An Optimal Structure-Discriminative Amino Acid Index for Protein Fold Recognition

R. H. Leary,* J. B. Rosen,*† and P. Jambeck‡

*San Diego Supercomputer Center, †Department of Computer Science and Engineering, and ‡Department of Bioengineering,
University of California-San Diego, La Jolla, California 92093-0505

ABSTRACT Identifying the fold class of a protein sequence of unknown structure is a fundamental problem in modern biology. We apply a supervised learning algorithm to the classification of protein sequences with low sequence identity from a library of 174 structural classes created with the Combinatorial Extension structural alignment methodology. A class of rules is considered that assigns test sequences to structural classes based on the closest match of an amino acid index profile of the test sequence to a profile centroid for each class. A mathematical optimization procedure is applied to determine an amino acid index of maximal structural discriminatory power by maximizing the ratio of between-class to within-class profile variation. The optimal index is computed as the solution to a generalized eigenvalue problem, and its performance for fold classification is compared to that of other published indices. The optimal index has significantly more structural discriminatory power than all currently known indices, including average surrounding hydrophobicity, which it most closely resembles. It demonstrates >70% classification accuracy over all folds and nearly 100% accuracy on several folds with distinctive conserved structural features. Finally, there is a compelling universality to the optimal index in that it does not appear to depend strongly on the specific structural classes used in its computation.

INTRODUCTION

As the volume of protein sequence data grows, the development of efficient methods to assign new sequences to structural and functional classes becomes increasingly important. In many sequence analysis problems, it is desirable to classify protein sequences into two or more categories whose characteristics are known ahead of time. For example in fold recognition, a protein sequence is evaluated with respect to a set of known three-dimensional structure classes, and is assigned to the class with which it is most compatible. Another example is secondary structure prediction, in which the residues of a sequence are classified as α-helices, β-sheets, or other local structural motifs.

Sequences can be assigned to classes by alignment or by pattern recognition approaches. Alignment has historically been the more common approach in bioinformatics research. In alignment-based fold recognition, the target sequence is aligned against class libraries of representative aligned sequences or a class profile extracted from such libraries, and then assigned to the class most similar to the target. Alignment-based methods generally rely on scoring matrices that encode the pairwise similarity of all possible amino acids at a given position in the target and class representative sequence. These scoring matrices are combined with algorithmic procedures, usually based on dynamic programming, to extract subsequences from the targets that best match the class libraries. Thus alignment methods are based primarily on a direct positional comparison of sequence information.

In contrast, statistical pattern recognition methods typically extract a vector of numerical features from the sequence. This vector is then compared to the vectors (or a statistical summary of the vectors) of known representatives from each class. The target sequence is assigned to the class with the closest members, in the case of nearest neighbor classifiers, or more generally on the basis of statistically based discriminant functions. For example, many authors have attempted to classify target sequences into one of four top-level SCOP (Murzin et al., 1995) structural classes (all-α, all-β, $\alpha + \beta$, α/β), based on a 20-dimensional vector representing the fractional representation of each residue type in the target sequence. Perhaps the best-known example is the component-coupled method (Chou, 1995), in which the classification is based on simple quadratic discriminants using the means and covariance matrices of the vectors in the training set in the form of the Mahalanobis distances of the target vector to the training set means. A complementary approach was described by Bahar and co-workers (Bahar et al., 1997), who applied a singular value decomposition to a matrix of amino acid compositions for the database of proteins used by Chou (1995) to analyze the relationship between the amino acid types and fold classes. Other more elaborate discriminant functions for this case are implicitly used by neural network and support vector machine approaches, e.g., Ding and Dubchak (2001) and Cai et al. (2003).

Hybrid methods for fold recognition combine both sequence information and extracted numerical features. For example, it is well known that there is a strong tendency for hydrophobic residues in a globular protein to reside at interior, solvent inaccessible positions in the folded structure, and similarly for hydrophilic residues to reside at solvent accessible surface positions. This has been exploited for fold recognition in the form of functions that measure the compatibility of a target sequence with sequences and structures

Submitted June 4, 2003, and accepted for publication August 26, 2003.

Address reprint requests to R. H. Leary, E-mail: leary@sdsc.edu.

