BIOINFORMATICS

Vol. 20 no. 17 2004, pages 3080–3098
doi:10.1093/bioinformatics/bth369

Developing optimal non-linear scoring function for protein design

Changyu Hu, Xiang Li and Jie Liang*

Department of Bioengineering, SEO, MC-063, University of Illinois at Chicago,
851 S. Morgan Street, Room 218, Chicago, IL 60607-7052, USA

Received on February 29, 2004; revised on June 9, 2004; accepted on June 10, 2004
Advance Access publication June 24, 2004

ABSTRACT

Motivation. Protein design aims to identify sequences compatible with a given protein fold but incompatible to any alternative folds. To select the correct sequences and to guide the search process, a design scoring function is critically important. Such a scoring function should be able to characterize the global fitness landscape of many proteins simultaneously.

Results: To find optimal design scoring functions, we introduce two geometric views and propose a formulation using a mixture of non-linear Gaussian kernel functions. We aim to solve a simplified protein sequence design problem. Our goal is to distinguish each native sequence for a major portion of representative protein structures from a large number of alternative decoy sequences, each a fragment from proteins of different folds. Our scoring function discriminates perfectly a set of 440 native proteins from 14 million sequence decoys. We show that no linear scoring function can succeed in this task. In a blind test of unrelated proteins, our scoring function misclassfies only 13 native proteins out of 194. This compares favorably with about three–four times more misclassifications when optimal linear functions reported in the literature are used. We also discuss how to develop protein folding scoring function.

Availability: Available on request from the authors.

Contact: jliangATuicDOTedu

1 INTRODUCTION

The problem of protein sequence design aims to identify sequences compatible with a given protein fold and incompatible with alternative folds (Drexler, 1981; Pabo, 1983; DeGrado *et al.*, 1999). It is also called the inverse protein folding problem. This is a fundamental problem and has attracted considerable interest (Yue and Dill, 1992; Shakhnovich, 1998; Li *et al.*, 1996; Deutsch and Kurosky, 1996; Koehl and Levitt, 1999a,b). The ultimate goal of protein design is to engineer protein molecules with improved activities or with acquired new functions. There have been many important design studies, including the design of novel hydrophobic core (Desjarlais and Handel, 1995; Lazar,G.A. *et al.*, 1997), the design and experimental validation of an entire protein for specified backbone (Dahiyat and Mayo, 1997), the design of a novel alpha helical protein (Emberly *et al.*, 2002), the design and validation of a protein adopting a completely new fold unseen in nature (Kuhlman *et al.*, 2003) and a soluble analog of membrane potassium channel (Slovic *et al.*, 2004).

A successful protein design strategy needs to solve two problems. First, it needs to explore both the sequence and the structure search space and efficiently generates candidate sequences. Second, a scoring function or fitness function needs to identify sequences that are compatible with the desired template fold (the ‘design in’ principle) but are incompatible with any other competing folds (the ‘design out’ principle) (Yue and Dill, 1992; Koehl and Levitt, 1999a,b). To achieve this, an ideal scoring function would maximize the probabilities of protein sequences taking their native fold, and reduce the probability that these sequences take any other fold. Because many protein sequences with low sequence identity can adopt the same protein fold, a full-fledged design scoring function should identify all sequences that fold into the same desired structural fold from a vast number of sequences that do fold into alternative structures, or that do not fold.

Several design scoring functions have been developed based on physical models. For redesigning protein cores, hydrophobicity and packing specificity are the main ingredients of the scoring functions (Desjarlais and Handel, 1995). Van der Waals interactions and electrostatics have also been incorporated for protein design (Koehl and Levitt, 1999a,b). A combination of terms including Lennard–Jones potential, repulsion, Lazaridis–Karplus implicit solvation, approximated electrostatic interactions and hydrogen bonds are used in an insightful computational protein design experiment (Kuhlman and Baker, 2000). Models of solvation energy based on surface area are a key component of several other design scoring functions (Wernisch *et al.*, 2000; Koehl and Levitt, 1999a,b).

A variety of empirical scoring functions based on known protein structures have also been developed for coarse-grained

*To whom correspondence should be addressed.

models of proteins. In this case, proteins are not represented in atomic detail but are represented at residue level. Because of the coarse-grained nature of the protein representation, these scoring functions allow rapid exploration of the search space of the main factors important for proteins, and can provide good initial solutions for further refinement where models with atomistic details can be used.

Many empirical scoring functions were originally developed for the purposes of protein folding and structure prediction. Because the principles are very similar, they are often used directly for protein design. One prominent class of empirical scoring functions are knowledge-based scoring functions, which are derived from statistical analysis of the database of protein structures (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1985; Samudrala and Moult, 1998; Lu and Skolnick, 2001). Here, the interactions between a pair of residues are estimated from its relative frequency in database when compared with a reference state or a null model. This approach has found many successfull applications (Miyazawa and Jernigan, 1996; Samudrala and Moult, 1998; Lu and Skolnick, 2001; Wodak and Rooman, 1993; Sippl, 1995; Lemer *et al.*, 1995; Jernigan and Bahar, 1996; Simons *et al.*, 1999; Li *et al.*, 2003). However, there are several conceptual difficulties with this approach. These include the neglect of chain connectivity in the reference state, and the problematic implicit assumption of Boltzmann distribution (Thomas and Dill, 1996a,b; Ben-Naim, 1997).

An alternative approach for empirical scoring function is to find a set of parameters such that the scoring functions are optimized by some criterion, e.g. maximized score difference between native conformation and a set of alternative (or decoy) conformations (Goldstein *et al.*, 1992; Maiorov and Crippen, 1992; Thomas and Dill, 1996a; Tobi *et al.*, 2000; Vendruscolo and Domany, 1998; Vendruscolo *et al.*, 2000a; Bastolla *et al.*, 2001; Dima *et al.*, 2000; Micheletti *et al.*, 2001). This approach has been shown to be effective in fold recognition, where native structures can be identified from alternative conformations (Micheletti *et al.*, 2001). However, if a large number of native protein structures are to be simultaneously discriminated against a large number of decoy conformations, no such scoring functions can be found (Vendruscolo *et al.*, 2000a; Tobi *et al.*, 2000). A similar conclusion is found in the present study for protein design, where we find that no linear design scoring function can simultaneously discriminate a large number of native proteins from sequence decoys. A recent criticism for contract potential is that it is impossible to predict stability changes due to mutation using contact-based scoring function (Khatun *et al.*, 2004).

There are three key steps in developing effective empirical scoring function using optimization: (1) the functional form, (2) the generation of a large set of decoys for discrimination and (3) the optimization techniques. The initial step of choosing an appropriate functional form is often straightforward. Empirical pairwise scoring functions are usually all in the form of weighted linear sum of interacting residue pairs (for an exception, see Fain *et al.*, 2002). In this functional form, the weight coefficients are the parameters of the scoring function, which are optimized for discrimination. The same functional form is also used in statistical potential, where the weight coefficients are derived from database statistics. The optimization techniques that have been used include perceptron learning and linear programming (Tobi *et al.*, 2000; Vendruscolo *et al.*, 2000a). The objectives of optimization are often maximization of score gap between native protein and the average of decoys, or score gap between native and decoys with the lowest score, or the Z-score of the native protein (Goldstein *et al.*, 1992; Koretke *et al.*, 1996, 1998; Hao and Scheraga, 1996; Mirny and Shakhnovich, 1996).

In this work, we study a simplified version of the protein design problem. Our goal is to develop a globally applicable scoring function for characterizng the fitness landscape of many proteins simultaneously. Specifically, we aim to identify a protein sequence that is compatible with a given three-dimensional coarse-grained structure from a set of protein sequences that are taken from protein structures of different folds. In conclusion, we discuss how to proceed to develop a full-fledged fitness function that discriminates similar and dissimilar sequences adopting the same fold against all sequences that adopt different folds and sequences that do not fold (e.g. all hydrophobes). In this study, we do not address the problem of how to generate candidate template fold or candidate sequence by searching either the conformation space or the sequence space.

To develop an empirical scoring function that improves discrimination of native protein sequence, we explore in this study an alternative formulation of protein scoring function, in the form of mixture of non-linear Gaussian kernel functions. We also use a different optimization technique based on quadratic programming. Instead of maximizing the score gap, here an objective function related to bounds of expected classification errors is optimized (Vapnik and Chervonenkis, 1974; Vapnik, 1995; Burges, 1998; Schölkopf and Smola, 2002).

Experimentation with the non-linear function developed in this study shows that it can discriminate simultaneously 440 native proteins against 14 million sequence decoys. In contrast, we cannot obtain a perfect weighted linear sum scoring function using the state-of-the-art interior point solver of linear programming following Tobi *et al.* (2000) and Meller *et al.* (2002). We also perform blind tests for native sequence recognition. Taking 194 proteins unrelated to the 440 training set proteins, the non-linear scoring function achieves a success rate of 93.3% in sequence design. This result compares favorably with optimal linear scoring function (80.9 and 73.7% success rate) and statistical potential (58.2%) (Tobi *et al.*, 2000; Bastolla *et al.*, 2001; Miyazawa and Jernigan, 1996).

The rest of the paper is organized as follows. We first describe the theory and model of linear and non-linear functions, including the kernel model and the optimization technique. We then explain details of computation. We further describe experimental results of learning and results of blind test. We conclude with a discussion on how these ideas may be applicable for developing protein folding scoring function.

2 THEORY AND MODELS

Modeling protein design scoring function. To model a protein computationally, we first need a method to describe its geometric shape and its sequence of amino acid residues. Frequently, a protein is represented by a d-dimensional vector $\boldsymbol{c} \in \mathbb{R}^d$. For example, a method that is widely used is to count non-bonded contacts of various types of amino acid residue pairs in a protein structure. In this case, the count vector $\boldsymbol{c} \in \mathbb{R}^d, d = 210$, is used as the protein descriptor. Once the structural conformation of a protein $\boldsymbol{s}$ and its amino acid sequence $\boldsymbol{a}$ is given, the protein description $f : (\boldsymbol{s}, \boldsymbol{a}) \mapsto \mathbb{R}^d$ will fully determine the d-dimensional vector $\boldsymbol{c}$. In the case of contact vector, f corresponds to the mapping provided by specific contact definition, e.g. two residues are in contact if their distance is below a specific cutoff threshold distance.

To develop scoring functions for our simplified problem, namely, a scoring function that allows the search and identification of sequences most compatible with a specific given coarse-grained three-dimensional structure, we use a model analogous to the Anfinsen experiments in protein folding. We require that the native amino acid sequence $\boldsymbol{a}_N$ mounted on the native structure $\boldsymbol{s}_N$ has the best (lowest) fitness score compared to a set of alternative sequences (sequence decoys) taken from unrelated proteins known to fold into a different fold $\mathcal{D} = \{\boldsymbol{s}_N, \boldsymbol{a}_D\}$ when mounted on the same native protein structure $\boldsymbol{s}_N$:

$$H(f(\boldsymbol{s}_N, \boldsymbol{a}_N)) < H(f(\boldsymbol{s}_N, \boldsymbol{a}_D)) \quad \text{for all } \boldsymbol{s}_N, \boldsymbol{a}_D \in \mathcal{D}.$$

Equivalently, the native sequence will have the highest probability to fit into the specified native structure. This is the same principle described by Shakhnovich and Gutin (1993), Deutsch and Kurosky (1996), Li *et al.* (1996). Sometimes we can further require that the score difference must be greater than a constant $b > 0$:

$$H(f(\boldsymbol{s}_N, \boldsymbol{a}_N)) + b < H(f(\boldsymbol{s}_N, \boldsymbol{a}_D)) \quad \text{for all } (\boldsymbol{s}_N, \boldsymbol{a}_D) \in \mathcal{D}.$$

A widely used functional form for protein scoring function H is the weighted linear sum of pairwise contacts (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1985; Tobi *et al.*, 2000; Vendruscolo and Domany, 1998; Samudrala and Moult, 1998; Lu and Skolnick, 2001). The linear sum score H is:

$$H(f(\boldsymbol{s}, \boldsymbol{a})) = H(\boldsymbol{c}) = \boldsymbol{w} \cdot \boldsymbol{c}, \tag{1}$$

where '·' denotes the inner product of vectors. As soon as the weight vector $\boldsymbol{w}$ is specified, the scoring function is fully defined. Much work has been done using this class of design function of linear sum of contact pairs (Shakhnovich and Gutin, 1993; Deutsch and Kurosky, 1996). For such linear scoring functions, the basic requirement for design scoring function is then:

$$\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) < 0,$$

or

$$\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0, \tag{2}$$

if we require that the score difference between a native protein and a decoy must be greater than a real positive value b. The goal here is to obtain a scoring function to discriminate native proteins from decoys. An ideal scoring function therefore would assign the value '-1' for native structure/sequence and the value '$+1$' for decoys.

