Proceedings of the 26th Annual International Conference of the IEEE EMBS
San Francisco, CA, USA • September 1-5, 2004

OPTIMAL NONLINEAR SCORING FUNCTION FOR GLOBAL FITNESS LANDSCAPE OF PROTEIN DESIGN

Changyu Hu, Xiang Li and Jie Liang

Department of Bioengineering, SEO, MC-063
University of Illinois at Chicago, Chicago, IL 60607–7052, U.S.A.

ABSTRACT

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. It is also important that a design scoring function can characterize the global fitness landscape of many proteins simultaneously. We describe how finding optimal design scoring functions can be understood from two geometric viewpoints, and propose a formulation using mixture of Gaussian kernel functions. We give results of distinguishing native sequences for a major portion of representative protein structures from a large number of alternative decoy sequences. We succeeded in deriving nonlinear scoring function that perfectly discriminate a set of 440 representative native proteins of known protein structures from 14 million sequence decoys. We show that no linear scoring function can have perfect discrimination. In an independent blind test using 194 unrelated proteins, our scoring function misclassfies only 13 native proteins. This compares favorably with 37 or 51 misclassifications when optimal linear functions reported in literature are used.

1. INTRODUCTION

The problem of protein sequence design or inverse folding aims to identify sequences compatible with a given protein fold and incompatible to alternative folds [1]. The ultimate goal is to engineer protein molecules with improved activities or with acquired new functions. A successful protein design strategy needs a scoring function or fitness function to identify sequences that are compatible with the desired template fold. To achieve this, an ideal fitness function would maximize the probabilities of protein sequences taking their native fold instead of other structures.

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 fintness landscape of many proteins simultaneously. Specifically, we aim to identify protein sequences that are compatible with given three-dimensional coarse-grained structures from a large set of protein decoy sequences that are taken from proteins of unrelated folds.

We formulate in this study a novel protein scoring function, in the form of mixture of nonlinear Gaussian kernel functions. Experimentation shows that this scoring function can discriminate simultaneous 440 native proteins against 14 million sequence decoys. In contrast, there is no linear scoring function with perfect discrimination. We also perform blind tests to identify native sequence compatible to a protein fold from other decoy sequences. Taking 194 proteins unrelated to the 440 training set proteins, the nonlinear 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%).

2. THEORY AND MODELS

Modeling Protein Design Scoring Function. For protein descriptor, we use the vector $\boldsymbol{c} \in \mathbb{R}^d, d = 210$ of number counts of the 210 types of amino acid residue contacts. Once the structural conformation of a protein $\boldsymbol{s}$ and its amino acid sequence $\boldsymbol{a}$ is given, the contact vector $\boldsymbol{c}$ is fully determined.

We use a model analogous to the Anfinsen experiments in protein folding. We require that the native amino acid sequence $\boldsymbol{a}_N$ from a set of native proteins $\mathcal{N}$ mounted on its native structure $\boldsymbol{s}_N$ has the best (lowest) fitness score compared to a set of alternative sequences $\mathcal{D}$ (sequence decoys) taken from unrelated proteins of different fold mounted on the same native protein structure $\mathcal{D} = \{\boldsymbol{s}_N, \boldsymbol{a}_D\}$. We have: $H(f(\boldsymbol{s}_N, \boldsymbol{a}_N)) + b < H(f(\boldsymbol{s}_N, \boldsymbol{a}_D))$, where $b > 0$ for all $(\boldsymbol{s}_D, \boldsymbol{a}_N) \in \mathcal{D}$. Equivalently, the native sequence will have the highest probability to fit into the specified native structure.

