Design of an optimal Chebyshev-expanded discrimination function for globular proteins

BORIS FAIN, YU XIA, AND MICHAEL LEVITT

Department of Structural Biology, Stanford University, Stanford University School of Medicine,
Stanford, California 94305, USA

(RECEIVED January 4, 2002; ACCEPTED May 8, 2002)

Abstract

We describe the construction of a scoring function designed to model the free energy of protein folding. An optimization technique is used to determine the best functional forms of the hydrophobic, residue-residue and hydrogen-bonding components of the potential. The scoring function is expanded by use of Chebyshev polynomials, the coefficients of which are determined by minimizing the score, in units of standard deviation, of native structures in the ensembles of alternate decoy conformations. The derived effective potential is then tested on decoy sets used conventionally in such studies. Using our scoring function, we achieve a high level of discrimination between correct and incorrect folds. In addition, our method is able to represent functions of arbitrary shape with fewer parameters than the usual histogram potentials of similar resolution. Finally, our representation can be combined easily with many optimization methods, because the total energy is a linear function of the parameters. Our results show that the techniques of Z-score optimization and Chebyshev expansion work well.

Keywords: Proteins; energy functions; optimization; Chebyshev

Any attempt to predict the three-dimensional structure of a protein's native state demands the knowledge of the interaction potential between the amino acid residues (Miyazawa and Jernigan 1985; Sippl 1990; Bowie et al., 1991; Goldstein et al. 1992a; Jones et al. 1992; Bryant and Lawrence 1993; Ouzonis et al. 1993; Bauer and Beyer 1994; Srinivasan and Rose 1995; Seno et al. 1998). In principle, the behavior of molecules, including proteins, is described by quantum mechanics and quantum field theory. In practice, however, both analytical solutions and computer simulations bog down for any molecule larger than a few dozen atoms. One way to get around the computational complexity is to replace the quantum description of atoms and molecules by points and spheres that interact with Newtonian forces that depend on distances and angles between atom centers. This approach—molecular dynamics—can describe the behavior of gases, crystals, and simple liquids with excellent accuracy (Levitt et al. 1995). Still, even with today's powerful computers, molecular dynamics cannot simulate nature for a long enough time to fold even a small protein surrounded by water. This forces researchers to develop simpler representations and effective energies that retain sufficient detail to keep the results biologically interesting, but are simple enough to make the protein-folding problem tractable. This work describes one such attempt.

Some of the techniques of this work originate in Fain et al. (2001). In our previous work, we used Chebyshev expansion and Z-score optimization to derive a scoring function that models the hydrophobic interaction. In this work, we expand the model to include hydrogen bonding and general pairwise interactions between amino acids. Using appropriate transformations, we represent hydrophobic and pairwise interactions as a sum of Chebyshev polynomials. The Chebyshev representation requires few constants to give excellent accuracy, allows an arbitrary shape of the underlying function, and allows the potential to be decomposed as a linear combination of parameters. Presenting the effective Hamiltonian and its Chebyshev expansion is one of the main points of this work.

Reprint requests to: Boris Fain, Department of Structural Biology, Stanford University, Stanford School of Medicine, Fairchild Bldg. D-109, Stanford, California 94305, USA; e-mail: bfain@stanford.edu; FAX: (650)723-8464.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.0200702.

Once the form of the effective potential is defined, one has to then determine the variable parameters that determine the function. We find the parameters by adapting the recently developed (Mirny and Shakhnovich 1996; Seno et al. 1998; Xia and Levitt 2001) method of Z-score minimization from its use in lattice models and discrete contact potentials to off-lattice models and continuous potentials. We train our procedure by minimizing the native Z-score with respect to energies of ensembles of decoys obtained by random perturbations of the native structure. Admittedly, in our current work, this training procedure places the native conformations in a local, rather than a global minimum. (The terms global and local minima are used in the context of discrimination. In other words, test sets sample the space, and not, for example, a minimization procedure.) However, our tests will show that satisfying this requirement produces a function possessing a high degree of discrimination.

After postulating and deriving an effective energy, we are obliged to test it. We evaluate the performance of our derived effective energy by applying it to decoys from the complete Park-Levitt (Park and Levitt 1996) set, as well as the sets from the Decoys'R'Us (or DD) (Samudrala and Levitt 1999) database. The decoy sets in this database are a challenging test for a discrimination function, because the sets include many near-native conformations (C^{α} RMSD of $< = 4$ Å). Other researchers have used the *DD* database to evaluate potentials, (Park and Levitt 1996; Betancourt and Thirumalai 1999; Huang et al. 1999; Simons et al. 1999; Toby and Elber 2000); hence, we can easily gage how our method compares with others.

Theory and models

Simplified representation of the protein

It is currently believed that all-atom potentials are required to properly model the dynamics of protein folding (Van Gunsteren 1989). The high level of detail combined with the (relatively) slow speed of today's computers limit the time scale over which we can follow the folding process to the order of one microsecond (1 μs). A common solution to the burden of computational complexity is to avoid a detailed description of the amino acid by representing the side-chain with a point approximating the side-chain's centroid (Levitt 1976). This level of detail is sufficient for our current goal of achieving 1 to 4 Å resolution.

Our model of the protein consists of the backbone heavy atoms and the virtual amino acid interaction centers. We also include the polar hydrogens involved in backbone hydrogen bonding. The model is shown in Figure 1. In our representation, the virtual side-chain (R) is a point 3.0 Å from the C^{α} along the C^{α}–C^{β} vector. (C^{α} and C^{β} refer to the alpha; and beta; carbons of the polypeptide chain.) We position the GLY centroid at the C^{α} atom. The model can be used (Bryant and Lawrence 1993; Huang et al. 1996; Simons et al. 1997) with either a fixed centroid distance or a sequence-dependent distance. We chose to use a fixed distance for two reasons: It is simpler, and the penalty in performance is minimal. The main-chain hydrogen atom coordinates are computed when the model is initialized and are subsequently moved along with the rest of the chain.

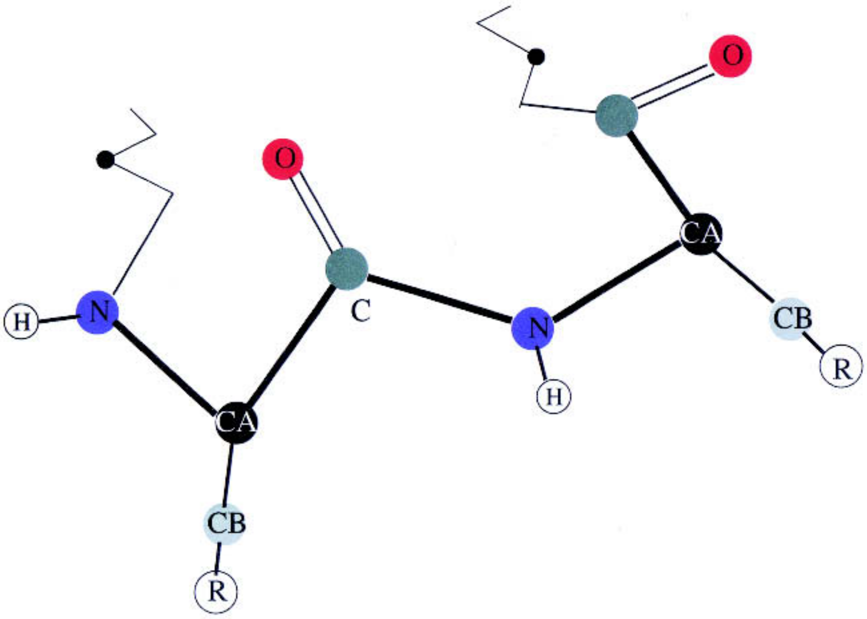

Fig. 1. Representation of the protein containing the main-chain heavy atoms C^{α}, C, N, and O. Also present are the virtual interaction centers representing the sidechain, R, as well as the main-chain hydrogens involved in forming hydrogen bonds. The C_{β} atoms are shown in the figure, but are not included in the calculation.