Two geometric views of linear protein design scoring function. There is a natural geometric view of the inequality requirement for weighted linear sum scoring functions. A useful observation is that each of the inequalities divides the space of $\mathbb{R}^d$ into two halves separated by a hyperplane (Fig. 1a). The hyperplane for Equation (2) is defined by the normal vector $(\boldsymbol{c}_N - \boldsymbol{c}_D)$ and its distance $b/||\boldsymbol{c}_N - \boldsymbol{c}_D||$ from the origin. The weight vector $\boldsymbol{w}$ must be located in the half-space opposite the direction of the normal vector $(\boldsymbol{c}_N - \boldsymbol{c}_D)$. This half-space can be written as $\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0$. When there are many inequalities to be satisfied simultaneously, the intersection of the half-spaces forms a convex polyhedron (Edelsbrunner, 1987). If the weight vector is located in the polyhedron, all the inequalities are satisfied. Scoring functions with such weight vector $\boldsymbol{w}$ can discriminate the native protein sequence from the set of all decoys. This is illustrated in Figure 1a for a two-dimensional toy example, where each straight line represents an inequality $\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0$ that the scoring function must satisfy.

For each native protein i, there is one convex polyhedron $\mathcal{P}_i$ formed by the set of inequalities associated with its decoys. If a scoring function can discriminate simultaneously n native proteins from a union of sets of sequence decoys, the weight vector $\boldsymbol{w}$ must be located in a smaller convex polyhedron $\mathcal{P}$ that is the intersection of the n convex polyhedra:

$$\boldsymbol{w} \in \mathcal{P} = \bigcap_{i=1}^{n} \mathcal{P}_i.$$

There is yet another geometric view of the same inequality requirements. If we now regard $(\boldsymbol{c}_N - \boldsymbol{c}_D)$ as a point in $\mathbb{R}^d$, the relationship $\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0$ for all sequence decoys and

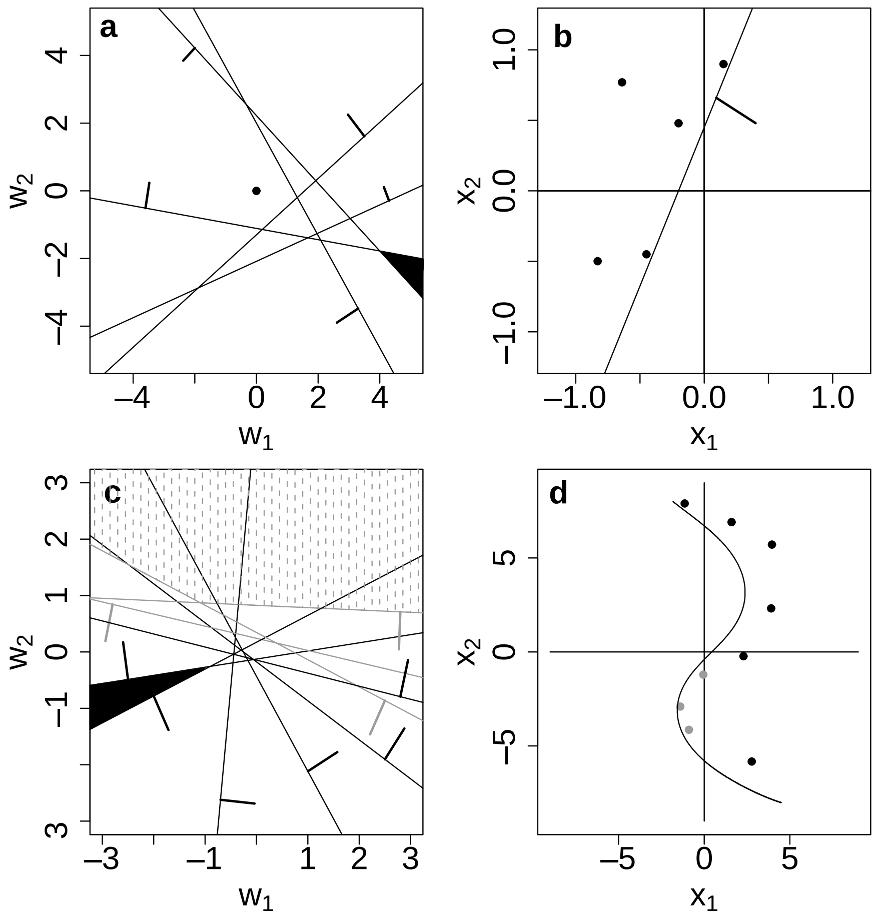

Fig. 1. Geometric views of the inequality requirement for protein scoring function. Here, we use a two-dimensional toy example for illustration. (**a**) In the first geometric view, the space $\mathbb{R}^2$ of $\boldsymbol{w} = (w_1, w_2)$ is divided into two half-spaces by an inequality requirement, represented as a hyperplane $\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0$. The hyperplane, which is a line in $\mathbb{R}^2$, is defined by the normal vector $(\boldsymbol{c}_N - \boldsymbol{c}_D)$, and its distance $b/||\boldsymbol{c}_N - \boldsymbol{c}_D||$ from the origin. In this figure, this distance is set to 1.0. The normal vector is represented by a short line segment whose direction points away from the straight line. A feasible weight vector $\boldsymbol{w}$ is located in the half-space opposite the direction of the normal vector $(\boldsymbol{c}_N - \boldsymbol{c}_D)$. With the given set of inequalities represented by the lines, any weight vector $\boldsymbol{w}$ located in the shaped polygon can satisfy all inequality requirement and provide a linear scoring function that has perfect discrimination. (**b**) A second geometric view of the inequality requirement for linear protein scoring function. The space $\mathbb{R}^2$ of $\boldsymbol{x} = (x_1, x_2)$, where $\boldsymbol{x} \equiv (\boldsymbol{c}_N - \boldsymbol{c}_D)$, is divided into two half-spaces by the hyperplane $\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0$. Here, the hyperplane is defined by the normal vector $\boldsymbol{w}$ and its distance $b/||\boldsymbol{w}||$ from the origin. All points $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ are located on one side of the hyperplane away from the origin, therefore satisfying the inequality requirement. That is, a linear scoring function $\boldsymbol{w}$ such as the one represented by the straight line in this figure can have perfect discrimination. (**c**) In the second toy problem, a set of inequalities are represented by a set of straight lines according to the first geometric view. A subset of the inequalities require that the weight vector $\boldsymbol{w}$ be located in the shaded convex polygon on the left, but another subset of inequalities require $\boldsymbol{w}$ to be located in the dashed convex polygon on the top. Since these two polygons do not intersect, there is no weight vector $\boldsymbol{w}$ that can satisfy all inequality requirements. That is, no linear scoring function can classify these decoys from native protein. (**d**) According to the second geometric view, no hyperplane can separate all points $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ from the origin. But a non-linear curve formed by a mixture of Gaussian kernels can have perfect separation of all vectors $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ from the origin: it has perfect discrimination.

native proteins requires that all points $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ are located on one side of a different hyperplane, which is defined by its normal vector $\boldsymbol{w}$ and its distance $b/||\boldsymbol{w}||$ to the origin (Fig. 1b). We can show that such a hyperplane exists if the origin is not contained within the convex hull of the set of points $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ (see Appendix section).

The second geometric view looks very different from the first view. However, the second view is dual and mathematically equivalent to the first geometric view. In the first view, a point $\boldsymbol{c}_N - \boldsymbol{c}_D$ determined by the structure–decoy pair $c_N = (\boldsymbol{s}_N, \boldsymbol{a}_N)$ and $c_D = (\boldsymbol{s}_N, \boldsymbol{a}_D)$ corresponds to a hyperplane representing an inequality, a solution weight vector $\boldsymbol{w}$

corresponds to a point located in the final convex polyhedron. In the second view, each structure–decoy pair is represented as a point $\boldsymbol{c}_N - \boldsymbol{c}_D$ in $\mathbb{R}^d$, and the solution weight vector $\boldsymbol{w}$ is represented by a hyperplane separating all the points $\mathcal{C} = \{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ from the origin.

Optimal linear scoring function. Several optimization methods have been applied to find the weight vector $\boldsymbol{w}$ of linear scoring function. The Rosenblantt perceptron method works by iteratively updating an initial weight vector $\boldsymbol{w}_0$ (Vendruscolo and Domany, 1998; Micheletti *et al.*, 2001). Starting with a random vector, e.g. $\boldsymbol{w}_0 = \boldsymbol{0}$, one tests each native protein and its decoy structure. Whenever the relationship $\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0$ is violated, one updates $\boldsymbol{w}$ by adding to it a scaled violating vector $\eta \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D)$. The final weight vector is therefore a linear combination of protein and decoy count vectors:

$$\boldsymbol{w} = \sum \eta(\boldsymbol{c}_N - \boldsymbol{c}_D) = \sum_{N \in \mathcal{N}} \alpha_N \boldsymbol{c}_N - \sum_{D \in \mathcal{D}} \alpha_D \boldsymbol{c}_D. \qquad (3)$$

Here, $\mathcal{N}$ is the set of native proteins, and $\mathcal{D}$ is the set of decoys. The set of coefficients $\{\alpha_N\} \cup \{\alpha_D\}$ gives a dual form representation of the weight vector $\boldsymbol{w}$, which is an expansion of the training examples including both native and decoy structures.

According to the first geometric view, if the final convex polyhedron $\mathcal{P}$ is non-empty, there can be an infinite number of choices of $\boldsymbol{w}$, all with perfect discrimination. But how do we find a weight vector $\boldsymbol{w}$ that is optimal? This depends on the criterion for optimality. For example, one can choose the weight vector $\boldsymbol{w}$ that minimizes the variance of score gaps between decoys and natives: $\arg_{\boldsymbol{w}} \min \frac{1}{|\mathcal{D}|} \sum [\boldsymbol{w} \cdot (c_N - c_D)]^2 - \left[\frac{1}{|\mathcal{D}|} \sum_D (\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D))\right]^2$ as used by Tobi *et al.* (2000), minimizes the Z-score of a large set of native proteins, minimizes the Z-score of the native protein and an ensemble of decoys (Chiu and Goldstein, 1998; Mirny and Shakhnovich, 1996) or maximize the ratio R between the width of the distribution of the score and the average score difference between the native state and the unfolded ones (Goldstein *et al.*, 1992; Hao and Scheraga, 1999). A series of important works using perceptron learning and other optimization techniques (Friedrichs and Wolynes, 1989; Goldstein *et al.*, 1992; Tobi *et al.*, 2000; Vendruscolo and Domany, 1998; Dima *et al.*, 2000) showed that effective linear sum scoring functions can be obtained.