A widely used functional form for protein scoring function H is the weighted linear sum of pairwise contacts. The linear sum score H is: $H(f(\boldsymbol{s}, \boldsymbol{a})) = H(\boldsymbol{c}) = \boldsymbol{w} \cdot \boldsymbol{c}$, where

Corresponding author. Phone: (312)355–1789, fax: (312)996–5921, email: `jliang@uic.edu`

0-7803-8439-3/04/$20.00©2004 IEEE

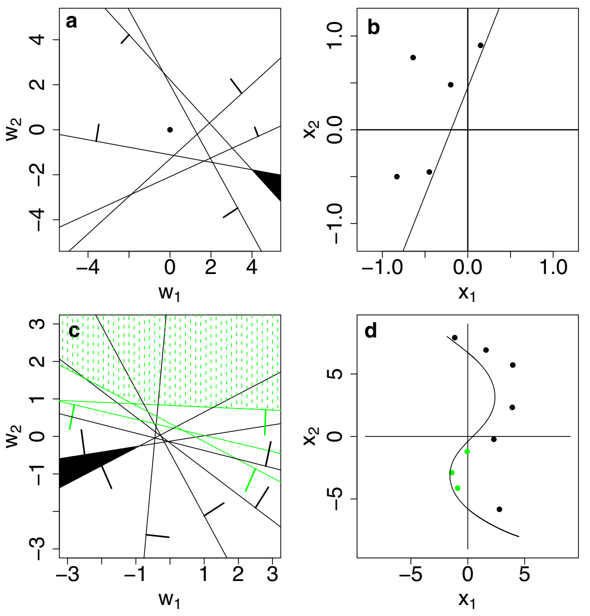

Fig. 1. Geometric views of the inequality requirement.

"·" denotes inner product of vectors. For such linear scoring functions, the basic requirement for design scoring function is then:

$$\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0, \qquad (1)$$

Our goal here is to obtain such a scoring function to discriminate native proteins from decoys.

Two Geometric Views of Linear Protein Scoring Potentials. There is a natural geometric view of the inequality requirement. Each of the inequalities divides the space of $\mathbb{R}^d$ ($\mathbb{R}^2$ of $\boldsymbol{w} = (w_1, w_2)$ in Fig 1a) into two halfs separated by a hyperplane. The hyperplane for Equation (1) 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 to 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 (filled polygon on the right in Fig 1a). 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.

For each native protein i, there is one convex polyhedron $\mathcal{P}_i$. To 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 $\mathbb{R}^d$ ($\mathbb{R}^2$ of $\boldsymbol{x} = (x_1, x_2)$ in Fig 1b, where $\boldsymbol{x} \equiv (\boldsymbol{c}_N - \boldsymbol{c}_D)$), the relationship $\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0$ for all sequence decoys and native proteins requires that all points $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ are located on one side of a hyperplane, which is defined by its normal vector $\boldsymbol{w}$ and its distance $b/||\boldsymbol{w}||$ to the origin. We can show that such a hyperplane exists if the origin is not contained within the convex hull of the set of points $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$. The second geometric view is dual and mathematically equivalent to the first geometric view.

Optimal Linear Scoring Function. According to the first geometric view, if the final convex polyhedron $\mathcal{P}$ is non-empty, there can be infinite number of choices of $\boldsymbol{w}$, all with perfect discrimination. But how do we find a weight vector $\boldsymbol{w}$ that is optimal? Here we describe an 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. 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 [2]. 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 mislabellings. The optimal weight vector $\boldsymbol{w}$ can be found by solving the following quadratic programming problem:

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}||\boldsymbol{w}||^2 \qquad (2) \\ \text{subject to} & \boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0 \text{ for all } \boldsymbol{c}_N \text{ and } \boldsymbol{c}_D.(3) \end{array}$$

We obtained the solution by solving the following support vector machine problem:

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}||\boldsymbol{w}||^2 \\ \text{subject to} & \boldsymbol{w} \cdot \boldsymbol{c}_N + d \leq -1 \\ & \boldsymbol{w} \cdot \boldsymbol{c}_D + d \geq 1, \end{array} \qquad (4)$$

where $d > 0$. Note that a solution of Problem (4) satisfies the constraints in Inequalities (3), since subtracting the second inequality here from the first inequality in the constraint conditions of (4) will give us $\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + 2 \leq 0$.

