we follow a discussion due to Penrose [10]. It is important to distinguish between the tolerance in the experimental measurement of the energy of the system, designated ΔE, and the energy spacing between microscopic energy levels, δE. Implicit in the averaging process of statistical mechanics, one assumes that there are many neighboring levels within the tolerance, ΔE, so that $\Delta E/\delta E$ is a large number. To determine the observational state, one typically uses a two-limit process. As the experimental tolerance vanishes, δE will also vanish. However, the number of microscopic energy levels in the system are still very dense. For standard statistical mechanical systems, a double limit is performed as:

$$\frac{\delta E}{\Delta E} \approx \lim_{\Delta E \to 0} [\lim_{\delta E \to 0} \frac{\delta E}{\Delta E}] = \lim_{\Delta E \to 0} [0] = 0 \tag{9}$$

For knowledge-based ensembles, one can, in principle, have near perfect accuracy in determining the energy levels of the system. This microscopic accuracy can exceed the accuracy of a macroscopic measurement. This situation is akin to the "ordinary mechanics limit" discussed by Penrose [10]. In such a case, $\Delta E \ll \delta E$ and the limit is taken as:

$$\frac{\Delta E}{\delta E} \approx \lim_{\delta E \to 0} [\lim_{\Delta E \to 0} \frac{\Delta E}{\delta E}] = \lim_{\delta E \to 0} [0] = 0 \tag{10}$$

Despite the fact that both ΔE and δE become small, their ratio is quite different in statistical mechanics as opposed to ordinary mechanics. When detailed knowledge of the systems exists, as in the protein zoo or in computer simulations, one has an ordinary mechanics limit. In this situation, one has precise knowledge of the state at a given energy level, but cannot infer details of the neighboring levels from this information. Also, any formulation of a phase space density must be done in terms of delta functions, rather than continuous functions. Such a formulation has not been previously developed. The above considerations do not invalidate the use of statistical mechanics for known systems. It does mean that care must be taken when performing averages and when using the knowledge of these systems.

3 Knowledge-Based Statistical Mechanics and MDL

In previous work, a statistical mechanical formalism was developed that could be used in either the statistical (9) or the classical (10) limit [11]. This work was based on defining the information content or complexity of phase space. For a variety of technical reasons, it is convenient to use a complexity measure developed by Rissanen [12] called the stochastic complexity. This parameter utilizes the principle of minimal description length (MDL). The basic concept is that the best statistical estimation scheme is one in which both the data and the model's structure and parameters are represented in the shortest binary string. If the number of parameters in the model is fixed, the MDL estimation reduces to the familiar maximum likelihood scheme. At the other extreme, when the parameters determine the data, MDL estimation is identical with Jayne's maximum entropy method [13]. Thus, this approach represent a generalization that encompasses more standard methods. The power of this scheme is that it permits estimates of the entire model, the data, the parameters and even the number of parameters. This means that the estimated parameters need not be tested by external hypothesis to determine if the model is over or under parameterized. A second advantage is that the MDL need not refer to continuous probability functions. It can be used to describe objects characterized by discrete values of their attributes.

In Rissanen's approach, a given parametric model will have a known functional form that allows the data set to be "encoded" by a set of parameter values. These values, along with a list of "errors" representing the difference between the fitted and observed data, capture the information content of the model. The contribution from the errors is the maximum likelihood term. In the case of a model in which the error in the data is a random variable, the minimum length of the system is a description of this stochastic variable, hence

the terminology: stochastic complexity. For a set of data $x \equiv (x_1, x_2, \ldots x_N)$ that is fit by a model with a parameter vector $\mathbf{\Theta}$, the minimal description length is:

$$L(\mathbf{x}, \mathbf{\Theta}) = L(\mathbf{x} \mid \mathbf{\Theta}) + L(\mathbf{\Theta}) \tag{11}$$

where $L(\mathbf{x} \mid \mathbf{\Theta})$ is the length of the binary string required to encode the data and $L(\mathbf{\Theta})$ is the length of the string encoding the parameters. Because the model will have a functional form that allows a predicted value to be calculated succinctly, the data description can be represented as a description of the errors, ξ, between fitted and observed data. This gives:

$$L(\mathbf{x} \mid \mathbf{\Theta}) = L(\xi \mid \mathbf{\Theta}) = -\ln_2 (P(\xi \mid \mathbf{\Theta})) \tag{12}$$

For the minimal description length, $P(\xi \mid \mathbf{\Theta})$ will be the maximum likelihood function and $L(\xi \mid \mathbf{\Theta})$ will be proportional to the sum of squared errors.