Here, we describe yet another optimality criterion according to the second geometric view. We can choose the hyperplane $(\boldsymbol{w}, b)$ that separates the points $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ with the largest distance to the origin. Intuitively, we want to characterize proteins with a region defined by the training set points $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$. It is desirable to define this region such that a new unseen point drawn from the same protein distribution as $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ will have a high probability to fall within the defined region. Non-protein points following a different distribution, which is assumed to be centered around the origin when no a priori information is available, will have a high probability to fall outside the defined region. In this case, we are more interested in modeling the region or support of the distribution of protein data, rather than estimating its density distribution function. For linear scoring function, regions are half-spaces defined by hyperplanes, and the optimal hyperplane $(\boldsymbol{w}, b)$ is then the one with maximal distance to the origin. This is related to the novelty detection problem and single-class support vector machine studied in statistical learning theory (Vapnik and Chervonenkis, 1964; Vapnik and Chervonenkis, 1974; Schölkopf and Smola, 2002). In our case, any non-protein points will need to be detected as outliers from the protein distribution characterized by $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$. Among all linear functions derived from the same set of native proteins and decoys, an optimal weight vector $\boldsymbol{w}$ is likely to have the least amount of mislabelings. The optimal weight vector $\boldsymbol{w}$ can be found by solving the following quadratic programming problem:

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}||\boldsymbol{w}||^2 \\ \text{Subject to} & \boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0 \\ & \quad \text{for all } N \in \mathcal{N} \text{ and } D \in \mathcal{D}. \end{array} \qquad (4)$$

The solution maximizes the distance $b/||\boldsymbol{w}||$ of the plane $(\boldsymbol{w}, b)$ to the origin. We obtained the solution by solving the following support vector machine problem:

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{Subject to} & \boldsymbol{w} \cdot \boldsymbol{c}_N + d \le -1 \\ & \boldsymbol{w} \cdot \boldsymbol{c}_D + d \ge 1, \end{array} \qquad (5)$$

where $d > 0$. Note that a solution of Problem (5) satisfies the constraints in Inequalities (4), since subtracting the second inequality here from the first inequality in the constraint conditions of (5) will give us $\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + 2 \le 0$.

Non-linear scoring function. However, it is possible that the weight vector $\boldsymbol{w}$ does not exist, i.e. the final convex polyhedron $\mathcal{P} = \bigcap_{i=1}^{n} \mathcal{P}_i$ may be an empty set. First, for a specific native protein i, there may be severe restriction from some inequality constraints, which makes $\mathcal{P}_i$ an empty set. Some decoys are very difficult to discriminate due to perhaps deficiency in protein representation. In these cases, it is impossible to adjust the weight vector so the native protein has a lower score than the sequence decoy. Figure 1c shows a set of inequalities represented by straight lines according to the first geometric view. A subset of inequalities (black lines) require that the weight vector $\boldsymbol{w}$ to be located in the shaded convex polygon on the left, but another subset of inequalities (green lines) require $\boldsymbol{w}$ to be located in the dashed convex polygon on the top. Since these two polygons do not intersect, there is no weight vector that can satisfy all these inequality requirements. That is, no linear scoring function can classify all decoys from native protein. According to the second geometric view (Fig. 1d), no hyperplane can separate all points (black and green) $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ from the origin.

Second, even if a weight vector $\boldsymbol{w}$ can be found for each native protein, i.e. $\boldsymbol{w}$ is contained in a non-empty polyhedron, it is still possible that the intersection of n polyhedra is an empty set, i.e. no weight vector can be found that can discriminate all native proteins against the decoys simultaneously. Computationally, the question whether a solution weight vector $\boldsymbol{w}$ exists can be answered unambiguously in polynomial time (Karmarkar, 1984), and results described later in this study show that when the number of decoys reaches millions, no such weight vector can be found.

A fundamental reason for this failure is that the functional form of linear sum is too simplistic. It has been suggested that additional decriptors of protein structures such as higher-order interactions (e.g. three-body or four-body contacts) should be incorporated in protein description (Betancourt and Thirumalai, 1999; Munson and Singh, 1997; Zheng *et al.*, 1997). Functions with polynomial terms using up to six degrees of Chebyshev expansion have also been used to represent pairwise interactions in protein folding (Fain *et al.*, 2002).

Here, we propose an alternative approach. In this study we still limit ourselves to pairwise contact interactions, although it can be naturally extended to include three or four body interactions (Li and Liang, 2004). We introduce a non-linear scoring function analogous to the dual form of the linear function in Equation (3), which takes the following form:

$$H[f(\boldsymbol{s},\boldsymbol{a})] = H(\boldsymbol{c}) = \sum_{D\in\mathcal{D}} \alpha_D K(\boldsymbol{c},\boldsymbol{c}_D) - \sum_{N\in\mathcal{N}} \alpha_N K(\boldsymbol{c},\boldsymbol{c}_N), \tag{6}$$

where $\alpha_D \geq 0$ and $\alpha_N \geq 0$ are parameters of the scoring function to be determined, $\boldsymbol{c}_D = f(\boldsymbol{s}_N, \boldsymbol{a}_D)$ from the set of decoys $\mathcal{D} = \{(\boldsymbol{s}_N, \boldsymbol{a}_D)\}$ is the contact vector of a sequence decoy D mounted on a native protein structure $\boldsymbol{s}_N$, and $\boldsymbol{c}_N = f(\boldsymbol{s}_N, \boldsymbol{a}_N)$ from the set of native training proteins $\mathcal{N} = \{(\boldsymbol{s}_N, \boldsymbol{a}_N)\}$ is the contact vector of a native sequence $\boldsymbol{a}_N$ mounted on its native structure $\boldsymbol{s}_N$. In this study, all decoy sequences $\{\boldsymbol{a}_D\}$ are taken from real proteins possessing different fold structures. The difference of this functional form from the linear function in Equation (3) is that a kernel function $K(\boldsymbol{x}, \boldsymbol{y})$ replaces the linear term. A convenient kernel function K is:

$$K(\boldsymbol{x},\boldsymbol{y}) = \mathrm{e}^{-||\boldsymbol{x}-\boldsymbol{y}||^2/2\sigma^2} \text{ for any vectors } \boldsymbol{x} \text{ and } \boldsymbol{y} \in \mathcal{N} \bigcup \mathcal{D},$$

where σ^2 is a constant. Intuitively, the surface of the scoring function has smooth Gaussian hills of height α_D centered on the location $\boldsymbol{c}_D$ of decoy protein D, and has smooth Gaussian cones of depth α_N centered on the location $\boldsymbol{c}_N$ of native structures N. Ideally, the value of the scoring function will be -1 for contact vectors $\boldsymbol{c}_N$ of native proteins and $+1$ for contact vectors $\boldsymbol{c}_D$ of decoys.

Optimal non-linear scoring function. To obtain the non-linear scoring function, our goal is to find a set of parameters $\{\alpha_D, \alpha_N\}$ such that $H[f(\boldsymbol{s}_N, \boldsymbol{a}_N)]$ has a value close to -1 for native proteins, and the decoys have values close to $+1$. There are many different choices of $\{\alpha_D, \alpha_N\}$. We use an optimality criterion originally developed in statistical learning theory (Vapnik, 1995; Burges, 1998; Schölkopf and Smola, 2002). First, we note that we have implicitly mapped each structure and decoy from $\mathbb{R}^{210}$ through the kernel function of $K(\boldsymbol{x},\boldsymbol{y}) = \mathrm{e}^{-||\boldsymbol{x}-\boldsymbol{y}||^2/2\sigma^2}$ to another space with dimension as high as tens of millions. Second, we then find the hyperplane of the largest margin distance separating proteins and decoys in the space transformed by the non-linear kernel. That is, we search for a hyperplane with equal and maximal distance to the closest native proteins and the closest decoys in the transformed high-dimensional space. Such a hyperplane can be found by obtaining the parameters $\{\alpha_D\}$ and $\{\alpha_N\}$ from solving the following Lagrange dual form of quadratic programming problem:

$$\begin{array}{ll} \text{Maximize} & \sum_{i\in\mathcal{N}\cup\mathcal{D},}\alpha_i \\ & -\frac{1}{2}\sum_{i,j\in\mathcal{N}\cup\mathcal{D}} y_i y_j \alpha_i \alpha_j e^{-||\boldsymbol{c}_i-\boldsymbol{c}_j||^2/2\sigma^2} \\ \text{Subject to} & 0 \leq \alpha_i \leq C, \end{array}$$

where C is a regularizing constant that limits the influence of each misclassified protein or decoy (Vapnik and Chervonenkis, 1964; Vapnik and Chervonenkis, 1974; Vapnik, 1995; Burges, 1998; Schölkopf and Smola, 2002), $y_i = -1$ if i is a native protein and $y_i = +1$ if i is a decoy. These parameters lead to optimal discrimination of an unseen test set (Vapnik and Chervonenkis, 1964; Vapnik and Chervonenkis, 1974; Vapnik, 1995; Burges, 1998; Schölkopf and Smola, 2002). When projected back to the space of $\mathbb{R}^{210}$, this hyperplane becomes a non-linear surface. For the toy problem of Figure 1, Figure 1d shows that such a hyperplane becomes a non-linear curve in $\mathbb{R}^2$ formed by a mixture of Gaussian kernels. It separates perfectly all vectors $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ (black and green) from the origin. That is, a non-linear scoring function can have perfect discrimination.

3 COMPUTATIONAL METHODS

Alpha contact maps. Because protein molecules are formed by thousands of atoms, their shapes are complex. In this study we use the count vector of pairwise contact interactions after normalization by the chain length of the protein (Edelsbrunner, 1995; Liang *et al.*, 1998). Here, contacts are derived from the edge simplices of the alpha shape of a protein structure (Li *et al.*, 2003). These edge simplices represent the nearest neighbor interactions that are in physical contact. They encode precisely the same contact information as a subset of the edges in the Voronoi diagram of the protein molecule. These Voronoi edges are shared by two interacting atoms from different residues, but intersect with the body of the molecule modeled as the union of atom balls. Statistical potential based

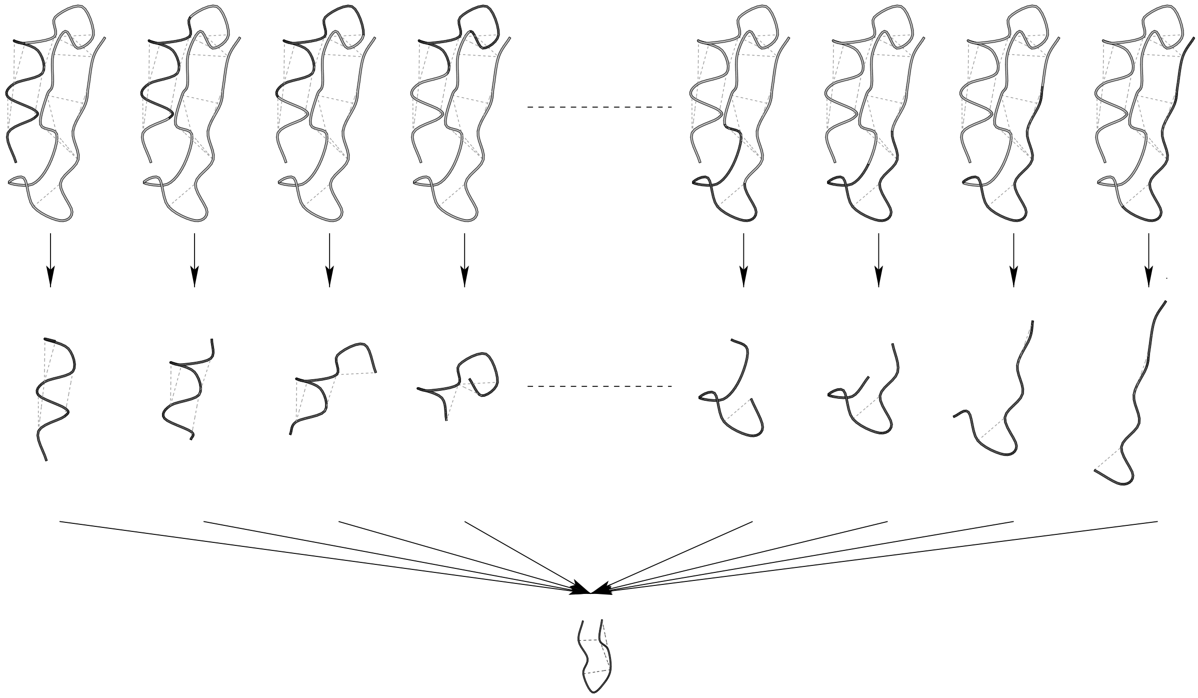

Fig. 2. Decoy generation by gapless threading. Sequence decoys can be generated by threading the sequence of a larger protein to the structure of an unrelated smaller protein.

on edge simplices has been developed (Li *et al.*, 2003). We refer to Edelsbrunner (1995) and Liang *et al.* (1998) for further theoretical and computational details.

