Neural Net Representations of Empirical Protein Potentials

Tal Grossman
Theoretical Division and CNLS
MS B213
Los Alamos National Laboratory
Los Alamos, NM 87544, USA
email: tal@t13.lanl.gov

Robert Farber and Alan Lapedes
Theoretical Division
Los Alamos National Laboratory
and The Santa Fe Institute
1399 Hyde Park Rd.
Santa Fe, NM 87501, USA

Abstract

Recently, there has been considerable interest in deriving and applying knowledge-based, empirical potential functions for proteins. These empirical potentials have been derived from the statistics of interacting, spatially neighboring residues, as may be obtained from databases of known protein crystal structures.

In this paper we employ neural networks to redefine empirical potential functions from the point of view of discrimination functions. This approach generalizes previous work, in which simple frequency counting statistics are used on a database of known protein structures. This generalization allows us to avoid restriction to strictly pairwise interactions. Instead of frequency counting to fix adjustable parameters, one now optimizes an objective function involving a neural network parameterized probability distribution.

We show how our method reduces to previous work in special situations, but also allows extensions to include orders of interaction beyond pairwise interaction. Given the close packing of proteins, steric interactions etc., the inclusion of higher order interactions is critical for developing an accurate potential. A key feature in the approach we advocate is the development of a representation to describe the spatial location of interacting residues that exist in a sphere of small fixed radius around each residue. This is a "shape representation" problem that has a natural solution for the interaction neighborhoods of protein residues. We demonstrate in a series of numerical experiments that the neural network approach improves discrimination over that obtained by previous methodologies limited to pair-wise interactions.

Introduction

Recently there has been considerable interest in knowledge-based, empirical potential functions for proteins. The idea of using a database of known protein structure/sequence pairs to derive an empirical potential describing the interaction between residues has a relatively long history. Some early work considered the frequency with which pairs of amino acids appeared within a certain "contact" distance of each other (Miyazawa & Jernigan 1985), and used a quasi-chemical approximation to relate this frequency to an approximate free-energy of interaction of a "gas" of residue pairs. New work also relies on an approximate statistical mechanical interpretation (Sippl 1990). These methods typically use simple probability approximation by frequency counting to derive approximate free energy formulae. Both older and more recent approaches fall into two main groups. The first group considers the observed frequency with which the distance between pairs of amino acids appear within one or more distance bins, in known crystal structures. This approach, by construction, is limited to considering pair interactions only. Except in a few exceptional cases there is insufficient data to consider triple and other higher order interactions (Godzik, Kolinski, & Skolnick 1992). The second group constructs a definition (by hand) of an "environment" for an amino acid based on polarity, secondary structure, water exposure etc (Bowie, Luthy, & Eisenberg 1991). These defining characteristics of an environment are coarsely binned so that one can approximate by frequency counting the conditional probability of a single amino acid appearing in an environment.

We describe below how the different forms of empirical potential functions (variously called "contact potentials" (Sippl 1990; Godzik, Kolinski, & Skolnick 1992), or in other versions where environments are predefined, "profile potentials" (Bowie, Luthy, & Eisenberg 1991)), may be derived as the solution to a discrimination problem. This opens the door to constructing discrimination functions using powerful machine learning techniques such as neural networks, as opposed to simple frequency counting methods. The discrimination task involves a log-likelihood function of certain probabilities that are also implicit in the prior "contact potential" and "profile" methodologies. Previous work of Bryant and Lawrence (Bryant & Lawrence 1993) also rely on a log-likelihood statistical interpretation of contact potentials, but restricts consideration to log-linear models truncated at the second order. One other approach (Goldstein, Luthey-Schulten, & Wolynes 1992) also optimise parameters in an empirical protein potential, and is therefore related to the approach we advocate. However, they use

a polynomial interaction term that stops at second order pair interactions, whereas our goal in using neural networks is to remove both the use of frequency counting and the assumption of second order interactions. Steric conflicts, volume constraints and other nonlocal interactions caused by close packing in a protein interior suggest that higher order interactions are the norm among residues in a protein, as well as the possible importance of the spatial configuration (i.e. the relative positions) of interacting residues. Neural networks efficiently include higher order interactions without the explosion of parameters plaguing polynomial representations. A neural network can employ hidden neurons to detect correlations higher than second order among interacting residues, and also does not rely on frequency counting to approximate probability distributions. We also describe how to use structural information beyond pair contacts.