The term $L(\mathbf{\Theta})$ will contain an encoding of the functional form and parameters of the model. Usually the functional form is so succinct that it is not considered to contribute significantly to the length. The fitted parameters can be encoded by a string of length $L(\mathbf{\Theta})$ according to:

$$L(\mathbf{\Theta}) = \sum_i \ln_2 \left(\frac{\Theta_i}{\delta_i} \right) \tag{13}$$

with δ_i being the precision of the ith parameter. Inclusion of the precision term allows real valued parameters to be represented as integers for the purpose of encoding

At first, this estimation method may not appear to be relevant to the statistical inference of traditional statistical mechanical methods. In statistical mechanics, one works with a family of models whose parameters are the phase space coordinates of an M particle system, $\mathbf{z} \equiv (\mathbf{q_1}, \mathbf{q_2}, \ldots \mathbf{q_M}, \mathbf{p_1}, \mathbf{p_2}, \ldots \mathbf{p_M})$ and with k observables represented as: $\mathbf{A}(\mathbf{z}) \equiv (\mathbf{A_1}(\mathbf{z}), \mathbf{A_2}(\mathbf{z}), \ldots \mathbf{A_k}(\mathbf{z}))$. The complexity of a many body system is given by the length of the minimal description required to encode both the parameters and the observables, $L(\mathbf{z}, \mathbf{A}(\mathbf{z}))$. In a standard ensemble representation of phase space, this length is given by:

$$L(\mathbf{z}, \mathbf{A}(\mathbf{z})) = -\log_2 P(\mathbf{A}(\mathbf{z}) | \mathbf{z}) + L(z) \tag{14}$$

where $P(\mathbf{A}(\mathbf{z}) | \mathbf{z})$ is the likelihood of the data $\mathbf{A}(\mathbf{z})$ for the parameters z and $L(\mathbf{z})$ is the minimal encoding of the parameters.

Invariably $M \gg k$ and the system is grossly under determined. Indeed statistical mechanics does not attempt to "fit" data in the statistical estimation sense. Rather it develops prior distributions that use the observables as

constraints. Constraining the variables to be compatible with the prior means that $P(\mathbf{A}(\mathbf{z})|\mathbf{z})$ is essentially constant. Thus, a maximum entropy encoding of statistical mechanics consists only of the $L(\mathbf{z})$ term along with the constraints. Rissanen showed that minimizing the length $L(\mathbf{z})$ is equivalent to maximizing an "entropy function" subject to constraints [12].

It has been explicitly demonstrated that an encoding of the Gibbs ensemble will give an $L(\mathbf{z})$ that is equivalent to the general expression for the thermodynamic entropy of the system [11]. This approach offers more than another way to derive statistical mechanics. It is seen that $L(\mathbf{z}, \mathbf{A}(\mathbf{z}))$ has the general properties of an entropy function. It is an extensive property of the size of the system and it is a concave function about an equilibrium point. This entropy function can than be used to describe microstates of ensembles in which one has detailed information of the state. Because of its generality, this approach is useful for discussing thermodynamic systems that obey both the classical limit (see 10) and the statistical limit (as in 9).