Generating sequence decoys by threading. Maiorov and Crippen introduced the gapless threading method to generate a large number of decoys (1992). The sequence of a smaller protein $\boldsymbol{a}_N$ is threaded through the structure of an unrelated larger protein and takes the conformation $\boldsymbol{s}_D$ of a fragment with the same length from the larger protein (Maiorov and Crippen, 1992). Along the way, the sequence of the smaller protein can take the conformations of many fragments of the larger protein, each becoming structure decoy.

We can generate sequence decoys in an analogous way, as already suggested by Jones *et al.* (1992) and Munson and Singh (1997). We thread the sequence of a larger protein through the structure of a smaller protein, and obtain sequence decoys by mounting a fragment of the sequence of the larger protein to the full structure of the smaller protein. We therefore have for each native protein $(\boldsymbol{s}_N, \boldsymbol{a}_N)$ a set of sequence decoys $(\boldsymbol{s}_N, \boldsymbol{a}_D)$ (Fig. 2). Because all native contacts are retained in this case, sequence decoys obtained by gapless threading are far more challenging than structure decoys generated by gapless threading.

Protein data. Following Vendruscolo *et al.* (2000b), we use protein structures contained in the WHATIF database (Vriend and Sander, 1993) in this study. WHATIF database contains a representative set of sequence-unique protein structures generated from X-ray crystallography. Structures selected for this study all have pairwise sequence identity <30%, R-factor <0.21 and resolution <2.1 Å. WHATIF database contains less structures than PDBSELECT because the R-factor and resolution criteria are more stringent (Vriend and Sander, 1993). Nevertheless, it provides a good representative set of all currently known protein structures.

We use a list of 456 proteins kindly provided by Dr Vendruscolo, which was compiled from the 1998 release (WHATIF98) of the WHATIF database (Vendruscolo *et al.*, 2000a). There are 192 proteins with multiple chains in this data set. Some of them have extensive interchain contacts. For these proteins, it is possible that their conformations may be different if there are no interchain contacts present. We use the criterion of contact ratio to remove proteins that have extensive interchain contacts. Contact ratio is defined here as the number of interchain contacts divided by the total number of contacts a chain makes. For example, protein 1ept has four chains A, B, C and D. The intra chain contact number of chain B are 397. Contacts between chain A and chain B are 178, between B and C they are 220 and between B and other heteroatoms, 11. The Contact ratio of chain B is therefore $(178 + 220 + 11)/(397 + 178 + 220 + 11) = 51\%$. Thirteen protein chains are removed because they all have Contact ratio >30%. We further remove three proteins because each has >10% of residues missing with no coordinates in the Protein Data Bank file. The remaining set of 440 proteins are then used as a training set for developing the design scoring functions. Using the threading method described earlier, we generated a set of 14 080 766 sequence decoys.

Learning linear scoring function. For comparison, we have also developed optimal linear scoring function following the method and computational procedure described by Tobi *et al.* (2000). We apply the interior point method as implemented in BPMD package by Mészáros (1996) to search for a weight vector $\boldsymbol{w}$. We use two different optimization criteria as described by Tobi *et al.* (2000). The first is:

$$\begin{array}{ll} \text{Identify} & \boldsymbol{w} \\ \text{Subject to} & \boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) < \epsilon \quad \text{and} \quad |w_i| \leq 10, \end{array}$$

where w_i denotes the i-th component of weight vector $\boldsymbol{w}$, and $\epsilon = 1 \times 10^{-6}$. Let $\mathcal{C} = \{c_N - c_D\}$, and $|\mathcal{C}|$ the number of decoys. The second optimization criterion is:

$$\begin{array}{ll} \text{Minimize} & \min \frac{1}{|\mathcal{C}|} \sum (\boldsymbol{w} \cdot (c_N - c_D))^2 \\ & \quad - \left[\frac{1}{|\mathcal{C}|} \sum (\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D)) \right]^2 \\ \text{Subject to} & \boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) < \epsilon. \end{array}$$

Learning non-linear kernel scoring function. We use SVMLIGHT (http://svmlight.joachims.org/) (Joachims, 1999) with Gaussian kernels and a training set of 440 native proteins plus 14 080 766 decoys to obtain the optimized parameter $\{\alpha_N, \alpha_D\}$. The regularization constant C takes a default value, which is estimated from the training set $\mathcal{N} \cup \mathcal{D}$:

$$C = |\mathcal{N} \cup \mathcal{D}|^2 \Bigg/ \left[\sum_{\boldsymbol{x} \in \mathcal{N} \cup \mathcal{D}} \sqrt{K(\boldsymbol{x}, \boldsymbol{x}) - 2 \cdot K(\boldsymbol{x}, \boldsymbol{0}) + K(\boldsymbol{0}, \boldsymbol{0})} \right]^2 . \tag{7}$$

Since we cannot load all 14 millions decoys into computer memory simultaneously, we use a heuristic strategy for training. Similar to the procedure reported by Tobi *et al.* (2000), we first randomly selected a subset of decoys that fits into the computer memory. Specifically, we pick every 51st decoy from the list of 14 million decoys. This leads to an initial training set of 276 095 decoys and 440 native proteins. An initial protein scoring function is then obtained. Next, the scores for all 14 million decoys and all 440 native proteins are evaluated. Three decoy sets were collected based on the evaluation results: the first set of decoys contains the violating decoys which have lower scores than the native structures; the second set contains decoys with the lowest absolute score and the third set contains decoys that participate in $H(\boldsymbol{c})$ as identified in the previous training process. The union of these three subsets of decoys is then combined with the 440 native proteins as the training set for the next iteration of learning. This process is repeated until the score difference to native protein for all decoys is greater than 0.0. Using this strategy, the number of iterations typically is between 2 and 10. During the training process, we set the cost factor j in SVMLIGHT to 120, which is the factor in which the training errors on native proteins outweighs the training errors on decoys.

The value of σ^2 for the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-||\boldsymbol{x}-\boldsymbol{y}||^2/2\sigma^2}$ is chosen by experimentation. If the value of σ^2 is too large, no parameter set $\{\alpha_N, \alpha_D\}$ can be found such that the fitness scoring function can perfectly classify the 440 training proteins and their decoys, i.e. the problem is unlearnable. If the value of σ^2 is too small, the performance in blind test will deteriorate. The final design scoring function is obtained with σ^2 set to 416.7.

4 RESULTS

Linear design scoring functions. To search for the optimal weight vector $\boldsymbol{w}$ for design scoring function, we use linear programming solver based on interior point method as implemented in BPMD by Mészáros (1996). After generating 14 080 766 sequence design decoys for the 440 proteins in the training set, we search for an optimal $\boldsymbol{w}$ that can discriminate native sequences from decoy sequences. That is, we search for parameters $\boldsymbol{w}$ for $H(\boldsymbol{s}, \boldsymbol{a}) = \boldsymbol{w} \cdot \boldsymbol{c}$, such that $\boldsymbol{w} \cdot \boldsymbol{c}_N < \boldsymbol{w} \cdot \boldsymbol{c}_D$ for all sequences. However, we fail to find a feasible solution for the weight vector $\boldsymbol{w}$. That is, no $\boldsymbol{w}$ exists capable of discriminating perfectly 440 native sequences from the 14 million decoy sequences. We repeat the same experiment using a larger set of 572 native proteins from Tobi *et al.* (2000) and 28 261 307 sequence decoys. The result is also negative.

Non-linear kernel scoring function. To overcome the problems associated with linear function, we use the set of 440 native proteins and 14 million decoys to derive non-linear kernel design functions. We succeed in finding a function in the form of Equation (6) that can discriminate all 440 native proteins from 14 million decoys.

Unlike statistical scoring functions where each native protein in the database contributes to the empirical scoring function, only a subset of native proteins contribute and have $\alpha_N \neq 0$. In addition, a small fraction of decoys also contribute to the scoring function. Table 1 lists the details of the scoring function, including the numbers of native proteins and decoys that participate in Equation (6). These numbers represent $\sim$ 50% of native proteins and <0.1% of decoys from the original training data.

Discrimination tests for design scoring function. Blind test in discriminating native proteins from decoys for an independent test set is essential to assess the effectiveness of design scoring functions. To construct a test set, we first take the entries in WHATIF99 database that are not present in WHATIF98. After eliminating proteins with chain length less than 46 residues, we obtain a set of 201 proteins. These

Table 1. Derivation of kernel scoring function

		Design scoring function $\sigma^2 = 416.7$	Folding scoring function $\sigma^2 = 227.3$
Number of vectors	Natives	220	214
	Decoys	1685	1362
Range of score values	Natives	0.9992–4.598	0.9990–4.215
	Decoys	−9.714–0.7423	−6.859–0.3351
Range of smallest score gap		0.2575–11.53	0.8446–9.816

Details of derivation of non-linear kernel design scoring functions. The numbers of native proteins and decoys with non-zero α_i entering the scoring function are listed. The range of the score values of natives and decoys are also listed, as well as the range of the smallest gaps between the scores of the native protein and decoy. Details for non-linear kernel folding scoring functon are also listed.

Table 2. Blind discrimination test of protein sequence design

	Misclassified natives	Misclassified natives
Kernel design scoring function (KDF)	13/194	19/201
Tobi and Elber	37/194	44/201
Bastolla *et al.*	51/194	54/201
Miyazawa and Jernigan	81/194	87/201

The number of misclassified protein sequences for the test set of 194 proteins and the set of 201 proteins using non-linear kernel design scoring function, two optimal linear scoring functions taken as reported in Tobi *et al.* (2000) and in table I of Bastolla *et al.* (2001) and Miyazawa–Jernigan statistical potential (Miyazawa and Jernigan, 1996). The non-linear kernel design scoring function has the best performance in blind test and is the only function that succeeded in perfect discrimination of the 440 native sequences from a set of 14 million sequence decoys.

proteins all have <30% sequence identities with any other sequence in either the training set or the test set proteins. Since 139 of the 201 test proteins have multiple chains, we use the same criteria applied in training set selection to exclude seven proteins with >30% Contact Ratio or with >10% residues missing coordinates in the PDB files. This leaves a smaller set of test proteins of 194 proteins. Using gapless threading, we generate a set of 3 096 019 sequence decoys from the set of 201 proteins. This is a superset of the decoy set generated using 194 proteins.

To test design scoring functions for discriminating native proteins from sequence decoys in both the 194 and the 201 test sets, we take the sequence $\boldsymbol{a}$ from the conformation–sequence pair $(\boldsymbol{s}_N, \boldsymbol{a})$ for a protein with the lowest score as the predicted sequence. If it is not the native sequence $\boldsymbol{a}_N$, the discrimination failed and the design scoring function does not work for this protein.

For comparison, we also test the discrimination results of optimal linear scoring function taken as reported in Tobi *et al.* (2000), as well as the statistical potential developed by Miyazawa and Jernigan. Here we use the contact definition reported in Tobi *et al.* (2000), i.e. two residues are declared to be in contact if the geometric centers of their side chains are within a distance of 2.0–6.4 Å.

The non-linear design scoring function capable of discriminating all of the 440 native sequences also works well for the test set (Table 2). It succeeded in correctly identifying 93.3% (181 out of 194) of native sequences in the independent test set of 194 proteins. This compares favorably with results obtained using optimal linear folding scoring function taken as reported in Tobi *et al.* (2000), which succeeded in identifying 80.9% (157 out of 194) of this test set. It also has better performance than optimal linear scoring function based on calculations using parameters reported in Bastolla *et al.* (2001), which succeeded in identifying 73.7% (143 out of 194) of proteins in the test set. The Miyazawa–Jernigan statistical potential succeeded in identifying 113 native proteins out of 194 (success rate 58.2%).