Nonlinear 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. Some decoys are very difficult to discriminate due to perhaps deficiency in protein representation. 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 filled convex polygon on the left, but another subset of inequalities (green lines) require that $\boldsymbol{w}$ 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. According to the second geometric view (Figure 1d), no hyperplane can separate all points (black and green) $\{\boldsymbol{c}_N - \boldsymbol{c}_D\}$ from the origin. In addition, even if a weight vector $\boldsymbol{w}$ can be found for each native protein, *i.e.*, $\boldsymbol{w}$ is contained in a nonempty 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.

A fundamental reason for this failure is that the functional form of linear sum is too simplistic. Here we introduce a nonlinear scoring function, which takes the following form: $H(f(\boldsymbol{s}, \boldsymbol{a})) = H(\boldsymbol{c}) = \sum_{D \in \mathcal{D}} \alpha_D K(\boldsymbol{c}, \boldsymbol{c}_D) - \sum_{N \in \mathcal{N}} \alpha_N K(\boldsymbol{c}, \boldsymbol{c}_N)$, where $\alpha_D \geq 0$ and $\alpha_N \geq 0$ are parameters to be determined, and $\boldsymbol{c}_D = f(\boldsymbol{s}_N, \boldsymbol{a}_D)$ for the set of decoys $\mathcal{D} = \{(\boldsymbol{s}_{\mathcal{N}}, \boldsymbol{a}_{\mathcal{D}})\}$ is the contact vector of a sequence decoy D mounted on a native protein structure $\boldsymbol{s}_N$, and $\boldsymbol{c}_N = f(\boldsymbol{s}_N, \boldsymbol{a}_N)$ for 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 $\boldsymbol{s}_N$. A convenient kernel function K is $K(\boldsymbol{x}, \boldsymbol{y}) = e^{-||\boldsymbol{x}-\boldsymbol{y}||^2/2\sigma^2}$ for any vectors $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathcal{N} \bigcup \mathcal{D}$, where σ^2 is a constant.

Optimal Nonlinear 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 value close to -1 for native proteins, and the decoys have values close to $+1$. We use an optimality criterion originally developed in statistical learning theory [2]. First, we note that we have implicitly mapped each structure and decoy from $\mathbb{R}^{210}$ through the kernel function of $K(\boldsymbol{x}, \boldsymbol{y}) = e^{-||\boldsymbol{x}-\boldsymbol{y}||^2/2\sigma^2}$ 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 nonlinear 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:

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

where C is a regularizing constant that limits the influence of each misclassified protein or decoy [2], and $y_i = -1$ if i is a native protein, and $y_i = +1$ if i is a decoy. When projected back to the space of $\mathbb{R}^{210}$, this hyperplane becomes a nonlinear surface. For the toy problem of Figure 1, Figure 1d shows that such a hyperplane becomes a nonlinear 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. A nonlinear scoring function in this case can have perfect discrimination.

Computational Methods. We use the count vector of pairwise contact interactions after normalization by the chain length of the protein. Here contacts are derived from the edge simplices of the alpha shape of a protein structure [3]. To obtain design decoys, 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 large protein to the full structure of the small protein. We therefore have for each native protein $(\boldsymbol{s}_N, \boldsymbol{a}_N)$ a set of sequence decoys $(\boldsymbol{s}_N, \boldsymbol{a}_D)$. Because all native contacts are retained in this case, sequence decoys obtained by gapless threading are challenging.

We use protein structures contained in the WHATIF database. It provides a good representative set of currently all known protein structures. We use a list of 440 proteins from WHATIF98 as training data. Using threading method, we generated a set of 14,080,766 sequence decoys.