0006-3495/04/01/411/09 $2.00

in a fold library. For example, Bowie et al. (1990) used the known structures of sequences in a fold library to compute the corresponding solvent accessibilities at each residue. A compatibility function was then constructed based on a dynamic programming alignment of the Fauchere-Pliska hydrophobicity profile (Fauchere and Pliska, 1983) of the test sequence to the solvent accessibility profiles of the library sequences. More recently, Mallick et al. (2002) proposed a compatibility function that incorporates sequence-sequence comparisons, sequence-derived information, and structural environmental information. They note that the most important contribution is the sequence profile of the environmental hydrophobicity surrounding each residue.

Clearly the hydrophobicity profile of a sequence can form at least a partial basis for a compatibility function. But there are many different hydrophobicity indices, not all of which are highly correlated. Which form is best? Are there other independent amino acid properties that also have a strong structural discriminative capability, and what is the discriminatory power of hydrophobicity relative to these other indices? In general an amino acid index can be represented as a vector r in the space R^{20} that assigns a numerical value to each of the 20 possible residues. In Tomii and Kanehisa (1996), cluster analysis was used to compare the 402 property indices contained in the then-current version of the AAindex database (Kawashima and Kanehisa, 2000), a compilation of published indices. Their analysis revealed that each of the indices could be categorized into one of six groups: hydrophobicity, α-helix and turn propensity, physiochemical properties, β-strand propensity, and other properties. AAindex has now grown to 494 indices, and in principle each of these could be tested within a given fold recognition framework. However, this is somewhat unsatisfying as it only provides a rank ordering for specific cases. Also, the space R^{20} of possible amino acid indices is enormous and only very sparsely sampled by the 494 entries in the AAindex database. This leaves open the possibility that there are perhaps even better novel indices that do not appear in the literature.

Similarly, it is difficult to draw definitive conclusions from the many published fold recognition studies because these are based on a variety of frameworks and contexts, with compatibility functions that are often not directly comparable. Also, most studies in effect consider only a composite measure of the performance of separate alignment and fold assignment phases. As noted in Fischer and Eisenberg (1996), the alignment problem is much more difficult than the fold assignment problem. Thus somewhat more definitive conclusions regarding the usefulness of various amino acid indices, at least with respect to fold assignment, may be obtainable by considering the fold assignment problem separately from the alignment problem in a more controlled environment.

In this article we introduce FoldID, a general method for classifying sequences, and demonstrate its use in a fold assignment task. Fold assignment is performed in the context of a supervised learning scenario for pattern classification (Duda and Hart, 1973). Supervised learning refers to the use of a library of patterns (the ‘‘training set’’) with known class labels as the basis for the training of a classification rule. Usually the training consists of determining parameter values in the classification rule such that the rule performs optimally with respect to a merit function when tested on the training set. Performance evaluation is then done with a cross-validation procedure.

Here the classification rule for a given target sequence of length L is applied to the sequence profile vector in R^L obtained by replacing each residue with the corresponding numerical property defined by a generic amino acid index r in R^{20}. The index vector r is considered as the parameter vector to be optimized by the training process. A simple nearest centroid classification rule assigns target sequences to folds based on the closest library fold class centroid vector, as measured by Euclidean distance. A merit function representing the power of the index r to discriminate between the folds is defined as the ratio $J(r) = S_B(r)/S_W(r)$ of the between-class variation to within-class variation of the corresponding library profile vectors. The index r_{opt} with optimal discriminatory power that maximizes the merit function is obtained as the maximal eigenvector of a generalized eigenvalue problem, with other possibly useful independent indices being defined by lower eigenvectors.

The training set consists of a library of sequences that have been structurally aligned and organized into 174 structural classes (‘‘folds’’) by Shindyalov and Bourne using their Combinatorial Extension (CE) algorithm (Shindyalov and Bourne, 1998, 2000). Because the sequences within each class are already structurally aligned, no alignment phase is necessary during training. In fact, the given alignments are optimal in the sense that they are precisely the ones used to define the fold classes. Thus the assignment problem can be considered in the absence of noise from possibly suboptimal alignments.

MATERIALS AND METHODS

CE fold library

The training set is the CE3291 4-Å Common Subdomain Library constructed by Shindyalov and Bourne using their Combinatorial Extension structural alignment algorithm (Shindyalov and Bourne, 1998, 2000). The library consists of a total of 3876 sequences from the Protein Data Bank (Berman et al., 2000) that are organized into 174 folds. Here we use the term fold to indicate a generic class of similar structures defined by some suitable structural similarity measure. Shindyalov and Bourne use the term ‘‘common substructure’’ for their library, and note that the size of the substructure is typically at or below the domain level. Within the J-th CE fold, the sequences have a fixed length L_J (including gaps) that ranges from 64 to 300 residues, a uniformly low (<20%) sequence identity, and a maximum pairwise RMSD of 4 Å. The number N_J of sequences in fold J decreases essentially monotonically with J from the most populous fold $J = 1$ with $N_1 = 153$ down to fold $J = 174$ with $N_{174} = 5$.