Using 14 one can obtain a generalized Bayesian form for the system. By analogy with continuous probability, the following functions can be defined:

$$P(\mathbf{z}, (\mathbf{z})) = 2^{-L(\mathbf{z}, \mathbf{A}(\mathbf{z}))} \tag{15}$$

$$P(\mathbf{z}) = 2^{-L(\mathbf{z})} \tag{16}$$

It is important to note that these functions are not necessary derived from a frequency distribution and can be used to represent encoded descriptions of single objects. In the Kolmogorov interpretation, these are the probabilities of randomly generating the binary string needed to describe the respective object. The generalized Bayes formula then takes the form:

$$P(\mathbf{z}, \mathbf{A}(\mathbf{z})) = P(\mathbf{A}(\mathbf{z})|\mathbf{z}) P(\mathbf{z}) \tag{17}$$

$$= P(\mathbf{z}|\mathbf{A}(\mathbf{z})) P(\mathbf{A}(\mathbf{z})) \tag{18}$$

The advantage of 17 is that it provides a Bayesian format for developing the analog of maximum entropy "image reconstruction" techniques.

It is interesting to view 17 and 18 in reference to the two descriptions, statistical and classical in 9 and 10, respectively. In both cases, $P(\mathbf{z}, \mathbf{A}(\mathbf{z}))$ provides an adequate description of the system. However, the weighting of the different probability functions in 17 is different between the statistical and the classical description. In probabilistic ensemble descriptions the thermodynamic parameters narrowly define a most probable configuration. In this case, $P(\mathbf{z})$ has a very narrow distribution and $P(\mathbf{A}(\mathbf{z})|\mathbf{z})$ is quite broad. The breadth of $P(\mathbf{A}(\mathbf{z})|\mathbf{z})$ is a result of the breadth in the tolerance in the macroscopic energy of the system, ΔE. It is because of this that $P(\mathbf{z})$ provides a good

approximation for $P(\mathbf{z}, \mathbf{A}(\mathbf{z}))$ and maximum entropy techniques can be used. From 18 it is seen that ensemble statistical mechanics could also be developed along the lines of the maximum likelihood method. In this case, $P(\mathbf{z}|\mathbf{A}(\mathbf{z}))$ is used to approximate $P(\mathbf{z}, \mathbf{A}(\mathbf{z}))$ and in the statistical limit $P(\mathbf{z}|(\mathbf{z}))$ will be distributed very narrowly while $P(\mathbf{A}(\mathbf{z}))$ will be quite broad. Again, the breadth of $P(\mathbf{A}(\mathbf{z}))$ is a reflection of ΔE.

In the classical limit, 10, the probability distribution function of the system's observables $P(\mathbf{A}(\mathbf{z})|\mathbf{z})$ is extremely narrow, possibly even narrower than the experimental tolerance as a result of our knowledge of the system. The distribution of the probability, $P(\mathbf{z})$ will, for most thermodynamic systems, be very broad. This is because there are many phase space configurations that could correspond to a set of thermodynamic observables. When the number of observables is much greater than the parameters, this situation is best treated with maximum likelihood methods. In such cases, $P(\mathbf{A}(\mathbf{z})|\mathbf{z})$ is determined and because of its breadth, $P(\mathbf{z})$ can be ignored.

When there is partial knowledge of the system, one does not have a single dominant probability function contributing to 17 or 18. It is this intermediate case that is most difficult for maximum likelihood and maximum entropy techniques. This situation is best treated using Rissanen's MDL estimation scheme. The case of the protein zoo will most closely resemble this type of system. This is because the content of the database does not greatly exceed the content of the inferred potentials. Thus, the MDL scheme provides the mathematical tools for determining the statistical mechanics (or more appropriately inferring thermodynamic relations) for the protein zoo. It should be emphasized that our knowledge of the system in no way changes the thermodynamics of it. All that is changed is how inferences are made. Regardless of the "distribution" of knowledge about the system, it will always be represented by some function, $P(\mathbf{z}, \mathbf{A}(\mathbf{z}))$. What changes with knowledge is the best way of calculating this function.

4 Reconstruction of Potential Functions from Structural Data

Jayne's showed that statistical mechanics can be viewed as a form of statistical inference rather than a physical theory [13]. This formulation of statistical mechanics became the origin of the maximum entropy techniques used in image reconstruction [14]. In this view of statistical mechanics, the "image" that one infers is that of the most probable phase space coordinates. The "imperfect data" is our limited knowledge of the system. Usually, this will consist of the macroscopic thermodynamic variables (such as E, V and T for a microcanonical ensemble) as well as the microscopic energy levels (ε_i) derived from some

quantum model or from spectroscopic measurements. This mapping from observable space to phase space is tricky because the number of observables, N, is much less than the number of variables in the image space (the M phase space coordinates). The maximum entropy technique achieves this mapping in the most unbiased way possible. Consequently, it is the method of choice in such problems.