Discriminating dissimilar proteins. As any other discrimination problems, the success of our classification strongly depends on the training data. If the scoring function is challenged with a drastically different protein than proteins in the training set, it is possible that the classification may fail. To further test how well the non-linear scoring function performs when discriminating proteins that are dissimilar to those contained in the training set, we take five proteins that are longer than any training proteins (lengths between 46 and 688). These are obtained from the list of 1261 polypeptide chains contained in the updated October 15, 2002 release of WHATIF database. The first test is to discriminate the five proteins from 1728 exhaustively generated design decoys using gapless threading. The second test is to discriminate these five proteins from exhaustively enumerated sequence decoys generated by threading 14 large protein sequences of unknown

Table 3. Discriminating large proteins from decoys.

pdb	N	n	(a) Design decoy by KDF		(b) Folding decoy by KFF		(c) SwissProt decoy by KDF		
			H	Δ_{score}	H	Δ_{score}	n	H	Δ_{score}
1cs0.a	1073	0	2.67	N/A	2.31	N/A	8232	2.67	2.42
1g8k.a	822	545	2.07	4.18	1.49	4.71	11 997	2.07	1.69
1gqi.a	708	1002	3.03	5.16	2.82	5.03	13 707	3.03	2.16
1kqf.a	981	93	2.19	5.17	1.85	4.95	9612	2.19	1.82
1lsh.a	954	148	1.97	4.57	1.66	4.02	10 017	1.97	2.01

Discrimination of five large proteins against (a) design decoys and (b) folding decoys generated by gapless threading, and against (c) additional design decoys generated by threading unrelated long proteins (length from 1124 to 2459) to the structures of these five proteins. Here pdb is the pdb code of the protein structure, N is the size of protein, n is the number of decoys, H is the predicted value of the scoring function, Δ_{score} is the smallest gap of score between the native protein and its decoys. The results show that all decoys can be discriminated from natives, and the smallest score gaps between native and decoys are large.

structures obtained from SwissProt database, whose sizes are between 1124 and 2459. This is necessary since structures of the longest chains otherwise have few or no threading decoys. Table 3 lists results of these tests, including the predicted score value and the smallest gap between the native protein and decoys. For the first test, the non-linear design scoring functions can discriminate these five native proteins from all decoys in the first test. For the second test, the design scoring function can also discriminate all five proteins from a total of 53 565 SwissProt sequence decoys, and the smallest score gaps between native and decoys are large.

We find that it is infrequent for an unknown test protein to have low similarity to all reference proteins. For each protein in the 440 training set, we calculate its Euclidean distance to the other 439 proteins. The distribution of the 440 maximum distances for each training protein to all other 439 proteins is shown in Figure 3a. We also calculate for each protein in the 201 test set its maximum distance to all training proteins (Fig. 3b). It is clear that for most of the 201 test proteins, the values of maximum distances to training proteins are similar to the values for training set proteins. The only exceptions are two proteins, ribonuclease inhibitor (1a4y.a) and formaldehyde ferredoxin oxidoreductase (1b25.a). Although they are correctly classified, the former has a significant amount of unaccounted interchain contact with another protein angiogenin, and the latter has iron/sulfur clusters. It seems that the set of training proteins provide an adequate basis set for characterizing the global fitness landscape of sequence design for other proteins.

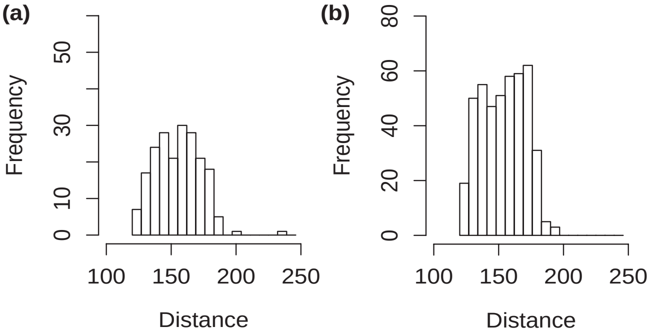

Fig. 3. The distribution of maximum distances of proteins to the set of training proteins. (**a**) The maximum distance for each training protein to all other 439 proteins. (**b**) The maximum distance for each protein in the 201 test set to all 440 training proteins. These two distributions are similar.

Nature of misclassification. We further distinguish misclassifications due to native protein being too close to a decoy and misclassifications due to decoys being too close to a native protein. Among the set of 201 test proteins, the native sequences of 13 proteins are not recognized correctly from design decoys. These 13 proteins are truly misclassifications because they do not have extensive unaccounted interchain interactions or cofactor interactions. We calculate the Euclidean distance of each of the 13 proteins from the 220 native proteins and 1685 decoys that participate in the kernel design scoring function. The results are shown in Table 4, where the number of native proteins among the top 3, 5 and 11 nearest neighboring vectors to the failed protein are listed. Except protein 1bx7, all misclassifications are due to native vectors being too close to decoys.

5 DISCUSSION

Formulation of non-linear scoring function. A basic requirement for computational studies of protein design is an effective scoring function, which allows searching and identifying of sequences adopting the desired structural templates. Our study follows earlier works such as Vendruscolo *et al.* (2000b), Tobi and Elber (2000) and Goldstein *et al.* (1992), where empirical scoring functions based on coarse residue level representation have been developed by optimization. The goal of this study is to explore ways to improve the sensitivity and/or specificity of discrimination.

There are several routes towards improving empirical scoring functions. One approach is to introduce higher-order interactions, where three-body or four-body interactions are

Table 4. Nearest neighbors of misclassified proteins

pdb	3-NN	5-NN	11-NN
1bd8	0	0	0
1bx7	2	3	5
1bxy.a	0	1	1
1cku.a	1	2	2
1dpt.a	2	2	2
1flt.v	1	3	3
1hta	0	0	0
1mro.c	0	0	0
1ops	1	1	1
1psr.a	1	1	1
1rb9	1	1	1
1ubp.b	1	1	1
3ezm.a	0	0	0

The nearest neighbors of the 13 proteins misclassified by design function. The number of native protein support vectors among the top 3, 5 and 11 nearest neighbors (NNs) are listed. Except protein 1bx7, the majority of nearest neighbors of all misclassified proteins are decoys.

explicitly incorporated in the scoring function (Zheng *et al.*, 1997; Munson and Singh, 1997; Betancourt and Thirumalai, 1999; Rossi *et al.*, 2001; Li *et al.*, 2003). A different approach is to introduce non-linear terms. Recently, Fain *et al.* (2002) use sums of Chebyshev polynomials up to order 6 for hydrophobic burial and each type of pairwise interactions.

In this work, we propose a different framework for developing empirical protein scoring functions, with the goal of simultaneous characterization of fitness landscapes of many proteins. We use a set of Gaussian kernel functions located at both native proteins and decoys as the basis set. Decoys set in this formulation are equivalent to the reference state or null model used in statistical potential. The expansion coefficients $\{\alpha_N\}$, $N \in \mathcal{N}$ and $\{\alpha_D\}$, $D \in \mathcal{D}$ of the Gaussian kernels determine the specific form of the scoring function. Since native proteins and decoys are non-redundant and are represented as unique vectors $\boldsymbol{c} \in \mathbb{R}^d$, the Gram matrix of the kernel function is full-rank. Therefore, the kernel function effectively maps the protein space into a high-dimensional space in which effective discrimination with a hyperplane is easier to obtain. The optimization criterion here is not Z-score, rather we search for the hyperplane in the transformed high-dimensional space with maximal separation distance between the native protein vectors and the decoy vectors. This choice of optimality criterion is firmly rooted in a large body of studies in statistical learning theory, where the expected number of errors in classification of unseen future test data is minimized probabilistically by balancing the minimization of the training error (or empirical risk) and the control of the capacity of specific types of functional form of the scoring function (Vapnik, 1995; Burges, 1998; Schölkopf and Smola, 2002).

This approach is general and flexible, and can accommodate other protein representations, as long as the final descriptor of protein and decoy is a d-dimensional vector. In addition, different forms of non-linear functions can be designed using different kernel functions, such as polynomial kernel and sigmoidal kernels. It is also possible to adopt different optimality criteria, e.g. by minimizing the margin distance expressed in 1-norm instead of the standard 2-norm Euclidean distance.

Folding scoring function. The geometric views of design scoring function and the optimality criterion also apply to the protein folding problem. For folding scoring function, the only difference from the design scoring function of Equation (6) is that here $\mathcal{D}$ is a set of structure decoys rather than a set of sequence decoys. Specifically, we generate for each native protein $(\boldsymbol{s}_N, \boldsymbol{a}_N)$ a set of structure decoys $\{(\boldsymbol{s}_D, \boldsymbol{a}_N)\}$, i.e. by mounting the native sequence on fragment of the structure of a large protein such that it contains exactly the same number of amino acid residues as the native protein. We use the same training set of 440 proteins from WHATIF98 and 14 080 766 structural decoys as in design study. The same optimization technique of margin maximization is used. The σ^2 value and the number of proteins and decoys entering the final folding scoring function are listed in Table 1.

For comparison, we also report discrimination results of the optimal linear scoring function taken as reported in Tobi *et al.* (2000), as well as the statistical potential developed by Miyazawa and Jernigan. Here we use the contact definition reported in Tobi *et al.* (2000), i.e. two residues are declared to be in contact if the geometric centers of their side chains are within a distance of 2.0–6.4 Å.

To test non-linear folding scoring functions for the same 194 and 201 test set proteins, we take the structure $\boldsymbol{s}$ from the conformation–sequence pair $(\boldsymbol{s}, \boldsymbol{a}_N)$ with the lowest score as the predicted structure of the native sequence. If it is not the native structure $\boldsymbol{s}_N$, the discrimination failed and the folding scoring function does not work for this protein. The results of discrimination are summarized in Table 5. There are four and eight misclassified native structures for the 194 set and 201 set, respectively. These correspond to a failure rate of 2.1 and 4.0%, respectively. The performance of the optimal non-linear kernel folding scoring function is better than the optimal linear scoring function of Tobi *et al.* (2000), based on calculation using values taken from Tobi *et al.* (2000) (failure rates 3.6 and 6.5% for the 194 set and 201 set, respectively), and is comparable to the results using values taken from Bastolla *et al.* (2001) (two and five misclassifications, failure rates of 1.0 and 2.5% for the 194 set and 201 set, respectively). Consistent with previous reports (Clementi *et al.*, 1998), statistical potential has ∼43.8% (81 out of 194) and 43.2% (87 out of 201) failure rates for the 194 set and the 201 set, respectively.

An updated study to Vendruscolo *et al.* (2000b) reported perfect discrimination for 1000 proteins from folding decoys (Bastolla *et al.*, 2001). Our results cannot be directly compared with this study, because many of the test proteins or their homologs in our study are likely to be included in the training

Table 5. Blind discrimination test of protein structures

	Misclassified natives	Misclassified natives
Kernel folding scoring function	4/194	8/201
Tobi and Elber	7/194	13/201
Bastolla *et al.*	2/194	5/201
Miyazawa and Jernigan	85/194	92/201
Kernel design scoring function	4/194	9/201

The number of misclassified protein structures for the test set of 194 proteins and the set of 201 proteins using non-linear kernel folding scoring function, two optimal linear scoring function taken as reported in Tobi *et al.* (2000) and in table I of Bastolla *et al.* (2001) and Miyazawa–Jernigan statistical potential (Miyazawa and Jernigan, 1996). The set of 201 proteins include those with >30% interchain contacts and those with >10% missing coordinates. We also list performance of kernel design scoring function for structure recognition.

set of Bastolla *et al.* (2001), as it is the union of proteins in the WHATIF database and the PDBSELECT database. In addition, it is not clear whether all decoys generated by gapless threading were tested by Bastolla *et al.* (2001). This makes a direct comparison of the two studies rather difficult.