Amino acid indices and the AAindex database

A collection of 494 published amino acid indices is available online as the AAindex database (Kawashima and Kanehisa, 2000). For purposes of displaying and comparing amino index vectors on a common unit-free numerical scale, we modify each AAindex index vector r by centering (subtracting the mean of the 20 components from each component) and normalizing to unit Euclidean length $\|r\| = 1$. Note that centering and normalization do not affect the correlation coefficient, the usual measure of similarity between two indices, or the performance of the classification rules used here.

FoldID supervised learning algorithm

In the discussion that follows, a sequence of length L is encoded as an L × 20 permutation matrix P such that if amino acid j appears in the i-th sequence position, then $P_{ij} = 1$, and otherwise $P_{ik} = 0$ for all $k \neq j$. Each amino acid index vector r in R^{20} thus defines a profile vector Pr in R^L where the matrix-vector product Pr simply substitutes the numerical index value for the corresponding amino acid at each sequence position. If there is a gap in position i, all entries in the i-th row of P are assigned a value of 0.05 and the i-th position in the profile vector is thus assigned a value equal to the arithmetic average of the amino index components (which is zero if the index is centered).

For simplicity of presentation, we assume all folds have the same sequence length $L_J = L$. The necessary modifications required for the more general case of unequal fold sequence lengths are straightforward and discussed at the end of this section. Let $\{P_i^J\}_{i=1}^{N_J}$ be the set of permutation matrices encoding the sequences in fold J. An index vector r defines a cluster of sequence profile vectors $\{P_i r\}_{i=1}^{N_J}$ in R^L with centroid $c_J(r)$ given by

$$c_J(r) = \frac{1}{N_J}\sum_{i=1}^{N_J} P_i^J r = \bar{P}^J r, \tag{1}$$

$$\text{where} \quad \bar{P}^J = \frac{1}{N_J}\sum_{i=1}^{N_J} P_i^J.$$

The centroid $c_J(r)$ can be thought of as a profile template for fold J, and the FoldID nearest centroid classification rule assigns a test sequence with profile Pr to the fold with the minimum Euclidean distance $\|Pr - c_J(r)\|$.

Let $N = \sum_{J=1}^{M} N_J$ be the total number of sequences in the training set, and M be the total number of folds. An ensemble centroid $\bar{c}(r)$ is defined as the mean profile over all sequences:

$$\bar{c}(r) = \frac{1}{N}\sum_{J=1}^{M} N_J c_J(r) = \bar{\bar{P}} r, \tag{2}$$

$$\text{where} \quad \bar{\bar{P}} = \frac{1}{N}\sum_{J=1}^{M} N_J \bar{P}^J.$$

The between-class variation $S_B(r)$ is the N_J-weighted sum of the squared Euclidean distances between the fold centroids and ensemble centroid:

$$S_B(r) = \sum_{J=1}^{M} N_J \|c_J(r) - \bar{c}(r)\|^2 = r^T S_B\, r,$$

$$S_B = \sum_{J=1}^{M} N_J (\bar{P}^J - \bar{\bar{P}})^T (\bar{P}^J - \bar{\bar{P}}). \tag{3}$$

Thus $S_B(r) = r^T S_B r$ is a positive semidefinite quadratic form corresponding to the 20 × 20 matrix S_B, i.e., a quadratic polynomial in the components of r that takes on only nonnegative values. Similarly, the within-class variation $S_W(r) = r^T S_W r$ is a positive semidefinite quadratic form representing the sum over all sequences of the squared distances of the sequence profile to its fold centroid:

$$S_W(r) = \sum_{J=1}^{M}\sum_{i=1}^{N_J} \|(P_i^J r - c_J(r))\|^2 = r^T S_W\, r,$$

$$S_W = \sum_{J=1}^{M}\sum_{i=1}^{N_J} (P_i^J - \bar{P}^J)^T (P_i^J - \bar{P}^J). \tag{4}$$