Potential function

At atomic resolution, biological structures and processes depend on the interplay of three fundamental noncovalent interactions and their corresponding entropic factors, electrostatics, hydrogen bonding, and the Van der Waals interaction. Our simplified effective energy depends on forces that mirror the three fundamental interactions. It includes backbone hydrogen bonding, hydrophobic burial, and a generalized pairwise interaction between residues. The pairwise interaction encompasses solvent-adjusted electrostatic interactions, as well as hydrogen bonding between residues. It also encompasses gradual excluded volume repulsion between residues, and any other solvent and entropic effects that we cannot describe explicitly.

Our initial assumption is, therefore, that the effective free energy of each residue depends, in a completely arbitrary way, on its degree of burial, on the distance and identity of its neighbors, and on the number and distribution of backbone hydrogen bonds as follows:

$$E_{tot} = E_H + E_{Burial} + E_{Pairwise} \tag{1}$$

Let us address each term in equation 1.

Hydrogen bonding

Our Hamiltonian contains a term for main-chain hydrogen bonding. We do not include side-chain hydrogen bonding interaction explicitly—it cannot be accurately reproduced with our simplified model (Fig. 1) because our side-chain representation lacks sufficient detail. The interaction is, however, implicitly included in pairwise energies (equation 3).

We model the free energy of main-chain hydrogen bonding by weighing each hydrogen bond with two terms.

$$E_H = \sum_{i=1}^{N_{res}} (E_{bond} + E_{env})_i \tag{2}$$

The first term is a step function in angle and distance from the donor and acceptor (Pauling 1960)—it simply records that a hydrogen bond has been formed. The bond is considered formed if the distance between H and O is <2.8 Å and the corresponding bond angle is <90 degrees (a perfect angle for the hydrogen bond is 180 degrees). The second term is a correction that depends on the bond's environment. We define the environment by the presence of another hydrogen bond in nearest and next-nearest neighbors on both sides of the bond in question. We choose five environment classes corresponding to the following situations: bonds present in (1) no neighbors, an isolated bond, (2) both nearest neighbors (e.g., in the middle of an α-helix), (3) both next nearest but not nearest neighbors (e.g., in an edge strand of a β-sheet), (4) no bonds on one side but a nearest-neighbor bond on the other (e.g., at the end of an α-helix), and (5) the same as 4, except with a next-nearest bond instead of a nearest bond. Table 1 illustrates the different environments a bond might have.

The environment term in equation 2 introduces cooperativity and encourages formation of hydrogen bond networks—a characteristic feature of proteins—at the small expense of four additional parameters (E_{env} for environments 2–5). An additional underlying physical reason for weighing the environments in Table 1 differently is conformational entropy. The formation of a hydrogen bond in the middle of a helix when all other bonds are formed is no great achievement. In contrast, the formation of a hairpin turn and the initialization of a β zipper from an extended configuration is far less likely considering the many alternate conformations available to the chain.

Pairwise interactions

To model the interaction between residues produced by solvent-mediated electrostatic forces, the Van der Waals interaction, and covalent bonds, we introduce a pairwise energy term, which is residue specific, and is an a-priori undetermined function of the distance between two residue centroids

$$E_{pairwise} = \sum_{i=1;j=i+1}^{N} [E_{pair}^{AB}(r_{ij})] \tag{3}$$

in which r_{ij} is the distance between the centroids of side-chains of residues i and j (see Fig. 1). Each of the 210 functions E_{pair}^{AB} (r_{ij}) is expanded by use of Chebyshev polynomials (defined in Appendix A). The Chebyshev expansion is a good choice for several reasons. First of all, it is a complete basis of functions on a finite interval, so that we will be able to recreate energies of arbitrary shape. Furthermore, the Chebyshev approximation is very close to the minimax polynomial, that is, they distribute the residual error almost equally throughout the approximated interval.

To represent the E_{pair} with Chebyshev polynomials, we have to transform the semi-infinite interval [1, ∞] into a

Table 1. *Classification of hydrogen bond environment*

Environment	Parameter	Description/example
00**1**00	1	An isolated hydrogen bond
x1**1**1x	C_2	Middle of helix or large sheet
10**1**01	C_3	Middle of edge strand
x1**1**00	C_4	End of helix or large sheet
10**1**00	C_5	End of edge strand

1 denotes a bond, 0 denotes no bond, x denotes either 0 or 1. The parameter for an isolated hydrogen bond equals 1 to set the scale for all other parameters. The bond under consideration is in boldface.

compact one. In proteins, the centroids are always farther than 1 Å apart. We choose the transformation $r \rightarrow 1/r$ which takes $[1, \infty]$ to $[0, 1]$. The harmonic transformation gives us an added advantage of allowing greater resolution on the smaller and, presumably, more important distances

$$E^{AB}_{pair}(r_{ij}) = \sum_{k=0}^{k_{max}} C^{AB}_k T_k(1/r)) \tag{4}$$

in which $T_k(x)$ is the Chebyshev polynomial of order k (see equation A2). We truncate the expansion at $k_{max} = 6$, which gives us 6 Chebyshev coefficients, C^{AB}_k for each of the 210 amino acid pairs. Consulting formulas A2 in Appendix A, we see that $k_{max} = 6$ gives us a resolution of less than 1 Å on the interval from 0 to 4 Å. The 0'th term is a constant and can be omitted because the amino-acid composition does not change in folding prediction. (This would not be true, for example, in sequence design.) Because the argument of the Chebyshev polynomials takes values on $[0, 1]$ only (this is because $r \geq 0$) we can pretend the desired function is even in $1/r$ and drop terms with $k = 1,3,5$. The pairwise potential is thus defined by $210 \times 3 = 630$ parameters. This number is less than, say, 2730 parameters used to define the very economical pairwise potential Toby and Elber (2000). In addition to the economy of term, our potential has the same resolution and is continuous.

Hydrophobic potential

We have studied the effect of the hydrophobic potential alone in a previous publication (Fain et al. 2001), and we have shown that this term, as expected (Kauzmann 1959; Perutz et al. 1965; Rose et al. 1985; Dill 1990), can have significant discriminating power. The following section is a short summary of our previous work.