We show how our new formulation reduces to previous work in special situations, and also extends the power of previous approaches. Finally, we demonstrate that the network approach has improved discrimination power over previous methodologies by using spatial information.

Empirical Potentials as Log-Likelihood Ratios

Various groups have employed differing approaches to define empirical potentials for proteins. These potentials are often considered to be approximations to free energies. We find it useful to reconsider these potentials in statistical, instead of statistical mechanical, terms to make the extension to using neural network techniques more natural. This view of previous work also serves to point out the relationships between previous approaches. The statistical approach involving log-likelihood potentials to construct potential functions was first used in this context by Bryant and Lawrence (Bryant & Lawrence 1993), but was of necessity restricted to pair-wise interactions. Our approach differs in our use of Bayes theorem to relate differing methods and our use of neural networks to capture higher order effects.

We briefly summarize the essential features of the "profile" approach (Bowie, Luthy, & Eisenberg 1991), as well as the "contact potential" approach (Bryant & Lawrence 1993; Sippl 1990; Godzik, Kolinski, & Skolnick 1992), in order to point out the different log-likelihood ratios that each approach implicitly uses. In Sippl (Sippl 1990), the probability for amino acid pairs, ab, to be separated by a (binned) distance, r, is approximated by frequency counting in a database of known structures. This probability is represented as the conditional probability, $P(r|ab)$, of finding the distance bin r given each of the order 20×20 amino acid pairs. Arguments are given by Sippl (Sippl 1990) relating

$$\log\left(\frac{P(r|ab)}{P(r)}\right) = \log\left(\frac{P(r,ab)}{P(r)P(ab)}\right) \qquad (1)$$

to approximate statistical mechanical free-energies. We note that this expression may alternatively be viewed as a statistical log-likelihood ratio. This latter interpretation quantifies the relation between $P(r)$ and $P(ab)$ by comparing the joint probability, $P(r,ab)$ to that obtained under the assumption of independence of the distance bin, r, and the pair ab.

After obtaining these approximate free energy or log-likelihood quantities from a "training-set" of proteins, one may evaluate the compatibility of any specific amino acid sequence with a given structure of interest. The known structure enables the distances (and hence distance bins) between all pairs of amino acids in that structure to be calculated. One may consider a sphere surrounding each residue of the protein that contains the interacting spatial neighbors for each residue. An approximate free-energy, or log-likelihood ratio, for each sphere can be calculated by summing the appropriate pair-wise free-energy contributions of residues in the sphere interacting with the central residue. The energy, or log-likelihood ratio, for the complete protein is approximated by adding together the energies of all the individual spheres.

Hence this process may be given a statistical interpretation in which one approximates the log-likelihood ratio for each sphere and assumes independence of the spheres, i.e. additivity of the sphere log-likelihood ratios, to derive a log-likelihood ratio for the complete protein. A key feature of the neural net approach introduced below is that while one still assumes that the spheres contribute additively to the overall log-likelihood or free-energy, the interactions of the residues within each sphere are *not* assumed to be the additive contribution of independent pair interactions, but rather includes high order interactions within the interaction radius for each residue. The statistical interpretation advocated here also serves to relate seemingly different approaches published in the literature. For example, Sippl (Sippl 1990) expresses the log-likelihood ratio as $\log\left(\frac{P(r|ab)}{P(ab)}\right)$, while Wilmanns and Eisenberg (Wilmanns & Eisenberg 1993) use $\log\left(\frac{P(ab|r)}{P(r)}\right)$. These, of course, are identical expressions when related by Bayes theorem

$$\frac{P(r|ab)}{P(r)} = \frac{P(ab|r)}{P(ab)} \ .$$

Still different log-likelihood expressions are used by Bryant and Lawrence (Bryant & Lawrence 1993) and implicitly by Skolnick et.al. (Godzik, Kolinski, & Skolnick 1992). In this work, the probabilities of the various pairs of amino acids to have inter-residue distances within certain distance bins are computed in a similar frequency counting fashion to the above. However,

the log-likelihood ratio that is used differs from the above. The probabilities of the various pairs of amino acids to have inter-residue distances within certain distance bins are compared to the probability assuming the sequence was randomly permuted and re-threaded through the protein. This approach leads to likelihood ratios, $\frac{P(ab|r)}{P(a)P(b)}$, different than the above.