An ideal mapping vector r results in widely separated centroids in R^L and tight clustering around each centroid, which corresponds to a large value of $S_B(r)$ relative to $S_W(r)$. Thus we define the merit criterion function $J(r) = r^T S_B\, r / r^T S_W\, r$, which is the ratio of two positive semidefinite quadratic forms. If $r^T S_W\, r$ is positive definite (i.e., vanishes only at $r = 0$), then the stationary points where the gradient of $J(r)$ is equal to zero are the eigenvectors of the generalized eigenvalue problem

$$S_B r = \lambda S_W r, \tag{5}$$

and the (globally) optimal index vector r_{opt}, which maximizes $J(r)$ is the eigenvector corresponding to the largest eigenvalue λ_{max}, with $J(r_{opt}) = \lambda_{max}$. In the case here, a small technical difficulty arises as S_W is typically only positive semidefinite with rank 19, because $S_W e = 0$, where $e^T = (1, 1, \ldots, 1)$ is a 20-dimensional vector of all 1's. However, S_W can be converted to a positive definite matrix by adding a small multiple of the rank one matrix ee^T. This does not change the stationary points of $J(r)$ or the largest 19 eigenvectors/eigenvalues (r_i, λ_i) The 20th eigenvalue λ_{20} is zero, and corresponds to the uninteresting eigenvector $r_{20} = e$, which assigns all components of the index vector the same value.

The resulting generalized eigenvalue problem is solved with the DGESV routine from the standard Fortran linear algebra subroutine library LAPACK (Anderson et al., 1999) with computation of all eigenvalues and eigenvectors. In addition to the optimal eigenvector r_{opt} corresponding to the maximal eigenvalue, other lower eigenvectors may also be useful in classification and assessing the significance of the maximal eigenvector. For example, the eigenvector r_2 corresponding to the second largest eigenvalue maximizes $J(r)$ over all vectors that satisfy the orthogonality condition $r^T S_W r_{opt} = 0$. Similarly the third eigenvector r_3 is the optimal solution among all vectors that are S_W-orthogonal to both the first and second eigenvectors, etc. Thus the eigenvectors define a hierarchy of independent, S_W-orthogonal amino acid indices in decreasing order of discriminatory power as defined by the merit criterion $J(r)$.

The procedures outlined above require minor adjustment in the usual case where the sequence lengths L_J are not all identical. In this case the ensemble mean $\bar{c}(r)$ is defined component-wise, with the i-th component $\bar{c}_i(r)$ defined as the average of the i-th components of all profile vectors of length at least equal to i:

$$\bar{c}_i(r) = \sum_{\{J:\, i \le L_J\}} N_J (c_J(r))_i \Big/ \sum_{\{J:\, i \le L_J\}} N_J. \tag{6}$$

Similarly, the matrix $\bar{\bar{P}}$ is adjusted to reflect row-by-row averaging of $\bar{P}^J$ over all folds of appropriate sequence length. Finally, the nearest centroid rule is now based on comparing the distance $d(Pr, c_J(r))$ from the profile vector Pr to a centroid vector $c_J(r)$ that may be of a different length. Let L be the minimum of the two vector lengths. The generalized distance function is then defined as the root mean square difference of the first L components:

$$d(Pr, c_J(r)) = \left(\sum_{i=1}^{L} \left[(Pr)_i - (c_J(r))_i\right]^2 / L\right)^{1/2}. \tag{7}$$

Also, we have found that omitting gapped positions in the target sequence from the summation in Eq. 7 and adjusting L accordingly considerably improves classification performance.

Performance evaluation by hold-out, cross-validation, and self-consistency tests

The performance of a classification rule is typically measured by either a hold-out (HO) or K-fold cross-validation (CV) test, and usually compared with the results of a simple self-consistency (SC) test. In the HO procedure, the data set is divided into two disjoint representative subsets, one that is used for training and the other for testing. This has the advantage of both conceptual and computational simplicity, but is somewhat inefficient in the use of training data and tends to produce pessimistic estimates of ''true'' performance. However, for some purposes, notably comparing the relative performances of different classification methodologies, it may be superior to less biased but more computationally complex and possibly higher variance methods such as K-fold CV with large values of K. There is no universally acknowledged ''best'' method of performance evaluation, and a discussion of the tradeoffs between bias and variance in various approaches and the implications for different purposes can be found in (Kohavi, 1995).