When a hydrophobic residue is buried in the interior of the protein, it will necessarily have many neighboring residues. Viswanadhan (1987) has shown that the average number of neighbors with 10 Å of a given residue correlates well with its hydrophobicity. Thus, we assume that the hydrophobic energy contribution from each residue will depend on the number of residues within a 10 Å shell surrounding it. Explicitly,

$$E_i = E_a(n), \tag{5}$$

where a denotes a specific amino acid type, and n is the number of centroids within 10 Å of a given residue centroid. $E_a(n)$ has some a priori unknown form, and we shall represent it as a linear combination of suitably chosen appropriate basis functions. Once again, we choose to represent the burial potential as a linear sum of Chebyshev polynomials (see Appendix A). Because the Chebyshev representation is most naturally applied to functions defined on the interval $[-1, 1]$, we transform the functional dependence of E on n as follows:

$$E_i = E_a \left(\frac{n - 10}{n + 10} \right). \tag{6}$$

The transformation $n \rightarrow (n - 10)/(n + 10)$ maps the possible number of near neighbors $[0, \infty]$ to an interval $[-1, 1]$. We chose 10 as the crossover point because 10 is roughly the number of neighbors with which an amino acid becomes buried. It is not necessary to choose this parameter exactly, but getting it in the right ballpark helps the Chebyshev expansion converge more rapidly.

The final functional form for the hydrophobic energy of a protein length N becomes

$$E_{burial} = \sum_{i=1}^{N} E_i = \sum_{i=1}^{N} \sum_{k} C_{A,k} T_k \left(\frac{n - 10}{n + 10} \right), \tag{7}$$

in which A indexes the amino acid type, k is the order of the Chebyshev polynomial T_k, and n is the number of neighbors within a 10 Å radius from the amino acid i.

Because we want the resolution of our potential to be of order 0.1 (which corresponds to being able to tell the difference between, say, 9 and 10 neighbors), we retain terms no higher than order 6 in the Chebyshev expansion. Once again, the 0th term is omitted because it is a sum of constants and contributes equally to any conformation of the same protein. We then have $20 \times 6 = 120$ coefficients $C_{A,k}$, which completely determine the burial potential. Just like the pairwise energies in equation 3, the representation of the burial potential assumes only that the burial energy depends on the amino acid type and its degree of burial. The effective energy can assume any functional form, which will be determined by optimizing its discriminating power.

Number of undetermined parameters

The total parameter count is as follows: 5 for hydrogen bonding, 630 for pairwise interactions, and 120 for burial interactions, totaling 755 parameters. Our potential has a spatial resolution of ¡1Å in the 1 to 4 Å range, and a burial resolution of at least ±1 neighbor. This compares favorably with the 2730 parameters of a detailed 1 Å histogram function (Toby and Eber 2000) and is many orders of magnitude smaller than the number of parameters in the excellent potentials developed by Simons et al. (1997) and Samudrala and Moult (1998).

Potential training and Z-score optimization

Currently, there are three distinct approaches to extracting coarse-grained potentials between pairs of amino acids. The first method, pioneered by Tanaka and Sheraga (1976) is based on the quasichemical approximation. It derives conformational energies by comparing the distributions of amino acids occurring in native structures of proteins with those in the random compact conformations. This approach has been used by many researchers (Tanaka and Sheraga 1976; Hendlich et al. 1990; Sippl 1990; Miyazawa and Jernigan 1996; Samudrala and Moult 1998), and has been well reviewed by Sippl (1995) and also by Wodak and Rooman (1993).

The main flaw of such potentials of mean force is the suspicion that the quasichemical approximation may not be valid (Ben-Naim 1997). Thomas and Dill (1996) tested the method on exactly solvable lattice models. They showed that although the extracted and exact potentials do have common elements (which accounts for the current popularity of potentials of mean force), the two do not correlate very well (Xia and Levitt 2001).

Another strategy is to insist that the native state is lower in energy than multiple decoys and to solve the resulting inequalities by use of Linear Programming. This approach, unlike the previous one, has no theoretical weaknesses because there is only one assumption—mainly that the native state is a minimum. We did not use LP because it occasionally fails to find a feasible solution (Toby et al. 2000) and because it is computationally intensive.

Still another strategy has also been a subject of considerable activity (Goldstein et al. 1992; Maiorov and Crippen 1992; Mirny and Shakhnovitch 1996; Chiu and Goldstein 1998; Xia et al. 2000). The basic idea is to parameterize a suitably chosen Hamiltonian, and then to adjust the parameters in such a way that a collection of native states assumes either the lowest or one of the lowest energies compared with an ensemble of incorrectly folded alternate structures. We shall use a variation of this method to optimize our potential.

We optimize our Hamiltonian by minimizing the median Z-score. We choose Z-score optimization over linear programming, not because it is fundamentally a more correct method—recent work actually suggests that the choice is not crucial (Vendruscolo 2000). We concentrate on the Z score because it is a statistical quantity, which implies that the computational burden does not increase with the number of decoys in each ensemble (See Appendix B for details). We also like Z-score optimization because this method possesses flexibility of scoring near-native and/or (allegedly) dynamically inaccessible conformations lower than native.

Our optimization scheme follows the general outline of the method of Mirny and Shakhnovich (1996) applied to a continuous potential and an off-lattice model. The procedure is identical to Fain et al. (2001) with the addition of new interaction terms.

We choose a training set and construct alternate structures for each chosen protein. We then optimize the parameters of our Hamiltonian to minimize the median Z score of the native structures relative to their corresponding alternates. (The details of the computation are described in Xia and Levitt 2001 and Fain et al. 2001, and in Appendix B.) We decided to optimize the median Z score and not the harmonic mean (Mirny and Shakhnovich 1996) or average, or the energy gap (Goldstein et al. 1992b), because the median is least affected by outliers; the average is sensitive to large values, the harmonic mean to small values; the median, however, is robust. We have tried all three, and the median gave us the best results, although the improvement was, admittedly, marginal.

Training and results

Construction of training sets

We trained our function on a set of 70 protein structures, listed in Table 2. The structures were selected by use of three criteria. First, to ensure variety in our training set, the sequence identity between any two proteins had to be <35%. Second, to reinforce structural variation, each protein was from a different SCOP (protein classification database) (Murzin et al. 1995) family. Finally, to avoid structures of low quality, we kept proteins with a SPACI (structure quality database) (Brenner et al. 2000) score of better than 0.25.

Each member of our training set consists of the native structure and a set of 1000 alternate conformations (decoys). For each sequence, we perturbed the native structure (starting the Monte Carlo trajectory with the native structure allows us to generate near-native decoys easily) with a simulated annealing routine (Metropolis et al. 1953; Nelder and Mead 1965) and produced 1000 alternate conformations. We designed the decoy-generation procedure to produce decoys for which the root mean square deviation (RMSD) from native ranged from 0 to the radius of gyration (RG) of the native structure. In addition, the simulated annealing was designed to produce structures with RG similar to that of the native conformation, thus ensuring compactness.