The relationship of the Bryant and Lawrence, and Skolnick et.al. approaches to that of Wilmanns and Eisenberg (Wilmanns & Eisenberg 1993), is clarified if one considers the central amino acid in each sphere to exist in an "environment" comprised of its spatially neighboring residues. Then the approach of (Wilmanns & Eisenberg 1993) may be viewed as discriminating between spheres that have the natural amino acid at their center, versus environment spheres that have an arbitrary and possibly incompatible amino acid at their center. The alternative log-likelihood ratio implicit in (Godzik, Kolinski, & Skolnick 1992; Bryant & Lawrence 1993), on the other hand, involves a discrimination task in which not only the central residue of each sphere is randomly permuted, but also permuted are the rest of the residues in the sphere that comprise the environment of the central residue.

Finally we note that "3D-1D profile methods" (Bowie, Luthy, & Eisenberg 1991) are related to the approaches of (Wilmanns & Eisenberg 1993; Sippl 1990) where the environment remains unchanged and the central amino acids in each sphere are permuted. These profile approaches don't rely on the statistics of residue pairs in inter-residue distance bins. Instead, a fixed set of environment classes are predefined (depending, e.g., on polarity, hydrophobicity and secondary structure characteristics). The conditional probability of each amino acid type to associate with a particular environment is then computed from a "training set" of proteins. Log-Likelihood ratios of a central amino acid in relation to its environment may be computed and used as a "score" to represent the compatibility of an amino acid with its environment.

The choice of which type of log-likelihood ratio (Bryant & Lawrence 1993; Wilmanns & Eisenberg 1993; Sippl 1990; Godzik, Kolinski, & Skolnick 1992), to use depends on the nature of the problem being considered. In this paper we consider the type of "scrambled" log-likelihood ratio used in (Godzik, Kolinski, & Skolnick 1992; Bryant & Lawrence 1993). Scrambling a sequence is a good approximation to both (1) threading that sequence through an arbitrarily selected non-native and inappropriate structure, in which case residues are arbitrarily put into contact and (2) threading an arbitrary and inappropriate sequence (but having the same amino acid distribution) through a native structure, in which case arbitrary residues are put into contact.

Neural Networks as Log-Likelihood Ratios

The view advanced above, that constructing empirical protein potentials is usefully viewed as a problem in statistical approximation of log-likelihood ratios, as opposed to a statistical mechanical problem of approximating free-energies, prompts the use of machine learning techniques. In this section we show how neural networks approximate log-likelihood ratios.

The basic point is simple. Leaving aside details of representation for the following section, consider the pair consisting of an Input vector, I, and an Output vector, O: (I,O). Normally, a neural network would be trained, using e.g. backpropagation in a feed-forward architecture, to predict Output from Input. If a suitable objective function is used for training (Hopfield 1987; Baum & Wilczek 1987), then the Output of the network approximates the conditional probability, $P(Output|Input)$, which can be used in log-likelihood ratios.

Various log-likelihood ratios may be obtained, which we label (A), (B) and (C) below, depending on what is being represented in the Input and Output vectors:

(A) If the Output represents the central amino acid in a sphere of fixed radius, and the Input represents the surrounding neighbors in the sphere, then a log-likelihood ratio, LLR, can be constructed as

$$LLR = \log\left(\frac{P(Output|Input)}{P(Output)}\right) \qquad (2)$$

which quantifies the relation of the central amino acid to its environment comprised of its spatial neighbors.

(B) One can also consider a particular two class (True, False) discrimination problem that yields the same log-likelihood ratio as (A) above. First concatenate the Input and Output pairs (i.e. environment and central residue) into a new Input vector, $NewInput$. Assign a new Output class label "True" to all such concatenated residue-environment vectors that may be obtained from a representative set of known crystal structures. To construct a new False set, randomly permute the central residue among the environment vectors of the True set, and assign the label "False" to all such vectors. These vectors have the relation of the central amino acid to its environment broken by the permutation, so that the probability of $NewInput$ given False class will factorize.