In the K-fold CV procedure, the CE data set is divided into K roughly equal representative sets, each one of which is held out in turn as a testing set for a classification rule trained on the sequences in the remaining K-1 sets. The average correct classification rate on the K testing sets is then reported as the performance measure. An extreme form of the K-fold CV is the leave-one-out CV method, in which K is set equal to $N = 3876$, the total number of sequences. Thus during each of N training-testing cycles, exactly one sequence is held out for testing and all of the remaining sequences are used for training. This has the obvious disadvantage of computational complexity, but uses the training data very efficiently and has relatively little bias compared to HO or K-fold validation with low values of K.

Here we use two methods at opposite ends of this spectrum, namely a leave-one-out CV test and a 50% HO test, in addition to the self-consistency test. In the 50% HO test, training is done on the odd-numbered sequences in each fold, and the testing is done on the even-numbered sequences. The complementary test with the roles of the even and odd sequences reversed is also done, and the performance values averaged. Thus technically this is a twofold cross-validation, but we use the HO designation to emphasize the near equivalence to a 50% HO test and fundamental difference with the leave-one-out CV test.

The self-consistency test uses the entire CE library both as a training and a testing set. This is computationally undemanding, but tends to produce overly optimistic estimates of performance rates on new data, particularly for classes with limited training data. The gaps between HO, CV, and SC error rates can be a useful indicator of presence of bias, training data adequacy, and potential headroom for improvement with more training data.

RESULTS

The optimal index r_{opt} was computed using all 3876 sequences from the 174 fold families. As noted above, the first nineteen eigenvectors $r_1 = r_{\mathrm{opt}}$ through r_{19} span the space of interest, with the 20-th eigenvector $r_{20} = e$ corresponding to $\lambda_{20} = 0$ being useless for classification purposes. The amino acid indices defined by the first 3 eigenvectors (normalized to unit Euclidean length) are shown in Table 1.

A search of the AAindex database of 494 published amino acid indices was made to identify indices with high absolute correlations to r_{opt}. The maximum absolute correlation of 0.959 was observed for the Average Surrounding Hydrophobicity (ASH) index in (Manavalan and Ponnuswamy, 1978). This ASH index was constructed by summing the hydrophobicity index in Jones (1975) for all residues within an 8-Å radius of the residue of interest, and then averaging over a small protein small database. A nearly identical correlation of 0.950 was observed with an updated ASH index (Ponnuswamy and Gromiha, 1993) computed from a larger database of structures. As can be seen from the stem plots in Fig. 1, the three indices are substantially identical for most residues, but with r_{opt} values significantly different from both ASH indices for Phe, Gly, Thr, and Trp. However, as seen below, these differences are quite significant in terms of the structural discriminatory power of the indices.

TABLE 1 First three eigenvectors of $S_B r = \lambda S_W r$

		$r_1 = r_{\mathrm{opt}}$ $\lambda_1 = 0.260$	r_2 $\lambda_2 = 0.162$	r_3 $\lambda_3 = 0.147$
1.	Ala(A)	0.0450	−0.0772	0.0189
2.	Cys(C)	0.2230	0.9042	−0.8108
3.	Asp(D)	−0.2718	−0.0892	−0.0017
4.	Glu(E)	−0.2373	−0.1837	−0.1453
5.	Phe(F)	0.2804	−0.0179	0.0924
6.	Gly(G)	−0.2061	0.1863	0.4754
7.	His(H)	−0.1165	−0.0107	0.0412
8.	Ile(I)	0.4045	−0.0844	0.0932
9.	Lys(K)	−0.2028	−0.1475	−0.0926
10.	Leu(L)	0.3335	−0.1172	0.0428
11.	Met(M)	0.1833	−0.0715	0.0557
12.	Asn(N)	−0.2314	−0.0302	0.0322
13.	Pro(P)	−0.2051	0.0270	0.1710
14.	Gln(Q)	−0.1791	−0.1443	−0.1027
15.	Arg(R)	−0.1515	−0.1404	−0.0989
16.	Ser(S)	−0.1687	0.0086	0.0482
17.	Thr(T)	−0.0505	−0.0317	0.0258
18.	Val(V)	0.3651	−0.0552	0.0770
19.	Trp(W)	0.0654	0.0744	0.0338
20.	Tyr(Y)	0.1206	0.0007	0.0443