It is informative to examine the four misclassified proteins by the kernel folding scoring function (1bx7, 1hta, 1ops and 3ezm.a). Hirustasin 1bx7 contain five disulfide bonds, which are not modeled explicitly by the protein description. 1hta (histone Hmfa) exists as a tetramer in complex with DNA under physiological condition. Its native structure may not be the same as that of a lone chain. The two terminals of this protein are rather flexible, and their conformations are not easy to determine. Among the 13 native sequences misclassified by the kernel design scoring function (1bd8, 1bx7, 1bxy.a, 1cku.a, 1dpt.a, 1flt.v, 1hta, 1mro.c, 1ops, 1psr.a, 1rb9, 1ubp.b, 3ezm.a), several have extensive interchain interactions, although the contact ratio is below the rather arbitrary threshold of 30%: Contact ratio of 24% for 1mor.c, 19% for 1upb.b, 24% for 1flt.v, 15% for 1psr.a and 13% for 1qav.a. It is likely that the substantial contacts with other chains would alter the confirmation of a protein. 1cku.a (electron transfer protein) contains an iron/sulfur cluster, which covalently binds to four Cys residues and prevent them from forming two disulfide bonds. These covalent bonds are not modeled explicitly. 1bvf (oxidoreductase) is complexed with a heme and an FMN group. The conformations of 1cku.a may be different upon removal of these functionally important hetero groups. Altogether, there is some rationalization for 8 of the 13 misclassified proteins.

In many cases, the misclassification of some native conformations is often indicative of the peculiar nature of the protein structures. This is true for both the linear scoring function reported in Vendruscolo and Domany (1998) and Vendruscolo *et al.* (2000b) and the non-linear kernel function developed in this study. For example, the misclassified proteins are often peptide chains stabilized by other chains, or by interactions with cofactors, or are small fragments whose interactions are modified by crystal lattice interactions, or are NMR structures which are less compact and less stable than X-ray structures. Although in this study we attempted to alleviate such complications by eliminating very short peptide fragments and excluding proteins with over 30% interchain contacts, it is unlikely all problematic protein structures can be completely eliminated from the training set. As shown by Bastolla *et al.* (2001), the design of optimized scoring function is likely to be open to the presence of wrong samples when a large training set is used.

For protein folding scoring functions derived from simple decoys generated by gapless threading, a more challenging test is to discriminate native proteins from an ensemble of explicitly generated three-dimensional decoy structures with a significant number of near-native conformations (Park and Levitt, 1996; Samudrala and Moult, 1998). Here, we evaluate the performance of non-linear scoring functions using three decoy sets from the database “DECOYS R US” (Samudrala and Levitt, 2000): the 4STATE_REDUCED set, the LATTICE_SSFIT set and the LMSD set. We compare our results in performance with results reported in the literature using optimal linear scoring function (Tobi and Elber, 2000) and statistical potential (Miyazawa and Jernigan, 1996) (Table 6). For the 4STATE_REDUCED set of decoys, non-linear folding scoring function has the best performance in terms of identifying the native structure, with only one misclassification (2cro). The correlation of root mean square distance (RMSD) of conformations to the native structure and score value in the 4STATE set is shown in Figure 4. Although the performance of discriminating explicitly generated challenging decoys is not as good as that of discriminating decoys generated by threading, it is likely that non-linear kernel scoring functions can be further improved if more realistic structural decoys are included in training. The generation of realistic structural decoys is more involved. Several methods have been developed for generating realistic decoys, including the original ‘build-up’ method (Park and Levitt, 1996), those with additional energy minimization (Loose *et al.*, 2004), and the method based on fragment assembly (Simons *et al.*, 1997). In addition, effective strategy of sequential importance sampling has also been proposed to generate protein-like long-chain compact

Table 6. Discriminating native structures from decoy structures

Protein	No. of decoys	KFF	KDF	MJ	TE-13	BFKV
1. 4state_reduced						
1ctf	631	1/3.64 (0.49)	1/3.14(0.55)	1/3.73	1/4.20	2/3.00
1r69	676	1/3.77 (0.45)	1/3.79 (0.55)	1/4.11	1/4.06	1/4.30
1sn3	661	1/2.15 (0.24)	27/1.79 (0.41)	2/3.17	6/2.70	1/2.89
2cro	675	3/2.57 (0.54)	1/2.66 (0.61)	1/4.29	1/3.48	2/2.91
3icb	654	1/2.56 (0.70)	1/2.68 (0.74)	—	—	1/2.96
4pti	688	1/4.17 (0.41)	1/2.79 (0.54)	3/3.16	7/2.43	1/3.49
4rxn	678	1/3.45 (0.47)	7/1.99 (0.53)	1/3.09	16/1.97	1/3.32
2. lattice_ssfit						
1beo	2001	15/2.45	1/3.94	—	—	1/3.70
1ctf	2001	1/3.76	1/5.35	1/5.35	1/6.17	1/4.66
1dkt	2001	17/2.42	8/2.64	32/2.41	2/3.92	4/3.38
1fca	2001	56/2.00	98/1.76	5/3.40	36/2.25	14/2.56
1nkl	2001	1/3.60	1/3.51	1/5.09	1/4.51	1/4.53
1pgb	2001	1/3.95	1/4.91	3/3.78	1/4.13	1/3.41
1trl	2001	56/1.97	18/2.67	4/2.91	1/3.63	90/1.75
4icb	2001	1/3.92	1/5.31	—	—	1/4.39
3. 1msd						
1b0n-B	498	406/−0.94	19/2.05	—	—	257/−0.03
1bba	501	500/−3.58	487/−1.83	—	—	500/−3.31
1ctf	498	1/3.62	1/3.31	1/3.86	1/4.13	1/2.92
1dtk	216	59/0.64	185/−1.11	13/1.71	5/1.88	54/0.74
1fc2	501	501/−3.08	486/−1.87	501/−6.24	14/2.04	501/−3.84
1igd	501	1/5.18	1/3.93	1/3.25	2/3.11	6/2.68
1shf-A	438	5/2.14	12/1.82	11/2.01	1/4.13	1/3.28
2cro	501	2/2.65	1/3.24	1/5.07	1/3.96	1/4.59
2ovo	348	1/3.11	38/1.21	2/3.25	1/3.62	40/1.15
4pti	344	1/3.14	108/0.62	—	—	10/1.86

The results of discrimination of native structures from decoys using non-linear kernel scoring functions. The decoy sets include 4state_reduced set, Lattice_ssfit set and lmsd set (Samudrala and Levitt, 2000). The rank of the native structure and its Z-score are listed. The correlation coefficient R is also listed in parentheses for the 4state_reduced set. KFF stands for kernel folding scoring function and KDF stands for kernel design scoring function. TE-13 scoring function is a linear distance-based scoring function optimized by linear programming, taken as reported in Tobi and Elber (2000), BFKV the linear scoring function reported in Bastolla *et al.* (2001) and MJ is the statistical scoring function as reported in Miyazawa and Jernigan (1996). The results for TE-13 scoring function and Miyazawa–Jernigan scoring function are taken from table II of Tobi and Elber (2000).

self-avoiding walk to overcome the attribution problem (Zhang *et al.*, 2003). This approach has been applied to generate realistic decoys. Preliminary results of deriving scoring function using such decoys can be found in Zhang *et al.* (2004).

Non-linear scoring function for folding and design. Sequence decoys and structure decoys in general lead to different scoring functions. For example, the contact count vectors $\boldsymbol{c}$ can be very different for a sequence decoy of a protein and a structure decoy of the same protein. The discrimination surface defined by the design scoring function and the folding scoring function therefore may be different (Table 7). There are 220 out of 440 native proteins participating in design scoring function, and 214 out of 440 native proteins participating in folding scoring function. There are 199 proteins that appear both in folding and design scoring functions. The majority of the native proteins have similar α values for both folding and design scoring functions. Figure 5 shows the difference $\Delta\alpha_i$ of the coefficient α_i for protein i appearing in both folding scoring function and design scoring function. In most cases, $\Delta\alpha_i$ values are small. That is, most native proteins contribute similarly in design scoring function and in folding scoring function. This is expected because the main differences between the two scoring functions are due to differences in decoys. Out of the top 20 proteins with the largest $|\alpha_i|$ values, 11 are common for both folding and design scoring functions. It is possible that the score values by kernel folding scoring function and by kernel design scoring function may be similar for many structure–sequence pairs $(\boldsymbol{s}, \boldsymbol{a})$. Figure 6a shows that the 194 proteins in the test set have similar score values by the kernel folding and kernel design scoring functions.

We also compare the values of the scoring functions for each of the vectors $\boldsymbol{c}_1 = \{1, 0, \ldots, 0\}^{\mathrm{T}}, \ldots, \boldsymbol{c}_{210} = \{0, \ldots, 1\}^{\mathrm{T}}$. We normalize these values so max $H(\boldsymbol{c}_i) = 1$ for both scoring functions (Fig. 6b). There is a strong correlation ($R = 0.94$) for folding and design scoring functions.

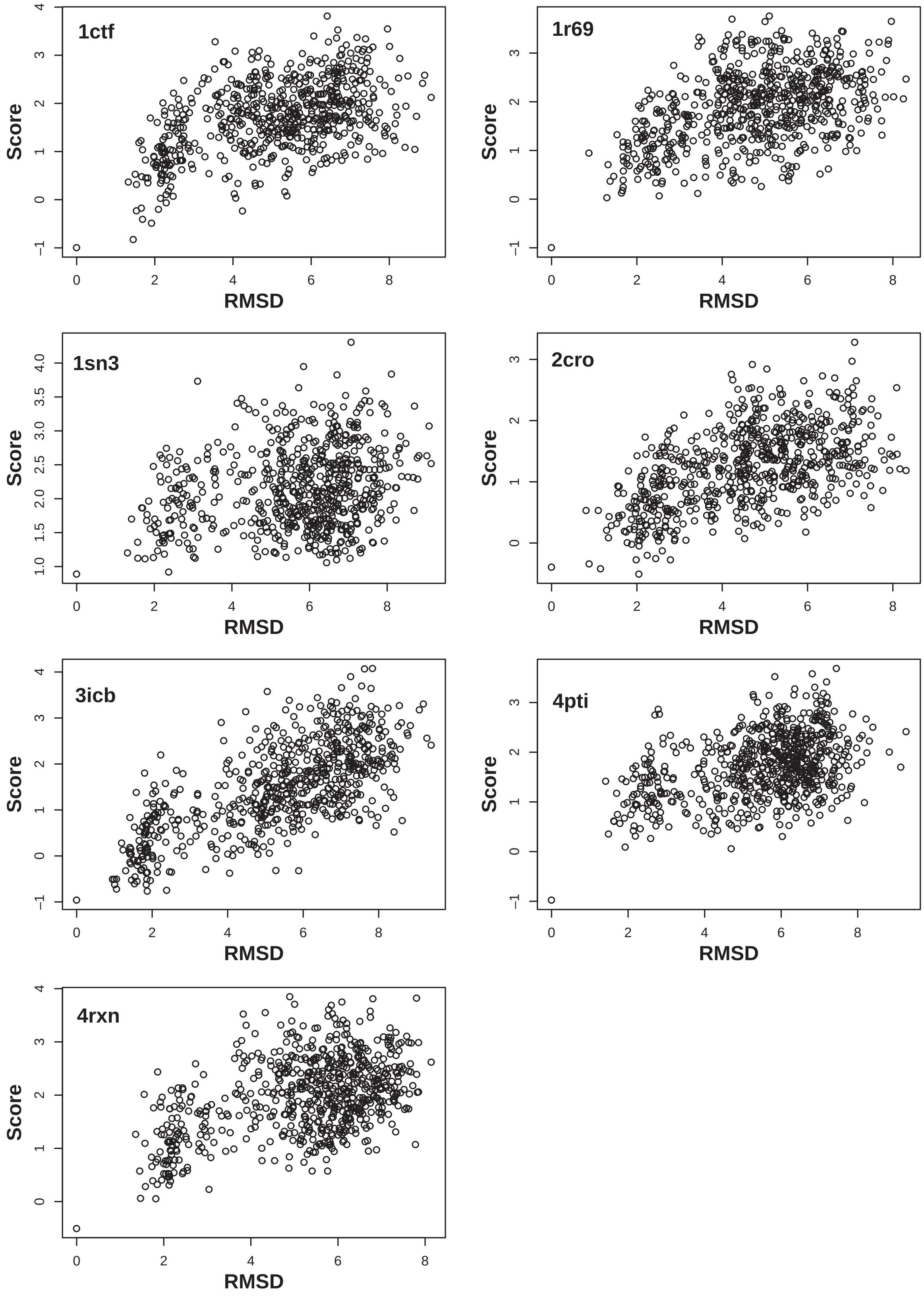

Fig. 4. Correlation of scores of decoys evaluated by non-linear kernel folding scoring function and their RMSD values to the native proteins in the 4STATE_REDUCED set.