Typically, training sets are generated by threading onto diverse alternate structures (Park and Levitt 1996). It thus becomes exponentially difficult to produce decoys with near-native RMSDs by threading (Reva et al. 1998); and the

Table 2. *PDB names of proteins in the initial training set*

1a1x	1a32	1aep	1aho	1aie	1ail	1ako	1aly	1amx	1arb
1bb9	1bd8	1bea	1bfg	1bgf	1ble	1bm8	1bv1	1c25	1cby
1cex	1cfr	1chd	1cyw	1dun	1fna	1gpr	1gvp	1hoe	1hyp
1ilb	1ifc	1kid	1koe	1kte	1lcl	1lfb	1lki	1lxa	1mjc
1msc	1nkd	1noa	1pbv	1pdo	1pht	1pne	1pft	3pte	1rcb
1rpl	1sfp	1tfe	1tig	1tlk	1tud	1tul	1utg	1vcc	1vie
1wer	1whi	1who	1xat	2end	2igd	2pii	2pth	2rn2	2tgi

resulting decoy set is not challenging. We did not use decoys generated by gapless threading, because we feel that such ensembles are not sufficiently challenging for either evaluation or training of potentials. Because we wanted to force the function to differentiate between native structure and alternate structures that have a relatively low C^{α} RMSD from the native (1 to 6 Å), we likewise wanted our training sets to have the same characteristics as our test sets. Perturbing the native structure automatically makes the near-native region of conformational space accessible. The disadvantage of this approach is that we are only enforcing the native structure to be in a local minimum of conformational space.

Potential optimization

Training and purification

The training procedure attempts to find the lowest median Z score for all of the ensembles corresponding to the training proteins (listed in Table 2). The reader should consult equation B6 and surrounding paragraphs.

The parameter space consists of $5 + 120 + 630 = 755$ variables. Because a priori we have no notion what the Z-score surface looks like, and also because 755 variables is still a relatively large number, we chose to minimize with a simplex version of the simulated annealing procedure (Metropolis et al. 1953). The temperature is decreased linearly from $\tau = 100$ (an arbitrary large value) to $\tau = 10^{-3}$. The annealing is restarted several times. At the end of the run, the result is refined with a downhill-simplex method (Nelder and Mead 1965) which is equivalent to setting $\tau = 0$.

We cannot be certain that each of the proteins in our training set represents a true minimum of energy. It is possible that cofactors unlisted in the PDB file were present either in vitro or in vivo. Alternately, the shape of the molecule may have been significantly distorted by crystallization. The molecule may also actually be a dimer or a multimer. To safeguard against these errors, we ensured self-consistency by removing from the training set all proteins that did not (simultaneously with the rest) achieve a Z score lower than −0.8 using the burial function alone (Fain et al. 2001). We reasoned that undesirable subunits will have good pairwise energies and good hydrogen bonding energies, but will have an exposed hydrophobic surface where they should join other pieces of the stable structure. The value −0.8 is high enough to allow small disulfied-rich proteins to remain in the training set.

Four proteins were removed from the initial set, leaving a final training set of seventy proteins. It is interesting to note that every rejected protein had one of two features; they were either biologically active as a dimer or a tetramer, or formed an extended structure (see Table 3). This was a comforting result, as (barring an unlikely event that unstable structures in our initial set outnumbered the stable ones) it was exactly what we expected.

Table 3. *Proteins rejected by the self-consistent burial optimization*

Name	Possible underlying reason
1gvp	Biologically active as a dimer
1utg	Biologically active as a dimer
1vie	Biologically active as a dimer
1aie	Biologically active as a tetramer

The derived potentials

After the self-consistent purification, the sets were optimized for the full potential of hydrogen bonding, burial, and pairwise interactions. The final median Z score for the training set is −6.21. A representative RMSD versus energy plot is shown in Figure 2.

A plot of the burial function for three representative amino acids is shown in Figure 3. Shown are the burial preferences for hydrophobic Valine, hydrophilic Arginine, and intermediate Glycine. Some anomalies in the burial energies are caused by the fact that our reference state consists only of compact structures. Notice that VAL prefers to have many rather than few neighbors. ARG would rather have a few neighbors than many, but, contrary to what one would expect, it also prefers few to none (this is because the best it can do is lie on the surface and still retain all of its neighbors to the interior). Finally, the burial preferences of GLY are between those of ARG and VAL. When examining Figure 3, the reader should note that energy values for less

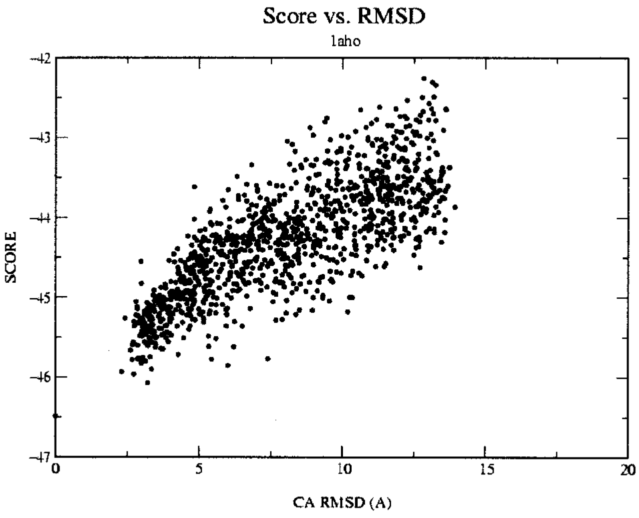

Fig. 2. A typical distribution of our derived energy vs. RMSD for an ensemble of decoys in the training set. These particular values come from the protein 1aho.

than 2 neighbors and 30 or more neighbors are arbitrary. All residues have at least 2 neighbors, and, because of their excluded volume, no residues have more than 30 neighbors. We have positioned the curves to coincide at 0,0 so that the relative energies can be easily compared.

Figure 4 shows the pairwise energy for CYS–CYS, VAL–LEU, TYR–VAL, VAL–VAL, and PHE–TRP interactions. The plots display several expected features. CYS–CYS has the deepest well because of the formation of disulfide bonds; the location of the minimum corresponds exactly to the distance of the disulfide bond (2.2 Å). Hydrophobic residues are frequently in the interior of the protein, and their contact energy is correspondingly lower. Even pairs of residues like TYR–VAL and VAL–VAL, which show no particular preference of being close, have a mild minimum at short distances to enforce overall compactness of the protein. A nice feature of both Z-score minimization and our representation is that the curves automatically enforce excluded-volume constraints by rising up swiftly to infinity at close distances.

One final observation is that the value of the pairwise interactions, the hydrophobic interactions, and the hydrogen bonding energy that our procedure derives are roughly similar, mirroring the fact that all three are important in protein folding.

The coefficients of the potential can be found on www.stanford.edu/fain or by e-mailing B.F.bfain@stanford.edu.