The second and third eigenvectors, which in principle represent the next two most structurally discriminative indices after r_{opt}, correlate at significantly lower levels with known indices. The second eigenvector shows a maximum absolute correlation of 0.795 with the index ''average relative fractional occurrence in EL(i)'', which is related to backbone geometry as described in Rackovsky and Scheraga (1982). The third eigenvector correlates most strongly (0.739) with an index representing residue frequencies at N''' helix capping positions (Aurora and Rose, 1998). Most of the remaining lower eigenvectors exhibit maximum

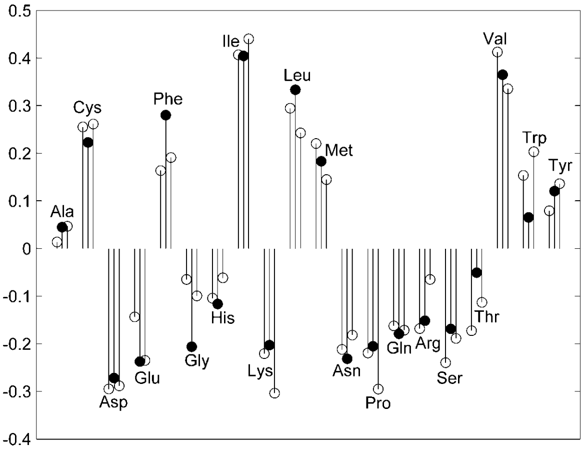


FIGURE 1 Centered and normalized indices: ASH 1978 (*left* ○), r_{opt}(●), and ASH 1993 (*right* ○).

absolute correlations below 0.6 with the entries in AAindex, and have relatively low discriminatory power. Notably, in both the second and third eigenvectors, the index value for Cys dominates the values for other residues. This possibly reflects the special structural role, outside of the context of hydrophobicity, of Cys in disulfide bonds, and a tendency toward conservation of such positions within at least some CE fold families. Similar anomalous behavior of Cys is observed in the Thomas-Dill contact potentials (Thomas and Dill, 1996), particularly for Cys-Cys interactions.

Fig. 2 shows the leave-one-out cross-validation correct classification rates as a function of the merit function values *J(r)* for a variety of indices *r* that cover the entire merit range. Included are several members of the hydrophobicity class, namely $r_{\rm opt}$ (EIG1, the first eigenvector), the Manavalan and Ponnuswamy (1978) ASH index (ASH), the Fauchere and Pliska (1983) hydophobicity index (FP-HP) used by Bowie et al. (1990) in their fold recognition procedure, a consensus hydrophobicity index (CONS-HP) due to Eisenberg et al. (1984), and the Jones (1975) hydrophobicity index (JONES-HP) that is the basis for ASH. Also included are results for a selection of eigenvectors $r_{\rm i}$ (EIGi) for several values of *i*, with associated merit function value $J(r_{\rm i}) = \lambda_{\rm i}$ for each eigenvector/eigenvalue pair ($r_{\rm i}$, $\lambda_{\rm i}$).

The monotonically increasing CV classification performance with $J(r) = S_{\rm B}(r)/S_{\rm W}(r)$validates the selection of the ratio of between-class variation to within-class variation as a merit function. Also, the optimal (first) eigenvector is seen to strongly outperform the second eigenvector, with relatively small performance differentials between successive lower eigenvectors. Thus the first eigenvector $r_{\rm opt}$ defines a particularly distinguished amino acid index with respect to structural discriminatory power, and the discriminatory power of this index is far superior to all other indices in the remaining $S_{\rm W}$-orthogonal subspace of vectors, which are typically nearly uncorrelated with $r_{\rm opt}$. There is surprisingly large performance variability within the hydrophobicity class of indices, with performance generally following the increase in correlation with $r_{\rm opt}$ from 0.628 for the JONES-HP index, 0.764 for CONS-HP, 0.849 for FP-HP, to 0.959 for ASH.

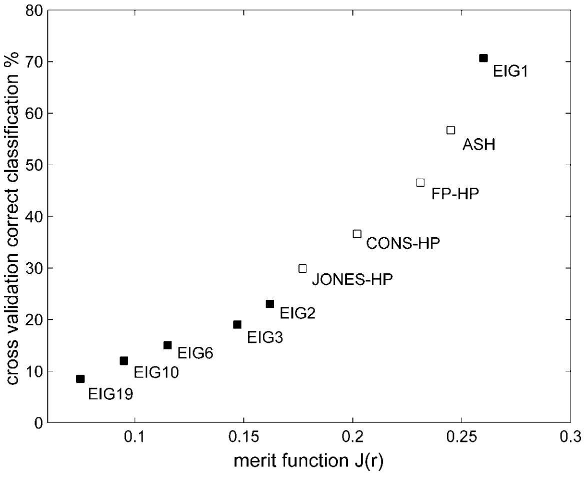


FIGURE 2 FoldID cross-validation correct classification performance for various indices (see text for identifications).