Consider the situation of equal numbers of True and False examples in the training set, and where the single output is a standard sigmoid function,

$$g(X) = \frac{1}{1+\exp(-X)} \quad , \qquad (3)$$

representing $P(TrueClass|NewInput)$.

Algebraically expressing X in terms of g yields

$$\begin{aligned} X &= \log\left(\frac{g(X)}{1-g(X)}\right) \\ &= \log\left(\frac{P(TrueClass|NewInput)}{P(FalseClass|NewInput)}\right) \\ &= \log\left(\frac{P(NewInput|TrueClass)}{P(NewInput|FalseClass)}\right) \end{aligned} \quad (4)$$

The False class was explicitly constructed by random permutation such that the expression $P(NewInput|FalseClass)$ factorizes into the product of $P(CentralResidue) \times P(Environment)$, while the expression $P(NewInput|TrueClass)$ is the joint probability $P(CentralResidue, Environment)$, of Central-Residue and Environment as seen in un-permuted data.

Hence, X is a neural expression of the same form of log-likelihood ratio,

$$LLR = \log\left(\frac{P(CentralResidue|Environment)}{P(CentralResidue)}\right) \quad (5)$$

used in (Wilmanns & Eisenberg 1993; Sippl 1990).

(C) Alternatively, if one constructs a "False" set by randomly permuting each protein sequence and rethreading it through its structure, then the form of the log-likelihood ratio of (Godzik, Kolinski, & Skolnick 1992; Bryant & Lawrence 1993) is recovered. This is the form of log-likelihood ratio we consider in this paper. As noted above, this provides a good approximation for problems in which sequences are threaded through a given structure, or the reverse, in which a given sequence is threaded through a variety of unrelated structures. Creating efficient threading algorithms that can handle gaps for the more complicated potentials we develop remains an interesting and open problem (Lathrop). Assuming that gapped threadings are of interest, then performing an exhaustive enumeration of gapped threadings, following the procedures of (Bryant & Lawrence 1993), is a possible solution until more efficient algorithmic methods are devised.

An essential advantage of the neural network approach is that a network is able to represent these forms of log-likelihood ratios without using the pairwise interaction assumption and frequency counting. Use of hidden neurons allows incorporation of interactions higher than pairwise, second order interactions. The ability of neural networks to "generalize" to input data not present in the training set replaces frequency counting. However, one needs a suitable representation of the data with which to train the network. We discuss data representation in the next section.

A Representation for the Spatial Neighborhood of an Amino Acid

The database of crystal structures contains full atomic information on the location of each atom of residues and the backbone chain to which each residue is connected. A key issue is how to *represent* this information. A simple list of, e.g., XYZ co-ordinates of each object of interest relative to the coordinates of the central residue of the sphere is insufficient without a method to invariantly order this list. This is the classic issue of representing shape.

We solve this problem by making use of a special and natural co-ordinate system implicit in protein backbones. Each residue is attached to the backbone of a protein at the C_α position. There are two "special directions" defined by the relative positions along the backbone of the two neighboring atoms to the C_α atom. If we choose each C_α atom to be the center of a local neighborhood, it is natural to define the reference frame of this neighborhood by the the relative directions of the two atoms, N and C, connected to the C_α origin. The angle between the $N - C_\alpha$ and $C_\alpha - C$ vectors is essentially constant by virtue of the nature of the chemical bond (Branden & Tooze 1991). Also constant to a high degree of accuracy is the distance between N, C, and C_α atoms. These atoms therefore define a special co-ordinate system, centered at each residue, spanned by the $N - C_\alpha$ and $C_\alpha - C$ vectors, and their cross product (which defines the perpendicular to the plane) as shown in Fig.1. We take the z axis to be in the $C_\alpha - N$ direction, the x axis is the z - orthogonal component of the $C_\alpha - C$ direction, and y is their vector product.

In this work we consider the C_α, or the C_β atoms of other residues as the objects of interest, and the "neighborhood sphere" is the structure defined by the positions of those that are within a distance d (the radius of the sphere) from the center, with respect to the local reference frame. If C_β atoms are used to note residue locations, instead of C_α locations, then glycine, which has no C_β, is represented by its C_α atom.