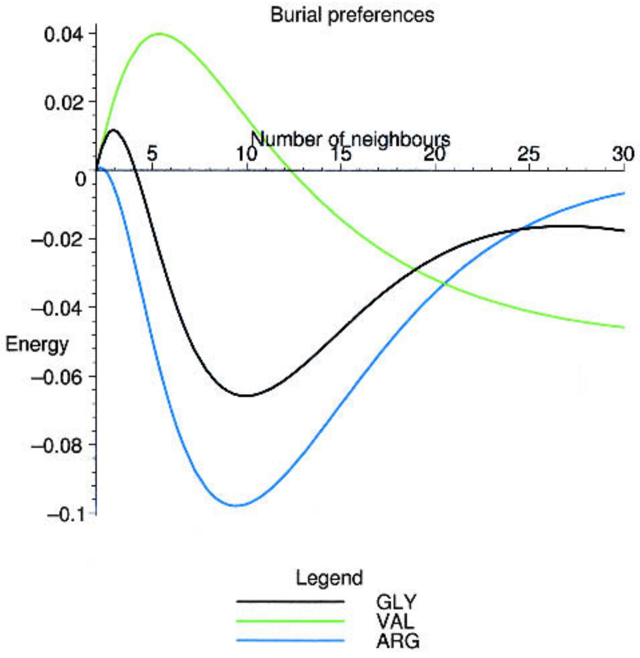

Fig. 3. Burial preferences for VAL, ARG, and GLY. Regions below 2 neighbors and above 30 neighbors do not contribute to discrimination. Note that hydrophilic proteins (GLY, ARG) prefer few neighbors, whereas hydrophobic residues (VAL) prefer many neighbors. The curves are positioned to coincide at 0,0 so that they can be compared easily.

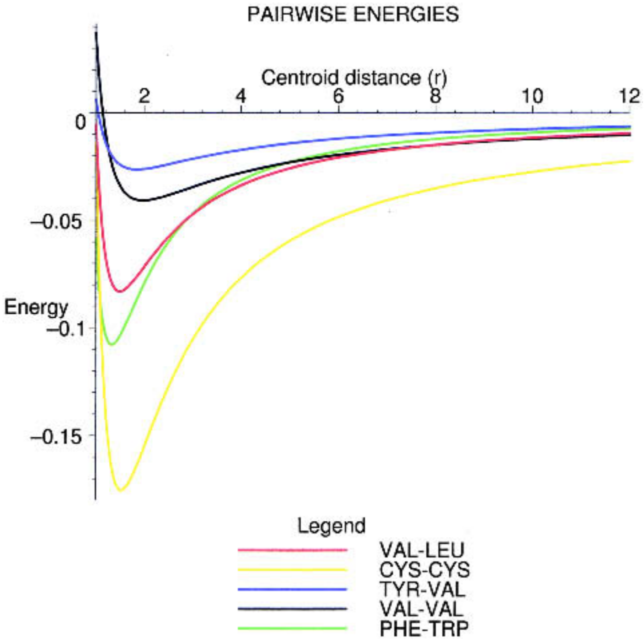

Fig. 4. Pairwise potential as a function of inter-centroid distance for CYS–CYS, VAL–LEU, TYR–VAL, VAL–VAL, and PHE–TRP interactions. The shape and the deep minimum of the CYS–CYS curve corresponds to disulfide bond formation. All of the curves have a minimum at small distances reflecting an overall preference for compact shapes and the periodicity of secondary structure. The deeper minimum of the PHE–TRP pair reflects the large number of atoms in these side chains. The curves have been positioned to approach 0 at ∞.

Discrimination power

We evaluated the performance of our function on several independently produced test sets that contain significant number of near-native decoys.

The first collection of decoys, also known as the Park Levitt (PL) set (Park and Levitt 1996), was produced by perturbing the loop degrees of freedom of the native structure and then selecting protein-like conformations. This set is the first collection of misfolded structures and has since been used in several comparisons of potentials (Park and Levitt 1996; Simons et al. 1999). The set is preconditioned by keeping only the conformations for which the radius of gyration is greater than 3 * $N_{res}^{1/3}$ (Park and Levitt 1996; Simons et al. 1999). We performed this compactness filtering for two reasons. First, in a discrimination scenario it is common and useful to precondition the set by first filtering out noncompact structures, instead of placing an extra requirement on the final discrimination function. Second, both Park and Levitt (1996) and Simons et al. (1999) had prefiltered the Park Levitt set for compact structures, and had we not done it, we would not have been able to compare our results to theirs.

The filtering reduces the set to ~400,000 structures in 8 sets. The full backbone is then reconstructed using the program ENCAD (Levitt 1992). (The complete set is too large to deposit on the web, but we will send it to anyone who wishes to use it.) The data from the full Park Levitt set is presented in Table 4. We compare our performance on the full Park-Levitt set with the results reported by Simons et al. (1999) (DB), and also with potential developed by Samudrala and Moult (1998) (SA). Judging by the most recent CASP results, DB and SA are among the best discriminating functions in current practice.

Table 4 shows that our Z scores (BF) are slightly worse than SA, and are considerably better than DB. This is very encouraging, especially when one considers that we have 755 parameters compared with well over 10^4 parameters for DB and 260,000 parameters for SA.

Evaluation of functions has been made convenient by a web-based database of decoy sets, Decoys'R'Us (Samudrala and Levitt 1999) (http://dd.stanford.edu). We tested our potential on three families of decoy sets from the following collection: 4-state-reduced, lmds, and lattice-ss-fit. The first set, 4-state-reduced, is the reduced version of the Park Levitt set with just 1000 select decoys per protein. The second collection of decoys from dd.stanford.edu, lmds, was produced by Keasar and Levitt by using minimization with a complex potential that contains a significant pairwise component as well as cooperative hydrogen bonds. The lmds family is significant because pairwise potentials are easily fooled by this set, possibly because each member of the set is itself a local minimum of a pairwise potential. The third and final set of decoys, lattice-ss-fit, was actually produced in a predictive scenario during CASP3 (Samudrala et al. 1999). Once again, none of the test set proteins had more than 35% identity to any proteins in our training set in Table 2. We also excluded proteins that were a single chain of a multimer, on the suspicion that the true ground state of those requires all of the chains to be present.

Table 4. *The performance of scoring functions on the full Park-Levitt decoy set*

Protein	No. decoys	BF Z	DB Z	SA Z
1ctf	61984	−3.8	n/a	−3.7
1erp	167532	−2.6	−3.5	−3.7
1r69	61206	−1.7	−2.0	−2.2
1sn3	28125	−4.7	−2.2	−5.3
1ubq	6996	−6.7	−3.1	−5.8
2cro	54677	−1.4	−1.2	−1.8
3icb	47732	−1.7	−2.8	−2.0
4pti	41764	−2.7	−2.3	−3.7
4rxn	44418	−6.4	−2.8	−4.0
ave		−3.5	−2.5	−3.6

Performance of the Baker DB set was taken from literature (Simons et al. 1999). The SA code was provided by Ram Samudrala. BF is the current potential. Z-score is defined in Appendix A. The Spearman Rank-order coefficient (S) is a robust measure of correlation.

Recently Toby and Elber (2000) derived an excellent pairwise potential and tested it on the same sets of decoys. We compare our results with Toby and Elber (2000) (TE-13) and, indirectly, all of the other potentials mentioned in Hinds and Levitt (1991); Godzik et al. (1995); Miyazawa and Jernigan (1996); Bahar and Jernigan (1997); Betancourt and Thirumalai (1999), all of which TE-13 were better than. We also compare our performance once again to SA (Samudrala and Moult 1998), and to results produced by MJ, a contact potential derived by Miyazawa and Jernigan (1996).