Table 2 summarizes the classification performance of $r_{\rm opt}$ for the SC, CV, and HO and performance validation tests, when the class libraries are restricted to the first (most populous) 10, 20, and 30 folds as well as the full 174-fold case. As seen from the table, the correct classification rates conform to the expected order HO < CV < SC, where HO and SC are known to have pessimistic and optimistic biases, respectively. However, in all cases, the gap between the HO and CV rates is quite small, generally at most a few percent, suggesting that the pessimistic bias of HO is relatively small. In the discussions that follow, we discuss performance primarily in terms of the CV rate, with the understanding that the HO rate is only marginally lower.

The nearest centroid rule based on $r_{\rm opt}$ has an overall average success rate of 70.7% in identifying the correct fold out of 174 candidate folds in the CV test over all 3876 sequences. This rate increases to over 90% as the number of folds is reduced to 10, reflecting both the reduced difficulty of the classification problem with fewer classes, as well the improved statistical base within each class as the smaller classes are removed. Similarly, the relatively large gap between the training set rate of 85.5% and CV rate of 70.7% for all 174 folds falls by roughly a factor of 5 as the training set is cut to only the 10 most populous folds. This is not surprising, because for the 100 least populous folds on average a single omitted sequence represents 8.3% of the total class population, whereas for the first 10 folds it averages 1.2%.

The overall CV successful classification rate of >70% for the full 174-class problem is remarkable, given the large number of classes. For >20 of the fold classes, the correct classification rate is extremely high (>98%). As seen below, the members of such classes typically have several distinctive, conserved structural features. Also, for all 174 folds, even when a misclassification is made, the rank orderings of centroid-test profile distances usually include the true class centroid among the more likely candidates. For example, the centroid of the true class lies among the closest three centroids at a >80% rate, and among the closest 10 at a >90% rate.

The CV correct classification rates for the first 20 folds (for the 20-class problem) are listed in Table 3, along with

TABLE 2 Correct classification rates of $r_{\rm opt}$ for various performance tests

Folds	Total sequences	SC success rate (%)	CV success rate (%)	HO success rate (%)
1–10	829	93.2	90.2	88.7
1–20	1292	87.7	83.3	80.6
1–30	1669	87.5	82.3	81.0
All (1–174)	3876	85.5	70.7	70.2

TABLE 3

Fold	CV correct class %	Near neighbors	Fraction α
1	84	3	0.00
2	77	2	0.36
3	58	13	0.82
4	100	1	0.46
5	97	1	0.33
6	99	0	0.49
7	90	1	0.30
8	85	3	0.32
9	75	3	0.30
10	78	3	0.67
11	81	1	0.35
12	100	3	0.27
13	61	5	0.81
14	83	1	0.18
15	91	2	0.34
16	100	3	0.35
17	77	5	0.42
18	74	2	0.36
19	100	3	0.32
20	77	3	0.33

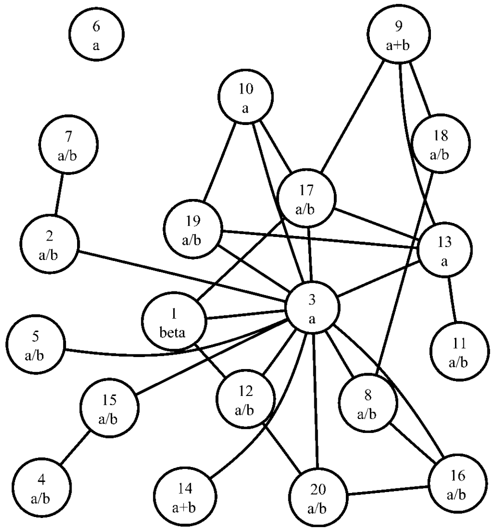
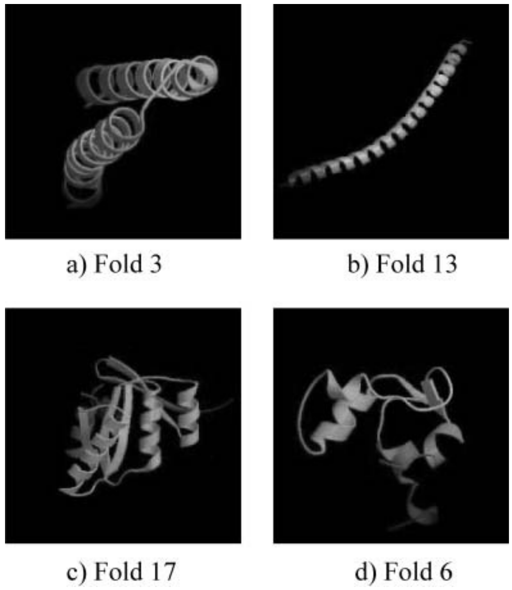