The maximum likelihood method, commonly used for fitting parameters to data, requires that "observable space" be larger than "parameter space", i.e., $N >> M$. While this method is inappropriate for inference in statistical mechanics, it may be appropriate for the analysis of computer simulations. If one has detailed knowledge of phase space coordinates and wants to infer thermodynamic parameters, one could appropriately use a maximum likelihood formalism.

In this work, our goal is to use structural data to infer statistical properties of intermolecular protein potentials. The observable space is the spatial coordinates of all the amino acids in all the proteins of the protein zoo. The image space is the potential function of interest for each amino acid pairing. Assuming that chain directionality is not important, there are 200 pairings that must be discretized along some relevant coordinate (angular and/or distance). Both data space, the protein zoo, and image space, the statistical potentials, will not be overwhelmingly different in size. Consequently, one cannot validly use either maximum entropy or maximum likelihood methods.

As seen in the previous sections, the MDL yields $L(\mathbf{z}, \mathbf{A}(\mathbf{z}))$ and $P(\mathbf{z}, \mathbf{A}(\mathbf{z}))$. These, in turn, are related to the entropy and the partition function, respectively [11]. To use the MDL estimation scheme in image reconstruction, one of the Bayesian representations (17 or 18) will be used:

$$P(\mathbf{z}, \mathbf{A}(\mathbf{z})) = P(\mathbf{A}(\mathbf{z}))\, P(\mathbf{z}|\mathbf{A}(\mathbf{z})) \tag{19}$$

or

$$P(\mathbf{z}, \mathbf{A}(\mathbf{z})) = P(\mathbf{z})\, P(\mathbf{A}(\mathbf{z})|\mathbf{z}) \tag{20}$$

The choice of representation will depend upon the specific problem of interest. Quite often the functions, $P(\mathbf{z})$ or $P(\mathbf{A}(\mathbf{z}))$, will not be of particular interest. For instance, $P(\mathbf{z})$ may represent some intrinsic quantum limit on the uncertainty in a variable. Often, $P(\mathbf{A}(\mathbf{z}))$ will be a experimental, measurement-limited distribution function. In such cases, these probabilities are considered fixed. The Bayesian representation containing them is chosen so that one focuses only on calculating the respective conditional probability.

For the protein problem, a structural database is used to generate a set of phase space coordinates (denoted $\mathbf{z_{zoo}}$), that allows the calculation of the

respective protein potentials $\mathbf{A}(\mathbf{z_{zoo}})$. While $L(\mathbf{z_{zoo}}, \mathbf{A}(\mathbf{z_{zoo}}))$ could be calculated and used as an estimate of $L(\mathbf{z}, \mathbf{A}(\mathbf{z}))$ and $P(\mathbf{z}, \mathbf{A}(\mathbf{z}))$, this approximation would lead to considerable variation from one databank to the next. There are better, alternate ways of determining the function $P(\mathbf{z}, \mathbf{A}(\mathbf{z}))$. In the present case, $P(\mathbf{z})$ is the natural distribution in protein coordinates, i.e., the spread in angular and spatial coordinates. While this is an interesting parameter, we are more concerned with the potentials themselves. Thus, this term is taken as fixed and the main computation task is determining $P(\mathbf{z}|\mathbf{A}(\mathbf{z}))$. This can be done using the Bayesian properties and a MDL maximization.