The results of scoring the Decoys'R'Us database are in Table 5 and the score versus RMSD graphs for the 4-state-reduced set are in Figure 5. The 4-state-reduced set contains a significant fraction of near-native decoys, therefore, in addition to the Z scores and native rank, we also report the C^{α} RMSD of our lowest energy conformation, and the Spearman rank-order coefficient (essentially a robust correlation coefficient) (R).

To show that the addition of the pairwise component improves the discrimination of native and native-like conformations, we compare the full potential presented in this work to one we have presented previously (Fain et al. 2001), which did not contain the pairwise interaction. The comparison of the two functions' performance on the reduced Park-Levitt set is presented in Table 6. The addition of pairwise interactions improves both the successful recognition of a near-native structure as the lowest energy state, as well as the Z score of the native conformation.

Our potential performs well. The native Z score of our function averaged over all sets is slightly lower than the other functions we have tested. The results with 4-state-reduced set are especially interesting because that set contains many near-native decoys. Although, in most cases, the native structure is not our lowest conformation, we always pick a decoy that is very close to the native state, which is a mistake we can easily live with. The correlation (R) and the score versus RMSD graphs in Figure 5 show that we successfully separate near-native structures from the rest of the decoys. However, in this set, the SA, TE-13, and MJ potentials discriminate the native better than we do. Results from the lmds collection are illuminating because on this set potentials of mean force—MJ and SA—begin to falter slightly. The two functions are often completely unable to distinguish natives from near-natives, and occasionally (1bba and 2fc2) endow the native with the highest energy. The last set, lattice-ss-fit, is made up of structures that are somewhat distant from the native state (although much closer than any threading set) and we, along with others, have no trouble picking out the native with very high probability (low Z scores). Overall, our discriminating function consistently and reliably separates native conformations from decoys.

Table 5. *The performance of scoring functions on the Decoys‘R’Us collection of decoys*

		BF				TE-13		SA			MJ	
Protein	Number decoys	rank	Z	R	best (Å)	rank	Z	rank	Z	R	rank	Z
4-state reduced												
1otf	631	1	−2.9	0.63	1.97Å	1	−4.2	1	−2.9	0.74	1	−3.7
1r69	676	11	−2.0	0.59	2.48Å	1	−4.6	1	−2.4	0.78	1	−4.1
1sn3	661	5	−2.7	0.40	1.31Å	2	−2.7	1	−3.3	0.45	2	−3.2
2cro	675	22	−1.7	0.42	3.82Å	1	−3.5	1	−2.5	0.72	1	−4.3
3icb	654	55	−1.4	0.76	1.92Å	—	—	22	−1.7	0.82	—	—
4pti	688	13	−2.0	0.29	1.49Å	7	−2.4	1	−3.3	0.60	3	−3.2
4rxn	678	2	−3.4	0.52	2.08Å	16	−2.0	1	−2.6	0.59	1	−3.1
1rnds												
1bba	501	35	−1.5			—	—	501	+9.4		—	—
1ctf	498	1	−3.4			1	−4.1	1	−3.7		1	−3.9
1dtk	216	10	−1.8			5	−1.9	23	−1.3		13	−1.7
1fc2	501	3	−2.6			14	−2.0	499	+2.7		501	+6.2
1gpt	101	2	−2.0			—	—	4	−1.8		—	—
1igd	501	1	−3.4			2	−3.1	1	−5.0		1	−3.3
1shf-A	438	1	−5.6			1	−4.1	1	−4.5		11	−2.0
1ubi	301	1	−4.3			—	—	1	−4.3		—	—
2cro	501	12	−3.1			1	−4.0	359	+0.6		1	−5.1
2ovo	348	3	−2.5			1	−3.6	18	−1.7		2	−3.3
4pti	344	16	−1.6			—	—	12	−1.7		—	—
lattice	ssfit											
1beo	2001	1	−10.1			—	—	1	−9.4		—	—
1ctf	2001	1	−6.6			1	−6.2	1	−6.7		1	−5.4
1dkt-A	2001	1	−5.4			2	−3.9	1	−5.6		32	−2.4
1fca	2001	1	−5.1			36	−2.3	1	−5.6		1	−3.4
1nkl	2001	1	−7.8			1	−4.5	1	−7.3		1	−5.1
1pgb	2001	1	−9.9			1	−4.1	1	−8.9		1	−2.2
1trl-A	2001	1	−4.9			1	−3.6	1	−3.9		4	−2.9
4icb	2001	1	−4.9			—	—	1	−4.3		—	—
ave			−3.95				−3.52		−3.14			−2.95

The definition of Z-score can be found in Appendix B. TE-13 is the potential reported by Tobi and Elber (2000), SA is the potential developed by Samudrala and Moult (1998), MJ is the potential from Miyazawa and Jernigan (1996), and BF is our potential. R is the Spearman rank-order coefficient (a robust measure of linear correlation).

Discussion

We had several goals when we started this work. First and most important, we wanted to devise an economical and flexible way to represent continuous potentials of arbitrary shape. Our method of expanding both burial (equation 7) and pairwise (equation 4) interactions using the Chebyshev polynomials achieves this. The expansion requires relatively few parameters to represent functional forms of arbitrary shape. This formalism can be useful to other researchers who are constructing effective potentials.

Our second goal was to test how well Z-score optimization can train a potential. Our training set has a serious theoretical drawback, namely, that the native state is required to only a local minimum of the energy surface (Betancourt and Thirumalai 1999). However, our final results suggest that this requirement is sufficient to produce a reasonably good discriminating function. Tables 4 and 5 show that the discriminating function we derived has considerable merit. The final Z scores are not nearly as important as the fact that we are able to separate native from nonnative states consistently across a variety of decoy sets.

Appendix A: Chebyshev expansion

Definition

An excellent exposition of the Chebyshev expansion can be found in Chapter 5 Press et al. (1992). We restate briefly some of the properties of the representation. The Chebyshev polynomial of degree n is defined by

$$T_n(x) = \cos (n \arccos x). \qquad \text{(A1)}$$

Explicitly, the polynomials are

$$\begin{aligned} T_0\,(x) &= 1; \\ T_1\,(x) &= x; \\ T_2\,(x) &= 2x^2 - 1; \\ T_3\,(x) &= 4x^3 - 3x; \\ T_4\,(x) &= 8x^4 - 8x^2 + 1. \end{aligned} \qquad \text{(A2)}$$