To solve the problem of how to usefully order the list of co-ordinates of the spatial neighbors we divide the sphere surrounding each residue into a small number of finite bins. The bins are invariantly defined according to the co-ordinate system described above. Various binnings are possible, and a number of suitable binnings are described in detail in the following section. If a bin is occupied by a spatially neighboring residue we increment an integer counter in that bin. A more elaborate partition may be needed for other applications, and the exact number of bins, and their boundaries, are parameters that can be optimized according to the problem considered. In contrast to other approaches, it is not necessary to ignore the sequence neighbors of a residue. We may include these chain-neighbors because they contain information on local secondary structure which is ultimately weighted by the neural network in an automatic fashion.

An integer valued vector therefore serves to invariantly represent the geometrical location of spatial neighbors within each sphere. Each sphere is allocated

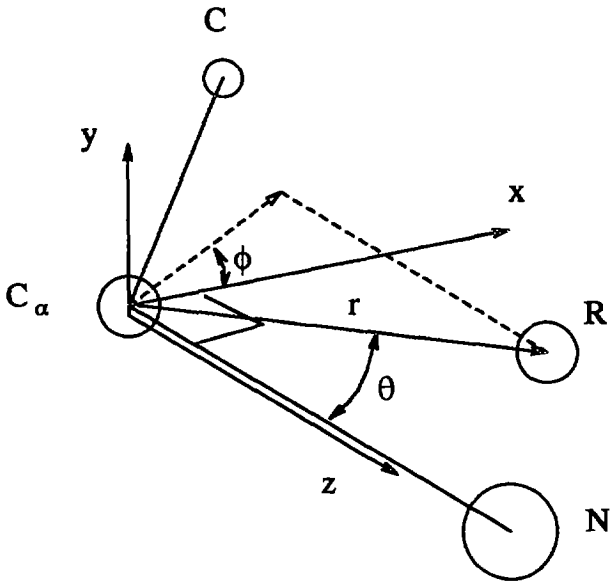

Figure 1: The definition of the reference frame around the backbone C_α. The polar coordinates of an object R (e.g. the C_α atom of another residue) within the sphere are also shown. θ is the angle between the vector R and the z axis (the $C_\alpha - N$ direction). ϕ is the angle between the projection of R on the x-y plane and the x axis.

a fixed number of bins regardless of how many neighbors are in the sphere (a variant of unary representation). Naturally, one can represent M bins in fewer than M integers, but this extended unary-style representation is useful for inputing data to a neural network, as described in the following section. One can augment this representation to also indicate the amino acid(s) residing in each bin. This is accomplished by using 20 integers per bin as a unary representation of the amino acid(s) in each bin. Thus, an empty bin is represented by 20 zero entries, a bin occupied by one residue with amino acid index j will have one as its j'th entry and zero for the other 19 components, etc. This representation has the advantage of keeping a fixed-size description of the neighborhood (even with different number of objects in a sphere), having a standard ordering of the objects (invariant spatial description), and uses a unary form suitable for neural networks.

Bin representations of Neighborhood Spheres

Construction of algorithms for determining the "optimal" bin representation containing relevant information is an interesting problem for which we will report results elsewhere. In this paper, we report results obtained with several different examples of binnings representing the identity and spatial position of neighboring residues within an interaction radius surrounding each central residue. There are several parameters which define the representation:

1. *Type of object:*. The "objects of interest" inside each sphere are either the backbone C_α of each amino acid, or its C_β. In the case of the C_β representation, only the position of the C_β atoms is recorded, and the reference frame produced by the backbone $N - C_\alpha - C'$ is translated to the C_β of the center residue (i.e. the one around which the neighborhood sphere is evaluated), which is the center of the sphere. Glycines have their "C_β" position set at the C_α position.

2. *Radius:* In this work we use a six Angstrom radius for each sphere.