Table 7. Proteins contributing most to scoring functions

Index	pdb	Kernel design scoring function			Kernel folding scoring function			
		α-value	Class	Number of resides	pdb	α-value	Class	Number of resides
1	2por	130.88	Membrane/cell	301	2spc.a	25.95	α	107
2	1prn	96.73	Membrane/cell	289	2por	23.19	Membrane/cell	301
3	2spc.a	52.27	α	107	1prn	14.31	Membrane/cell	289
4	1nsy.a	51.41	α/β	271	1rop.a	13.28	α	56
5	3pch.m	45.22	β	236	2wrp.r	11.41	α	104
6	1bkj.a	40.37	$\alpha+\beta$	239	1nsy.a	10.68	α/β	271
7	1xjo	36.02	α/β	276	1apy.a	10.12	$\alpha+\beta$	161
8	1bdb	34.26	α/β	276	1tgs.i	9.83	Small	56
9	1ppr.m	31.70	α	312	3pch.m	9.66	β	236
10	1fiv.a	27.48	β	113	1dan.l	8.80	Small	132
11	1hcz	27.23	β	250	7ahl.a	8.78	Membrane/cell	293
12	1tta.a	27.16	β	127	2ilk	8.72	α	155
13	7ahl.a	26.69	Membrane/cell	293	1ppr.m	8.25	α	312
14	2rhe	26.24	β	114	1bkj.a	8.09	$\alpha+\beta$	239
15	3pch.a	26.23	β	200	1cot	8.04	α	121
16	1snc	26.10	β	135	1wht.b	7.54	α/β	153
17	1wht.b	24.69	α/β	153	1vps.a	7.25	β	285
18	1cot	23.80	α	121	1vls	7.05	α	146
19	1bv1	23.58	β	159	1snc	6.48	β	135
20	2kau.b	22.45	β	101	1cmb.a	6.48	α	104

The top 20 proteins with the largest α-value among 199 proteins entering both kernel folding scoring function and kernel design scoring function. The α-value, the protein class as defined by SCOP and the number of residues are also listed.

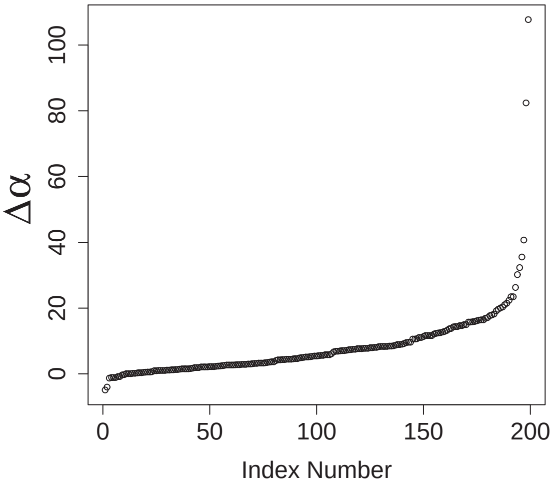

Fig. 5. The difference in contribution to the scoring function for the 199 native protein structures that participate in both folding and design scoring functions. They are sorted by $\Delta\alpha = \alpha_{\text{design}} - \alpha_{\text{folding}}$. The majority of them have $\Delta\alpha$ close to 0.

However, other methods reveal that kernel folding and design scoring functions are different. One method is to compare the scores of a subset of decoy structures that are challenging. That is, we compare the evaluated scores of decoys with $\alpha_i \neq 0$. Figure 6c shows that for decoys appearing in the design scoring functions, there is little correlation in the scores calculated by design scoring function and by folding scoring function. Similarly, there is no strong correlation between the scores calculated by folding scoring function and by design scoring function for the set of structure decoys entering the design scoring function (Fig. 6d). It seems that although the values of α_Ns are similar for the majority of the native proteins, design scoring function and folding scoring function can give very different score values for some conformations. This suggests that the overall fitness for design and folding potential may be different. However, since all empirical scoring functions derived from optimization and protein structures depend on the choice of traning set proteins and decoys, we cannot rule out the alternative explanation that the observed difference between design and folding scoring functions may be due to the difference in the decoy sets.

Remarks. Our goal in this study is to explore an alternative formulation of scoring function and assess the effectiveness of this new approach with experimental data. The non-linear scoring functions obtained in this study should be further improved. For example, unlike the study of optimal linear scoring function (Tobi *et al.*, 2000), where explicitly generated three-dimensional decoys structures are used in training, we used only structure decoys generated by threading. The test results using the 4STATE_REDUCED set and the LATTICE_SSFIT are comparable or better with other residue-based scoring functions (Fig. 4 and Table 6). It is likely

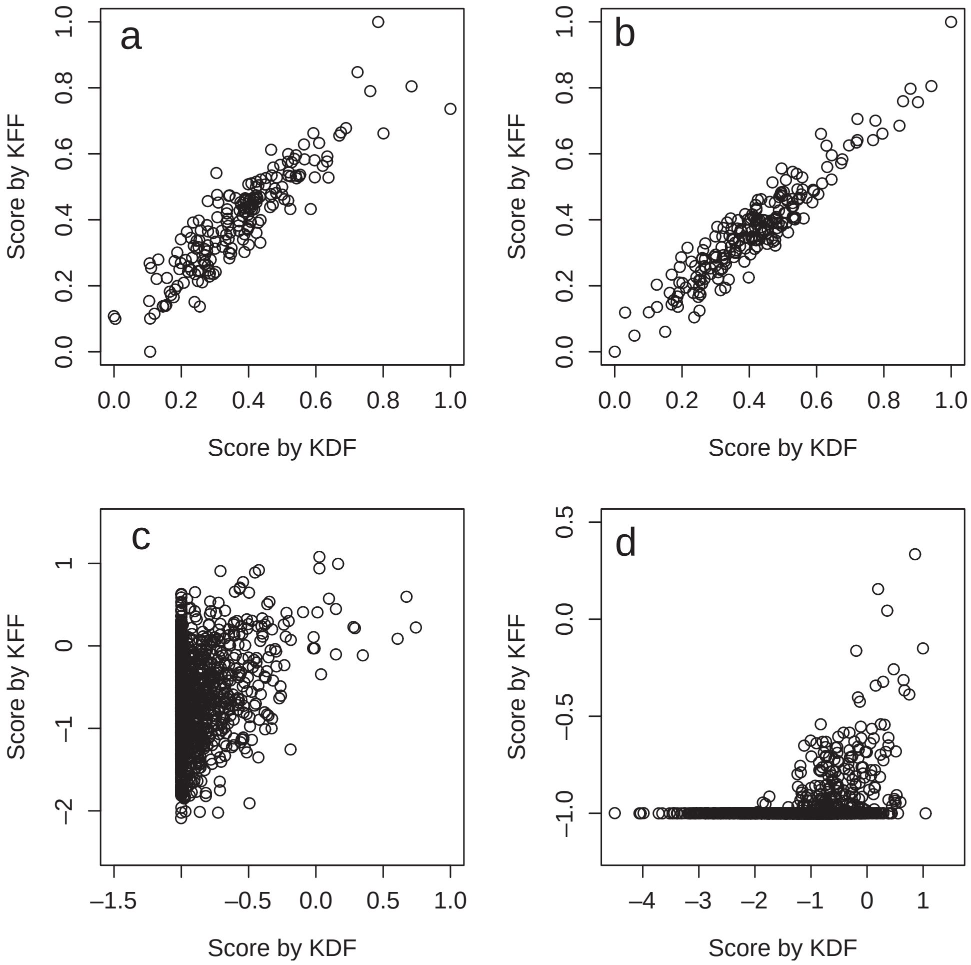

Fig. 6. Comparison of kernel design scoring function (KDF) and non-linear kernel folding scoring function (KFF). (**a**) The score values by KFF and by KDF for the 194 proteins are strongly correlated. The correlation coefficient is $R = 0.90$. (**b**) The score values of the non-linear design and folding scoring functions for the 210 unit vectors are strongly correlated ($R = 0.94$). (**c**) The score values by both design scoring functions and folding scoring functions for decoys that enter the non-linear design functions are poorly correlated. (**d**) The score values for decoys that enter the non-linear folding scoring functions are also poorly correlated.

that further incorporation of explicit three-dimensional decoy structures in the training set would improve the protein scoring function.

The evaluation of the non-linear scoring function requires more computation than for the linear function, but the time required is moderate: on an AMD Athlon MP1800+ machine of 1.54 GHz clock speed with 2 GB memory, we can evaluate the scoring function for 8130 decoys per minute.

Overfitting can be a problem in discrimination. Overfitting occurs when the scoring function predicts accurately the outcomes of training set data, but performs poorly when challenged with unrelated and unseen test data. Although our scoring function involves a large number of basis set proteins and decoys, it does not suffer from overfitting, because it has good performance in blind test of discriminating native proteins from both structural and sequence decoys.

In pursuit of improved sensitivity and specificity in discrimination, the number of reference decoys and native structures currently entering the scoring function is large (e.g. 1685 decoys and 220 native proteins for design scoring function). However, we expect the scoring function to be significantly simplified and the number of basis proteins and decoys reduced considerably. The use of 1-norm instead of 2-norm in the objective function of Equation (4) will automatically reduce the number of vectors (Schölkopf and Smola, 2002). In addition, new techniques such as finite Newton method for reduced support vector machine have recently shown great promise in further reducing the number of support vectors, where a reduction ratio of 1% has been reported (Lee and Mangasarian, 2001; Fung and Mangasarian, 2002).

Conclusions. We found in this study that no linear scoring function exists that can discriminate a training set of 440 native sequences from 14 million sequence decoys generated by gapless threading. The success of non-linear scoring function in perfect discrimination of this training set proteins and its good performance in an unrelated test set of 194 proteins is encouraging. It indicates that it is now possible to characterize simultaneously the fitness landscape of many proteins, and non-linear kernel scoring function is a general strategy for developing an effective scoring function for protein sequence design.

Our study of scoring function for sequence design is a much smaller task than developing a full-fledged fitness function, because we study a restricted version of the protein design problem. We need to recognize only one sequence that folds into a known structure from other sequences already known to be part of a different protein structure, whose identity is hidden during training. However, this simplified task is challenging, because the native sequences and decoy sequences in this case are all taken from real proteins. Success in this task is a prerequisite for further development of a full-fledged universal scoring function. A complete solution to the sequence design problem will need to incorporate additional sequences of structural homologs as native sequences, as well as additional decoy sequences that fold into different folds, and decoy sequences that are not proteins (e.g. all hydrophobes). It is our hope that the functional form and the optimization technique introduced here will also be useful for such purposes.

In summary, we show in this study an alternative formulation of scoring function using a mixture of Gaussian kernels. We demonstrate that this formulation can lead to an effective design scoring function that characterizes fitness landscape of many proteins simultaneously, and performs well in blind independent tests. Our results suggest that this functional form different from the simple weighted sum of contact pairs can be useful for studying protein design and protein folding. This approach can be generalized for any other protein representation, e.g. with descriptors for explicit hydrogen bond and higher-order interactions.

ACKNOWLEDGEMENTS

We thank Dr Hao Li for very helpful comments on protein design, Dr Yang Dai for discussion on linear programming, Dr Bosco Ho for suggestions on the presentation and Dr Mészáros for help in using the BPMD package. Dr Michele Vendruscolo kindly provided the list of the set of 456 proteins. This work is supported by funding from National Science Foundation CAREER DBI0133856, DBI0078270, Office of Naval Research N000140310329 and National Health Institute NIGMS GM68958-01.

REFERENCES

Bastolla,U., Farwer,J., Knapp,E. and Vendruscolo,M. (2001) How to guarantee optimal stability for most representative structures in the Protein Data Bank. *Proteins*, **44**, 79–96.

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 interaction schemes. *Protein Sci.*, **8**, 361–369.