We use SVMLIGHT [4] with Gaussian kernels and the 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 default value. The value of σ^2 for the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{y}) = e^{-||\boldsymbol{x}-\boldsymbol{y}||^2/2\sigma^2}$ is chosen by experimentation. The final design scoring function is obtained with σ^2 set to 416.7.

3. RESULTS

We succeeded in finding a nonlinear function that can discriminate all 440 native proteins from 14 million decoys. Unlike statistical scoring functions where each native protein in the database contribute 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. About 50% of native proteins and $< 0.1\%$ of decoys from the original training data enter the scoring function. No linear scoring function can be found by solving Eqn (4) that can perfectly discriminates 440 proteins from decoys.

Table 1. Discrimination results.

	Misclassified Natives	Misclassified Natives
KDF	13/194	19/201
TE	37/194	44/201
BFKV	51/194	54/201
MJ	81/194	87/201

We conduct a blind test in discriminating native proteins from decoys for an independent test set. To construct such a test set, we first take the entries in WHATIF99 database that are not present in WHATIF98. We then eliminate proteins with chain length less than 46 residues and those with $>$ 30% inter chain contact, and obtain a set of test proteins of 201 proteins. Further elimnation of structures with $>$ 10% missing coordinates give a smaller set of 194 proteins. Using gapless threading, we generate a sets of 3,096,019 sequence decoys.

To test design scoring functions for discriminating native proteins from sequence decoys, 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 reference [5, 6], as well as the statistical potential developed by Miyazawa and Jernigan.

Our nonlinear kernel design scoring function (KDF) is the only function capable of discriminating all of the 440 native sequences. It also works well for the test set (Table 1). 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 TE taken as reported in [5], which succeeded in identifying 80.9% (157 out of 194) of this test set. It also has better performance than optimal linear scoring function BFKV based on calculations using parameters reported in reference [6], which succeeded in identifying 73.7% (143 out of 194) of proteins in the test set. The Miyazawa-Jernigan statistical potential (MJ) succeeded in identifying 113 native proteins out of 194) (success rate 58.2%).

4. DISCUSSION

We found in this study that no linear scoring function exists that can discriminate a training set of 440 native sequence from 14 million sequence decoys generated by gapless threading. The success of nonlinear 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 nonlinear kernel scoring function is a general strategy for developing effective scoring function for protein sequence design.

In summary, we show in this study two geometric criteria for defveloping scoring function, and propose an alternative formulation of scoring function using a mixture of Gaussian kernels. We demonstrate that this formulation can characterize fitness landscape of many proteins simultaneously, and it performs well in blind independent tests. Our results suggest that this functional form 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.

5. ACKNOWLEDGMENT

We thank Dr. Hao Li for very helpful comments on protein design, Dr. Yang Dai for discussion on linear programming, and Dr. Bosco Ho for suggestion in presentation. This work is supported by funding from NSF CAREER DBI0133856, DBI0078270, ONR N000140310329, and NIH GM68958-01.

6. REFERENCES

[1] K.E. Drexler, "Molecular engineering: an approach to the development of general capabilities for molecular manipulation," *Proc Natl Acad Sci USA*, vol. 78, pp. 5275–5278, 1981.

[2] B. Schölkopf and A.J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, The MIT Press, Cambridge, MA, 2002.

[3] X. Li, C. Hu, and J. Liang, "Simplicial edge representation of protein structures and alpha contact potential with confidence measure," *Proteins*, vol. 53, pp. 792–805, 2003.

[4] T. Joachims, *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-scale SVM learning practical, MIT Press, 1999.

[5] D. Tobi, G. Shafran, N. Linial, and R. Elber, "On the design and analysis of protein folding potentials," *Proteins*, vol. 40, pp. 71–85, 2000.

[6] U. Bastolla, J. Farwer, E.W. Knapp, and M. Vendruscolo, "How to guarantee optimal stability for most representative structurs in the protein data bank," *Proteins*, vol. 44, pp. 79–96, 2001.