3. *Sequence neighbors:* Sequence neighbors may be included as "special objects" in the bin representation, or not. Like the center residue, the nearest sequence neighbors are "special" objects in the neighborhood sphere, since they are connected by peptide bonds to the center amino acid. Therefore when included (the other option is not to include them at all), they are each represented by a 20 integers representation of the residue identity, plus 4 bits each to represent their position. The 4 bit representation for position was determined by a simple, manual procedure in which we examined the the distribution of positions in a subset of the training set of proteins to determine representative spatial bins. This binning, which is somewhat different for the C_α and C_β representations, is given by the following partitions. Each partition into 4 bins is accomplished by testing two conditions (except for the N neighbor in the C_α representation where we use only 2 bins) on the x,y,z or θ, ϕ coordinates of the neighbors' C_α (or C_β) positions:

   C' neighbor, C_β representation: $z \leq y$, $z \leq -y - 2.0$.

   N neighbor, C_β representation: $x \leq 0.5$, $y \leq -2.0$.

   C' neighbor, C_α representation: $\theta \leq 1.9$, $\phi \leq 1.43\theta - 3.0$.

   N neighbor, C_α representation (only 2 bins, 1 condition): $\phi \leq 0$.

4. *Binning resolution:* The partition of the sphere into bins should provide relevant information about the local structure, but, on the other hand, the resolution is limited by the available data. In principle, one can try to optimize the binning by adaptive clustering techniques. We plan to explore this direction in the future. In the experiments reported here we tested a few basic options.

   (a) Radial shell partitions: the distance of the object from the center of the sphere is partitioned into several shells. Here we use either a two shell partition in which the inner shell is 0 - 5 Å, and the

outer shell is 5 -6 Å; or a single shell partition (i.e. no binning of the distance).

(b) Octants partition: when this option is used, the angular position of an object is provided by the octant in which it is located. The octants are determined by the local reference frame. For example, when both the shell and octants partitions are used, we have 16 bins in our bin representation, and when neither are used we have only one.

As noted earlier, each spatial bin has attached a "composition vector" of 20 integers, representing the number of residues of a given type occurring in the bin. Occurances of two residues of the same type occupying the same bin is possible, and hence values greater than one may occur in the composition vector of a bin (with the finer bin representations, this is a rare event). Thus the input representation to the network is not strictly a unary representation.

Data preparation

The October 1994 EMBL list of proteins with less than 25% sequence homology (Hobohm & Sander 1994) was used to obtain a reasonably structurally diverse set of proteins. The chain selection suggestion for each protein in the EMBL list was used. This choice suffers from at least two possible defects, shared by most other investigations: (1) the list may not contain as structurally a diverse set of proteins as desired (2) Non-monomeric proteins, and complexes etc. may not have their full co-ordinates properly represented.

This set of 365 protein structures/sequences contained five proteins that are no longer found in the current (October 1994 release) Brookhaven PDB (Bernstein & et al. 1977), contained 17 proteins with only the C_α co-ordinates (no C_β, N, C'), and contained 30 proteins with more than one structural model. These proteins were not retained. The remaining list of 312 proteins was partitioned at random into 3 sets: a training set, containing 188 proteins, a cross-validation set with 62 proteins, and a prediction set with 62 proteins.

For every protein in the data, the neighborhood sphere around each residue was calculated, except for those for which we determined that the local structure was not reliable. These "unreliable" cases include residues which don't have one or more of the backbone atoms, or for which the coordinates of the backbone triplet $N - C_\alpha - C'$ do not have canonical values. We eliminated neighborhood spheres around residues which did not have two valid sequence neighbors and consequently none of the terminal residues is used in the data. If any residue in the neighborhood sphere of a central residue was an "invalid" residues (for any of the reasons mentioned above) then the whole neighborhood sphere was eliminated from consideration.

This procedure resulted in 81,402 spheres (using the C_α representation) in the training set, 26,290 spheres in the cross-validation set and 26,998 in the prediction set. For the C_β representation, this procedure resulted in 81,570 spheres in the training set, 26,300 spheres in the cross-validation set, and 27,048 in the prediction set. The numbers of examples are a bit larger when using the C_β representation, since the C_β-C_β distances are usually larger than the C_α-C_α distances (for the same pair of amino acids) due to the extra freedom of the residues around the backbone. As a result, less spheres are rejected due to "invalid" neighbors in the C_β representation (we use the same sphere radius).