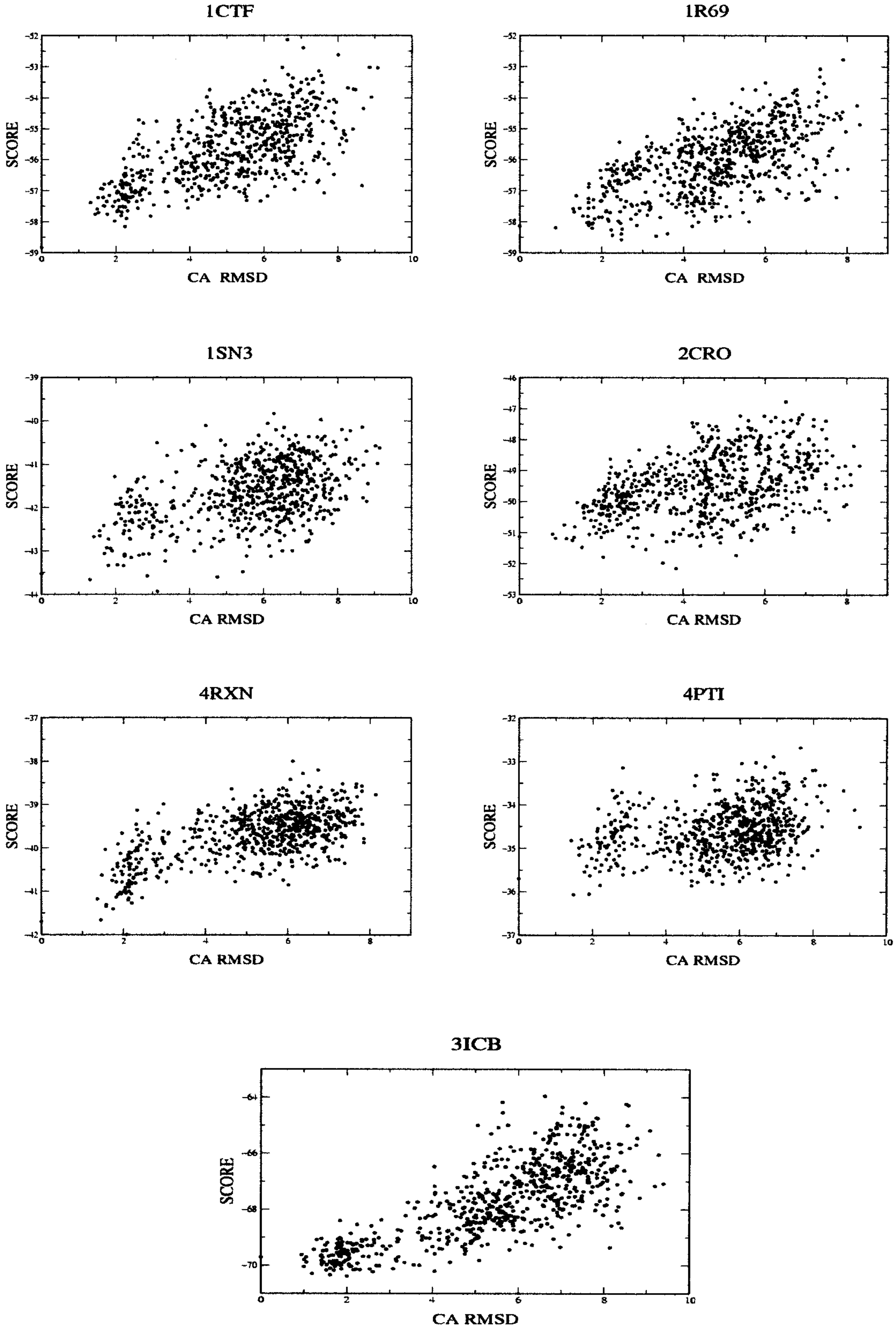

Fig. 5. Plots of score vs. RMSD for the 4-state-reduced decoy sets of Park and Levitt (1996). In all cases, native-like structures (≤4 Å RMSD) are distinguished from the ensembles.

Table 6. *This table shows the improvement in discrimination achieved by adding a pairwise component*

Protein	BF01 near-native rank	Z	BF (inc. pairwise) near-native rank	Z
1ctf	1	−2.9	1	−2.9
1r69	1	−2.4	1	−2.0
1sn3	1	−2.1	1	−2.7
2cro	1	−1.3	1	−1.7
3icb	1	−1.6	1	−1.4
4pti	9	−1.7	1	−2.0
4rxn	2	−2.2	1	−3.4
ave		−2.0		−2.3

BF is the current discrimination function, and BF01 is the function from Fain et al. (2001) that does not contain a pairwise component. Both functions are operating on the Park Levitt set. The table shows the rank of the lowest energy near-native structure as well as the native Z-score. Both recognition of near-native structures and the Z-score of the native conformation improve by the addition of pairwise interactions.

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)\; n \geq 1. \qquad \text{(A3)}$$

Although the T_n are defined only on the interval [−1, 1], a simple change of variable allows the expansion to be used to represent a function between two arbitrary limits, [*a, b*]:

$$y = \frac{x - \frac{1}{2}(b+a)}{\frac{1}{2}(b-a)}. \qquad \text{(A4)}$$

The Chebyshev expansion is useful for three main reasons. First, the error is spread out smoothly over the approximately interval; the expansion is nearly identical to the minimax polynomial. In addition, the Chebyshev approximation usually converges more rapidly than most other expansions and thus allows us to keep fewer coefficients to achieve the desired accuracy. Finally, recurrence relation A2 allows the Chebyshev polynomials to be computed very quickly.

Appendix B: Z-score optimization

Our optimization scheme minimizes the average Z score of the training set native structures relative to their respectively decoys. For each ensemble, the Z score is defined as

$$Z \equiv \frac{V(0) - \langle V(i) \rangle}{\sigma}, \qquad \text{(B1)}$$

in which the <> and σ are, respectively, the ensemble average and standard deviation of some function *V*. The 0 denotes the native conformation, and $0 \leq i \leq N$ runs over all of the conformations in the ensemble.

Because our potential function is a linear combination of terms

$$V(i) = \sum_{k=0}^{k_{max}} c_k V_k(i) \qquad \text{(B2)}$$

in which the c_k are the parameters we are optimizing, we can achieve significant simplification. We substitute equation B2 into equation B1. When the order of summation on *i* and *k* is interchanged, the numerator of equation B1 becomes

$$\sum_{k=0}^{k_{max}} c_k \left[V_k(0) - \frac{1}{N} \sum_{i=0}^{N} V_k(i) \right]. \qquad \text{(B3)}$$

Next, we consider the denominator, which is the square root of the variance. The variance of a linear sum can be decomposed as follows:

$$var\,(\sum_{k=0}^{k_{max}} c_k V_k) = \sum_{m=0}^{k_{max}} \sum_{n=0}^{k_{max}} c_m c_n cov\,(V_m, V_n), \qquad \text{(B4)}$$

in which the ensemble covariance matrix is defined as

$$cov\,(V_m, V_n) \equiv \langle (x_m - \mu_m)(x_n - \mu_n) \rangle, \qquad \text{(B5)}$$

in which μ is the ensemble mean. Putting equations B3 and B4 together, we get

$$Z = \frac{\sum_{k=0}^{k_{max}} c_k \left(V_k(0) - \frac{1}{N} \sum_{i=0}^{N} V_k(i) \right)}{\sqrt{\sum_{m=0}^{k_{max}} \sum_{n=0}^{k_{max}} c_m c_n cov\,(V_m, V_n)}}. \qquad \text{(B6)}$$

The value of equation B6 is that one can precalculate the actual basis functions V_k and the covariance matrix *cov* (V_m, V_n) for each ensemble. Consequent adjustment of the parameters c_k requires us to simply perform matrix and vector multiplication.