Burges,C.J.C. (1998) A Tutorial on Support Vector Machines for Pattern Recognition, Knowledge Discovery and Data Mining, 2.

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

Clementi,C., Maritan,A. and Banavar,J. (1998) Folding, design, and determination of interaction potentials using off-lattice dynamics of model heteropolymers. *Phys. Rev. Lett.*, **81**, 3287–3290.

Dahiyat,B. and Mayo,S. (1997) *De novo* protein design: fully automated sequence selection. *Science*, **278**, 82–87.

DeGrado,W., Summa,C., Pavone,V., Nastri,F. and Lombardi,A. (1999) *De novo* design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.*, **68**, 779–819.

Desjarlais,J. and Handel,T. (1995) *De novo* design of the hydrophobic cores of proteins. *Protein Sci.*, **19**, 244–255.

Deutsch,J. and Kurosky,T. (1996) New algorithm for protein design. *Phys. Rev. Lett.*, **76**, 323–326.

Dima,R., Banavar,J., Cieplak,M. and Maritan,A. (2000) Scoring functions in protein folding and design. *Protein Sci.*, **9**, 812–819.

Drexler,K. (1981) Molecular engineering: an approach to the development of general capabilities for molecular manipulation. *Proc. Natl Acad. Sci., USA*, **78**, 5275–5278.

Edelsbrunner,H. (1987) *Algorithms in Combinatorial Geometry.* Springer-Verlag, Berlin.

Edelsbrunner,H. (1995) The union of balls and its dual shape. *Discrete Comput. Geom.*, **13**, 415–440.

Emberly,E.G., Wingreen,N.S. and Tang,C. (2002) Designability of alpha-helical proteins. *Proc. Natl Acad. Sci., USA*, **99**, 11163–11168.

Fain,B., Xia,Y. and Levitt,M. (2002) Design of an optimal Chebyshev-expanded discrimination function for globular proteins. *Protein Sci.*, **11**, 2010–2021.

Friedrichs,M. and Wolynes,P. (1989) Toward protein tertiary structure recognition by means of associative memory hamiltonians. *Science*, **246**, 371–373.

Fung,G. and Mangasarian,O.L. (2002) Finite newton method for lagrangian support vector machine classification. *Technical Report 02-01, Data Mining Institute, Computer Sciences Department, University of Wisconsin*.

Lazar,G.A., Desjarlais,J. and Handel,T. (1997) *De novo* design of the hydrophobic core of ubiquitin, *Protein Sci.*, **6**, 1167–1178.

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

Hao,M. and Scheraga,H. (1996) How optimization of potential functions affects protein folding. *Proc. Natl Acad. Sci., USA*, **93**, 4984–4989.

Hao,M.-H. and Scheraga,H. (1999) Designing potential energy functions for protein folding. *Curr. Opin. Struct. Biol.*, **9**, 184–188.

Jernigan,R. and Bahar,I. (1996) Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.*, **6**, 195–209.

Joachims,T. (1999) *Advances in Kernel Methods—Support Vector Learning*. Chapter 11, Making large-scale SVM learning practical. MIT Press.

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

Karmarkar,N. (1984) A new polynomial-time algorithm for linear programming. *Combinatorica*, **4**, 373–395.

Khatun,J., Khare,S.D. and Dokholyan,N.V. (2004) Can contact potentials reliably predict stability of proteins? *J. Mol. Biol.*, **336**, 1223–1238.

Koehl,P. and Levitt,M. (1999a) *De novo* protein design. I. In search of stability and specificity. *J. Mol. Biol.*, **293**, 1161–1181.

Koehl,P. and Levitt,M. (1999b) *De novo* protein design. II. Plasticity of protein sequence. *J. Mol. Biol.*, **293**, 1183–1193.

Koretke,K., Luthey-Schulten,Z. and Wolynes,P. (1996) Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci.*, **5**, 1043–1059.

Koretke,K., Luthey-Schulten,Z. and Wolynes,P. (1998) Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. *Proc. Natl Acad. Sci., USA*, **95**, 2932–2937.

Kuhlman,B. and Baker,D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci., USA*, **97**, 10383–10388.

Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L. and Baker,D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.

Lee,Y.-H. and Mangasarian,O.L. (2001) RSVM: reduced support vector machines. *Proceedings of the First SIAM International Conference on Data Mining*, Chicago, IL.

Lemer,C., Rooman,M. and Wodak,S. (1995) Protein-structure prediction by threading methods—evaluation of current techniques. *Proteins*, **23**, 337–355.

Li,H., Helling,R., Tang,C. and Wingreen,N. (1996) Emergence of preferred structures in a simple model of protein folding. *Science*, **273**, 666–669.

Li,X., Hu,C. and Liang,J. (2003) Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins*, **53**, 792–805.

Li,X. and Liang,J. (2004) Cooperativity and anti-cooperativity of three-body interactions in proteins. *J. Phys. Chem. B.*, In review.

Liang,J., Edelsbrunner,H., Fu,P., Sudhakar,P. and Subramaniam,S. (1998) Analytical shape computing of macromolecules I: molecular area and volume through alpha-shape. *Proteins*, **33**, 1–17.

Loose,C., Klepeis,J. and Floudas,C. (2004) A new pairwise folding potential based on improved decoy generation and side-chain packing. *Proteins*, **54**, 303–314.

Lu,H. and Skolnick,J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, **44**, 223–232.

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

Meller,J., Wagner,M. and Elber,R. (2002) Maximum feasibility guideline in the design and analysis of protein folding potentials. *J. Comput. Chem.*, **23**, 111–118.

Mészáros,C. (1996) Fast Cholesky factorization for interior point methods of linear programming. *Comput. Math. Appl.*, **31**, 49–51.

Micheletti,C., Seno,F., Banavar,J. and Maritan,A. (2001) Learning effective amino acid interactions through iterative stochastic techniques. *Proteins*, **42**, 422–431.

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 inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Miyazawa,S. and Jernigan,R. (1996) Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term. *J. Mol. Biol.*, **256**, 623–644.

Munson,P. and Singh,R. (1997) Statistical significane of hierarchical multi-body potential based on delaunay tessellation and their application in sequence–structure alignment. *Protein Sci.*, **6**, 1467–1481.

Pabo,C. (1983) Designing proteins and peptides. *Nature*, **301**, 200.

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.

Rossi,A., Micheletti,C., Seno,F. and Maritan,A. (2001) A self-consistent knowledge-based approach to protein design. *Biophys J.*, **80**, 480–490.

Samudrala,R. and Levitt,M. (2000) Decoys 'R' us: a database of incorrect conformations to improved protein structure prediction. *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.

Schölkopf,B. and Smola,A. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA.

Shakhnovich,E. (1998) Protein design: a perspective from simple tractable models. *Fold. Des.*, **3**, R45–R58.

Shakhnovich,E. and Gutin,A. (1993) Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci., USA*, **90**, 7195–7199.

Simons,K.T., 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 functions. *J. Mol. Biol.*, **268**, 209–225.

Simons,K.T., 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*, **34**, 82–95.

Sippl,M. (1995) Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, **5**, 229–235.

Slovic,A.M., Kono,H., Lear,J.D., Saven,J.G. and DeGrado,W.F. (2004) Computational design of water-soluble analogues of the potassium channel KcsA. *Proc. Natl Acad. Sci., USA*, **101**, 1828–1833.

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. (1996a) An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl Acad. Sci., USA*, **93**, 11628–11633.

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

Tobi,D. and Elber,R. (2000) Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins*, **41**, 40–46.

Tobi,D., Shafran,G., Linial,N. and Elber,R. (2000) On the design and analysis of protein folding potentials. *Proteins*, **40**, 71–85.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

Vapnik,V. and Chervonenkis,A. (1964) A note on one class of perceptrons. *Automat. Remote Contr.*, 25.

Vapnik,V. and Chervonenkis,A. (1974) *Theory of Pattern Recognition [in Russian]*, Nauka, Moscow (German Translation: Wapnik,W. and Tscherwonenkis,A. *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin, 1979).

Vendruscolo,M. and Domany,E. (1998) Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.*, **109**, 11101–11108.

Vendruscolo,M., Najmanovich,R. and Domany,E. (2000a) Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins*, **38**, 134–148.

Vendruscolo,M., Najmanovich,R. and Domany,E. (2000b) Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Prot. Struct. Funct. Genet.*, **38**, 134–148.

Vriend,G. and Sander,C. (1993) Quality control of protein models—directional atomic contact analysis. *J. Appl. Cryst.*, **26**, 47–60.

Wernisch,L., Hery,S. and Wodak,S. (2000) Automatic protein design with all atom force-fields by exact and heuristic optimization. *J. Mol. Biol.*, **301**, 713–736.

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

Yue,K. and Dill,K. (1992) Inverse protein folding problem: designing polymer sequences. *Proc. Natl Acad. Sci., USA*, **89**, 4163–4167.

Zhang,J., Chen,R. and Liang,J. (2004) Potential function of simplified protein models for discriminating native proteins from decoys: combining contact interaction and local sequence-dependent geometry. *26th Annual International Conference, IEEE Engineering in Medicine and Biology Society, 2004*. IEEE.

Zhang,J., Chen,R., Tang,C. and Liang,J. (2003) Origin of scaling behavior of protein packing density: a sequential Monte Carlo study of compact long chain polymers. *J. Chem. Phys.*, **118**, 6102–6109.

Zheng,W., Cho,S., Vaisman,I. and Tropsha,A. (1997) A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In Altman,R., Dunker,A., Hunter,L. and Klein,T. (eds), *Pacific Symposium on Biocomputing'97*. World Scientific, Singapore, pp. 486–497.

APPENDIX

LEMMA 1. *For a scoring function in the form of weighted linear sum of interactions, a decoy always has score values higher than the native structure by at least an amount of $b > 0$, i.e.*

$$\boldsymbol{w} \cdot (\boldsymbol{c}_D - \boldsymbol{c}_N) > b \quad \textit{for all } \{(c_D - c_N) | D \in \mathcal{D} \textit{ and } N \in \mathcal{N}\} \tag{8}$$

if and only if the origin $\mathbf{0}$ *is not contained within the convex hull of the set of points* $\{(\boldsymbol{c}_D - \boldsymbol{c}_N) | D \in \mathcal{D} \textit{ and } N \in \mathcal{N}\}$.

PROOF. Suppose that the origin $\mathbf{0}$ is contained within the convex hull $\mathcal{A} = \text{conv}(\{\boldsymbol{c}_D - \boldsymbol{c}_N\})$ of $\{\boldsymbol{c}_D - \boldsymbol{c}_N\}$ and Equation (8) holds. By the definition of convexity, any point inside or on the convex hull $\mathcal{A}$ can be expressed as a convex combination of points on the convex hull. Specifically, we have:

$$\mathbf{0} = \sum_{(c_D - c_N) \in \mathcal{A}} \lambda_{c_D - c_N} \cdot (\boldsymbol{c}_D - \boldsymbol{c}_N), \qquad \text{and}$$

$$\sum \lambda_{c_D - c_N} = 1, \lambda_{c_D - c_N} > 0.$$

That is, we have the following contradiction:

$$\begin{aligned} 0 = \boldsymbol{w} \cdot \mathbf{0} = \boldsymbol{w} \cdot \sum_{c_D - c_N} \lambda_{c_D - c_N} \cdot (\boldsymbol{c}_D - \boldsymbol{c}_N) \\ = \sum_{c_D - c_N} \lambda_{(c_D, c_N)} \cdot \boldsymbol{w} \cdot (\boldsymbol{c}_D - \boldsymbol{c}_w) \\ > \sum_{c_D - c_N} \lambda_{c_D - c_N} \cdot b = b. \end{aligned}$$

Because the convex hull can be defined as the intersection of half-hyperplanes derived from the inequalities, if a half-hyperplane has a distance $b > 0$ to the origin, all points contained within the convex hull will be on the other side of the hyperplane (Edelsbrunner, 1987). Therefore, $\boldsymbol{w} \cdot (\boldsymbol{c}_D - \boldsymbol{c}_N) > b$ will hold for all $\{(\boldsymbol{c}_D - \boldsymbol{c}_N)\}$.