The training and testing methodology used in all the experiments reported here is as follows: for each input representation and network architecture a network was trained on the training set using back-propagation and the relative-entropy error function (Hopfield 1987; Baum & Wilczek 1987). During the training process, "snapshots" of the network weights were examined at 10 iteration intervals of a conjugate gradient minimization algorithm. These weights were used to perform predictions on the cross-validation set. The network performing best on the validation set was selected as the final predictor. The prediction performance of the chosen network on the (totally disjoint) prediction set was then evaluated. A network architecture with twenty hidden units was evaluated for each representation below. In order to demonstrate the effect of higher order correlations, we report for the last two representations (representation 6 and 7) the results obtained for a net with no hidden units (a simple perceptron).

Results

Six different representations were tested. For each we report the prediction rate per sphere obtained by the evaluation procedures described above. In all these experiments, the final prediction rates for True and False examples were almost identical, with up to 1 or 2% difference.

Representation 1: The simplest representation used included only the unary 20 bit representation for the central residue identity, as well as the 20-integer composition vector for the non-neighbor residues in the sphere, based on C_β positions. Therefore there was one radial shell – which was the interior of the interaction sphere out to six Angstroms. The sequence neighbors were not included. The input size is therefore forty. The prediction accuracy was 60%.

Representation 2: This representation uses the 2 radial shells to identify radial positions of the non-neighbor residues, as described in section 4. Other parameters are the same as Representation (1). The input size is sixty. The prediction accuracy was 61%.

Representation 3: This representation uses only 1 shell, similar to Representation (1), but adds the residue identity for the two sequence neighbors (i.e. two additional 20 bit vectors are used in the input representation). The input size is eighty. The prediction accuracy was 61%.

Representation 4: This representation uses both the

2 shell representation for the radial positions of the non-sequence neighbors, the representation of the identity of the central residue, and the representation of the identity of the sequence neighbors. The input size is 100. The prediction accuracy was 61%.

Representation 5: This representation uses the C_β representation. It includes 2 radial shells, octant binning for non sequence-neighbor positions, and also the residue identity and 4 bin spatial representation for the sequence neighbors. Together with the representation of the identity of the central residue, there are 388 inputs $(20 + 2 \times 24 + 16 \times 20)$. The prediction accuracy was 68%.

Representation 6: This representation is identical to Representation 5, but with additional information on exposure and secondary structure of the central residue included. The program Dssp (Kabsch & Sander 1983) was used to calculate the exposure and secondary structure (alpha helix, beta strand, and coil) of the central residue. The secondary structure was represented by a 3 bit unary representation. The calculated solvent accessibility was histogrammed and binned into 4 bins (0-50,50-100,100-150,>150) which were represented by a 4 bit unary representation. This additional secondary structure and accessibility information was concatenated on to the input representation used in Representation (5). The input size that resulted was 395 (388+7). The prediction accuracy was 71% and is the highest accuracy we were able to achieve with this data.

To test the effect of higher order interactions we repeated the same run using a no-hiddens architecture. The prediction accuracy dropped from 71% to 58%, indicating the importance of higher order interactions.

Representation 7: This representation tests the accuracy achievable using only pair-wise information about contacting residues, and first order information about secondary structure and accessibility. This presents the network with the explicit first order and second order pair-wise information that is used in previous constructions of pair-wise contact potentials. No additional spatial information is available to the network. The pair-wise contacts of residues with the central residue were represented by $20 \times 20 = 400$ inputs neurons, corresponding to the number of pairs of each type of the 400 possible (central residue, spatial neighbor) contacts. In addition the input contains the secondary structure classification of the central residue and its solvent accessibility, calculated by the Dssp program (Kabsch & Sander 1983). Solvent exposure and secondary structure were represented as before, using four bits and three bits, respectively. The total input size was therefore 407.

The prediction accuracy of the network using twenty hidden units was 62%, which is significantly less than the 71% accuracy achieved with spatial information explicit in Representation 6. An architecture with no hidden units achieved an accuracy of 57%.

Table 1: The input size and prediction accuracy for each representation (see text).

rep.no.	size	network arch.	accuracy
1	40	20 hidden	0.60
2	60	20 hidden	0.61
3	80	20 hidden	0.60
4	100	20 hidden	0.61
5	388	20 hidden	0.68
6	395	20 hidden	0.71
6	395	perceptron	0.58
7	407	20 hidden	0.62
7	407	perceptron	0.57
8 (C_α)	386	20 hidden	0.65

Representation 8: This is the only C_α representation reported here. It includes the same information as representation 5. The only difference is that we use only 2 bins (instead of 4) to represent the position of the N neighbor, which makes the input size 386. The prediction rate obtained was 65%.