Acknowledgments

We thank the present and former members of the Levitt group for help and discussions. We thank Patrice Koehl for providing excellent routines for constructing structures from dihedral angles. B.F. thanks the A.P. Sloan Foundation, and the U.S. Department of Energy for financial support, and Professor J. Rudnick for useful discussions. Y.X. is a Howard Hughes Medical Institute Predoctoral Fellow. This work was supported in part by NIH grant GM-63817 to M.L.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Bahar, I., and Jernigan, R. 1997. Inter-residue potentials in global proteins and the dominance of highly specific hydrophilic interactions at close separations. *J. Mol. Biol.* **266:**195–214.

Bauer, A., and Beyer, A. 1994. An improved pair potential to recognize native protein folds. *Proteins: Struct. Funct. Genet.* **18:** 254–261.

Ben-Naim, A. 1997. Statistical potentials extracted from protein structures: Are these meaningful potentials. *J. Chem. Phys.* **107:** 3698–3706.

Betancourt, M. and Thirumalai, D. 1999. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interactions schemes. *Protein Sci.* **2:** 361–369.

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

Brenner, S., Koehl, P., and Levitt, M. 2000. The astral compendium for sequence and structure analysis. *Nucleic Acid Res.* **28:** 254–256.

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

Chiu, T. and Goldstein, R. 1998. Optimizing energy potentials for success in protein tertiary structure prediction. *Folding and Design* **3:** 223–228.

Dill, K. 1990. Dominant forces in protein folding. *Biochemistry* **29:** 7133–7155.

Fain, B., Xia, Y., and Levitt, M. 2001. Determination of optimal chebyshev-expanded hydrophobic discrimination function for globular proteins. *IBM Systems J., special issue, Deep computing in life sciences,* (in press).

Godzik, A., Kolinski, A., and Skolnick, J. 1995. Are protein ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* **4:** 2107–2117.

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

———. 1992b. Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci.* **89:** 4918–4922.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M. 1990. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216:** 167–180.

Hinds, D. and Levitt, M. 1991. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci.* **89:** 2536–2540.

Huang, E., Subbiah, S., and Levitt, M. 1996. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* **252:** 709–720.

Huang, E., Samudrala, R., and Ponder, J. 1999. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J. Mol. Biol.* **290:** 261–281.

Jones, T., Taylor, W., and Thornton, J. 1992. A new approach to protein fold recognition. *Nature (London)* **358:** 86–89.

Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Advan. Prot. Chem.* **14:** 1–63.

Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104:** 59–107.

———. 1992. Accurate modelling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226:** 507–533.

Levitt, M., Hirshberg, R., and Daggett, V. 1995. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp. Phys. Comm.* **91:** 215–231.

Maiorov, V. and Crippen, G. 1992. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227:** 876–888.

Metropolis, N., Rosenbluth, M., Teller, A., and Teller, E. 1953. Rapidly convergent method for boltzmann-weighted ensemble generation in free energy simulations. *J. Chem. Phys.* **21:**

Mirny, L. and Shakhnovich, E. 1996. How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* **264:** 1164–1179.

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

———. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256:** 623–644.

Murzin, A., Brenner, S., and Hubbard, T. 1995. Scop: A structural classification of proteins database for investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Nelder, J., and Mead, R. 1965. A simplex method for function minimization. *Computer J.* **7:** 308–313.

Ouzounis, C., Sander, C., Sharf, M., and Schneider, R. 1993. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from three-dimensional structure. *J. Mol. Biol.* **232:** 805–825.

Park, B. and Levitt, M. 1996. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* **258:** 367–392.

Pauling, L. 1960. *The nature of the chemical bond and the structure of molecules and crystals: An introduction to modern structural chemistry.* Cornell University Press, Ithaca, New York.

Perutz, M., Kendrew, J., and Watson, H. 1965. Structure and function of hemoglobin. ii. some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **13:** 669–678.

Press, W., Teukolsky, S., Vetterling, W., and Flannery. B. 1992. *Numerical recipes in C,* 2nd ed., Cambridge University Press, New York.

Reva, B., Finkelstein, A., and Skolnik, J. 1998. What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? *Folding and Design* **3:** 141–147.

Rose, G., Geselowitz, A., Lesser, G., Lee, R., and Zehfus, M. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* **229:** 834–838.

Samudrala, R. and Levitt, M. 1999. Decoys'R'Us: A database of incorrect protein conformations for evaluating scoring functions. *Protein Sci.* **9:** 1399–1401.

Samudrala, R. and Moult, J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275:** 895–916.

Samudrala, R., Xia, Y., Huang, E., and Levitt, M. 1999. Bona fide ab initio protein structure prediction using a combined hierarchical approach. *Proteins: Struct. Funct. Genet.* **3:** 194–198.

Seno, F., Maritan, A., and Banavar, J. 1998. Interaction potentials for protein folding. *Proteins: Struct. Funct. Genet.* **30:** 224–248.

Simons, K., Kooperberg, C., Huang, E., and Baker, D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring function. *J. Mol. Biol.* **268:** 209–225.

Simons, K., Ruczinski, I., Kooperberg, C., Fox, B., Bystroff, C., and Baker, D. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Struct. Funct. Genet.* **34:** 82–95.

Sippl, M. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213:** 859–883.

Sippl, F. 1995. Knowledge based potentials for proteins. *Curr. Opin. Struct. Biol.* **5:** 229–235.

Srinivasan, R. and Rose, G. 1995. Linus: A hierarchic procedure to predict the fold of a protein. *Proteins: Struct. Funct. Genet.* **22:** 81–88.

Tanaka, S. and Scheraga, H. 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **9:** 945–950.

Thomas, P. and Dill, K. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* **257:** 457–469.

Toby, D. and Elber, R. 2000. Distance-dependent, pair potential for protein folding: Results from linear optimization. *Proteins: Struct. Funct. Genet.* **41:** 40–46.

Toby, D., Safran, G., Linial, D., and Elber, R. 2000. On the design and analysis of protein folding potentials. *Proteins: Struct. Funct. Genet.* **40:** 71–85.

Van Gunsteren, W. 1989. *Computer simulation of biomolecular systems.* (ed. B.V. Leiden) ESCOM Science Publishers, The Netherlands.

Vendruscolo, M., Mirny, L., Shakhnovitch, E., and Domany, E. 2000. Comparison of two optimization methods to derive energy parameters for protein folding: Perceptron and z-score. *Proteins: Struct. Funct. Genet.* **41:** 192–201.

Viswanadhan, V. 1987. Hydrophobicity and residue-residue contacts in globular proteins. *Int. J. Biol. Macromol.* **9:** 39–48.

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

Xia, Y., Huang, E., Levitt, M., and Samudrala, R. 2000. Ab initio construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* **300:** 171–185.

Xia, Y. and Levitt, M. 2001. Extracting knowledge-based energy functions from protein structures by error rate minimization. Comparison of methods using lattice model. *J. Chem. Phys.* **113:** 9318–9330.