Returning to the "probability functions", one has:

$$P(\mathbf{A}(\mathbf{z}), \mathbf{z_{zoo}}|\mathbf{z}) = P(\mathbf{A}(\mathbf{z})|\mathbf{z}) P(\mathbf{z_{zoo}}|\mathbf{A}(\mathbf{z}), \mathbf{z}) \tag{21}$$

Using:

$$P(\mathbf{A}(\mathbf{z})|\mathbf{z}) = \int P(\mathbf{A}(\mathbf{z}), \mathbf{z_{zoo}}|\mathbf{A}(\mathbf{z}))\, d\mathbf{z_{zoo}} \tag{22}$$

one has the common image reconstruction equation [14]:

$$P_{new}(\mathbf{A}(\mathbf{z})|\mathbf{z}) = \int P_{old}(\mathbf{A}(\mathbf{z})|\mathbf{z}) P(\mathbf{z_{zoo}}|\mathbf{A}(\mathbf{z}), \mathbf{z})\, d\mathbf{z_{zoo}} \tag{23}$$

where $P_{old}(\mathbf{A}(\mathbf{z})|\mathbf{z})$ is the prior. The quantity $P(z_{zoo}|A(z), z)$ is related to how well the prior predicts the phase space data of the protein zoo. This quantity is often assumed to take the form:

$$P(\mathbf{z_{zoo}}|\mathbf{A}(\mathbf{z}), \mathbf{z}) = \exp\left\{-(\mathbf{z} - \mathbf{z_{zoo}})^2 / \sigma^2\right\} \tag{24}$$

where σ^2 is some prescribed variance. Such a function gives a simple Gaussian distribution for the errors between the data space and the image space and is commonly used in image reconstruction. A direct physical justification for the form of 24 can be derived from a path integral formulation of polymer statistics. For the new potential function to be the best fit to the data, one maximizes the integrand in 23 with respect to $\mathbf{z}$. This gives the best potential function consistent with the prior. Thus, 23 provides the basic algorithm for calculating protein potentials.

5 Summary

In this work, the basic assumptions behind the use of the Boltzmann device to derive protein potentials have been examined. It was seen that there is

no justification for this device based on traditional statistical mechanical ensembles. To create a statistical mechanics that handles knowledge, a new formalism is employed that is based on the MDL principle. With this approach, knowledge-based potentials can be derived by methods that formally resemble image reconstruction techniques. Future work will focus on practical applications of this approach.

1. P.D. Thomas and K.A. Dill, *J. Mol. Biol.* **257**, 457 (1996).
2. A.V. Finkelstein, A. M. Gutun and A.Y. Badretdinov, *FEBS Lett.* **325**, 23 (1993).
3. M.J. Sippl, S. Weitckus and H. Flöckner,in "The Protein Folding Problem and Tertiary Structure Predicition" ed. K.M. Merz, Jr. and S. M. LeGrand pp. 353 (Birhäuser, Boston 1994).
4. M. J. Sippl, *J. Mol. Biol.* **213**, 859 (1990).
5. A. Godzik, A. Kolinski and J. Skolnick, *Protein Sci.* **4**, 2107 (1995).
6. J.-P. A. Kocher, M. J. Rooman and S. J. Wodak, *J. Mol. Biol.* **235**, 1598 (1994).
7. P.D. Thomas and K.A. Dill, *Proc. Natl. Acad. Sci. USA* **93**, 11628 (1996).
8. L.A. Mirny and E.I. Shakhnovich, *J. Mol. Biol.* **264**, 1164 (1996).
9. L.M. Grossman, Thermodynamics and Statistical Mechanics, pp. 115 (McGraw Hill, New York, 1969).
10. O. Penrose, Foundations of Statistical Mechanics, (Pergamon Press, Oxford, 1970).
11. T. G. Dewey, (submitted, 1998).
12. J. Rissanen, *Automatica* **14**, 465 (1978); J. Rissanen, *Ann. Stat.* **14**, 1080 (1986); J. Rissanen, *J. R. Statist. Soc. B* (1987) **49**, 223 (1987).
13. E. T. Jaynes, *Phys. Rev.* **106**, 620 (1957); E. T. Jaynes, *Phys. Rev.* **108**, 171 (1957).
14. B. Buck and V. A. Macaulay, Maximum Entropy in Action, (Clarendon Press, Oxford, 1991).

Acknowledgement

This work was supported in part by NIH grant 1R15GM55610,