Conclusions

We have introduced a statistical, instead of a statistical mechanical, formalism with which to construct "contact potentials". This formalism has numerous advantages over previous approaches. First, clear distinctions between various contact potentials already proposed in the literature may be easily illuminated using this approach. Furthermore, the formalism allows the introduction of powerful machine learning techniques, such as neural networks, for constructing contact potentials that include higher than second order pair-wise interactions. An essential ingredient of our approach is the development of a natural spatial representation, based on back bone co-ordinates, of the residues within an interaction radius of a central residue.

In a series of numerical investigations we compared the accuracy achieved using the new statistical formalism coupled with neural networks, to that achieved by limiting interaction information at second order. A significant increase in predictive accuracy (using Representation 6) was obtained, compared to a representation that limited interaction information to second order (pair-wise Representation 7).

The better result obtained with the C_β representation 5, compared to the similar C_α representation 8, suggests that, as expected, the positions of the C_β atoms provide more information about the interactions between the residues in a local environment. The C_α representation, however, may become useful in scenarios where the C_β positions are unknown.

We also tested the network using the more conventional "structure-matches-sequence" protocol on a smaller, but still representative selection of data. In this test the network correctly identified, for each pro-

tein, the native sequence from among the other sequences belonging to the other proteins in the predict set. All sequences equal to, or longer, than each test protein structure were evaluated in all (ungapped) alignments, and the log-likelihood for the sequence was taken to be the sum of the log-likelihoods for the individual spheres and was used as a discriminant. These results will be reported elsewhere, as well as more rigorous comparisons with standard empirical potentials.

We are currently investigating additional representations for the spatially neighboring residues within an interaction radius of a central residue, and improved methods to form the optimal spatial binning. Techniques for optimizing the network architecture (or combining networks) for better prediction will be tried as well. Other directions of research we are pursuing now following the work described here are: building a "library" of neighborhood spheres that serve as tertiary structure building blocks (which can be used in protein design), and using the neural net potential together with structural models to predict secondary structure (and other 3D structural motifs) from sequence.

Acknowledgments

The authors wish to thank the Santa Fe Institute where part of this research was performed. In addition, the authors wish to express their appreciation to the Advanced Computing Laboratory at Los Alamos National Laboratory for use of their facilities. This research was funded by the Department of Energy.

References

Baum, E., and Wilczek, F. 1987. Supervised learning of probability distributions by neural networks. In Anderson, D., ed., *Neural Information Processing Systems*, 52–61. AIP Press.

Bernstein, F., and et al. 1977. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Bowie, J.; Luthy, R.; and Eisenberg, D. 1991. A method to identify protein sequence that fold into a known three-dimensional structure. *Science* 253:164–170.

Branden, C., and Tooze, J. 1991. Gerland Publishing, New York.

Bryant, S., and Lawrence, C. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins: Strucutre, Function and Genetics* 16:92–112.

Godzik, A.; Kolinski, A.; and Skolnick, J. 1992. A topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* 227:227–238.

Goldstein, R.; Luthey-Schulten, Z.; and Wolynes, P. 1992. Protein tertiary structure recognition using optimized hamiltonians with local interaction. *Proc. Natl. Acad. Sci. USA* 89:9029–9033.

Hobohm, U., and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Science* 3:522.

Hopfield, J. 1987. Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proc. Natl. Acad. Sci. USA* 84:8429–8433.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *BioPolymers* 22:2577–2637.

Lathrop, R. work in progress (private communication).

Miyazawa, S., and Jernigan, R. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 18:534–552.

Sippl, M. 1990. Calculation of conformational ensembles from potentials of man force. *J. Mol. Biol.* 213:859–883.

Wilmanns, M., and Eisenberg, D. 1993. Three-dimensional profiles from residue-pair preferences: Identification of sequences with β/α-barrel fold. *Proc. Natl. Acad. Sci. USA* 90:1379–1383.