FIGURE 3 Near-neighbor relationships for folds 1–20.

the number of near neighbor fold classes defined by a RMS centroid-centroid distance of <0.115 (compared to a mean RMS distance over all centroid pairs of 0.134), and the mean fraction of α-helical secondary structure composition. There is a large variation in correct classification rates from class to class, ranging from a low of 58% for fold 3 (which falls to 26% for the full 174-class problem) to 100% for folds 4, 12, 16, and 19 (all of which remain >95% for the 174-class problem). This variability is due to a number of factors, but can be related most directly to the distance relationships between class centroids represented in Fig. 3 using the projection method of Gansner and North (2000). Arcs have been inserted between class nodes to indicate a near neighbor relationship defined by an RMS distance threshold of 0.115. The most highly connected folds in terms of near neighbors are 3, 13, and 17, which are also the folds with some of the highest error rates. At the other extreme, fold 6 has no near neighbors and one of the lowest error rates.

Fold 3 has the largest number of near neighbors, smallest mean RMS distance from its centroid to the centroids of other classes, and highest classification error rate. Similarly, fold 13 has the second largest number of near neighbors, second smallest mean distance to other class centroids, and second highest error rate. Folds 3 and 13 are distinguished among the first 20 folds as being the only folds composed almost entirely of α-helical structures (Fig. 4, *a* and *b*). Because of the sequence degeneracy of α-helices drawn from unrelated proteins, the centroids of such folds tend to have small norms and the individual profiles show relatively few consistently positioned distinguishing features. Consequently, the centroid-centroid distances involving these classes are usually smaller than average and the successful classification rates relatively poor. Essentially all folds with similarly high fractions of α-helical secondary structure in the full set of 174 classes display error rates that are among the highest observed for the 174-fold problem. (Ribbon diagrams for all folds are available at the CE3291 4 Å Common Subdomain Library website http://cl.sdsc.edu/subdomains/gal_dc4_new/gal_dc.html.) Fold 17 (Fig. 4 *c*) has an equally high connectivity to that of fold 13, but this is balanced by a larger number of distinguishing features in its three-layer $\alpha/\beta/\alpha$ sandwich architecture, leading to considerably improved (but still below average) FoldID performance.



a) Fold 3 b) Fold 13

c) Fold 17 d) Fold 6

FIGURE 4 The poor classification performance of FoldID for predominantly α-helical folds 3 (*a*) and 13 (*b*) contrasts with the fair performance for $\alpha/\beta/\alpha$ sandwich fold 17 (*c*) and excellent performance for EF-hand fold 6 (*d*).

In contrast, the centroid for fold 6 (Fig. 4 *d*) has no near neighbors within the cutoff distance, and fold 6 is readily discriminated by the optimal FoldID classifier. This result is consistent with the highly distinctive geometry of the EF-hand domain displayed by all members of fold 6, which contains two tandem copies of two helices joined by a calcium-binding loop. Thus several structural features are highly conserved within fold 6, giving rise to relatively consistently positioned corresponding profile features such as minima and maxima outside of the α-helical regions.

This fold 6 example is one of the cases in which the CE common substructure assignment agrees with the expert-determined taxonomy of SCOP (Murzin et al., 1995) at the superfamily level, which classifies all members of CE fold 6 as members of the EF-hand superfamily. Note, however, that the top-level SCOP classes (all-α, all-β, $\alpha + \beta$, α/β) given in Fig. 3 for each of the first 20 CE folds are primarily defined by secondary structure content and topology, and are basically independent of any considerations of alignment. Thus SCOP class definitions bear little relationship to CE structural classes, which are defined by similarity of backbone geometry and are fundamentally tied to structural alignment. Other than the difficulty that FoldID has with all-α folds with very high percentages of α-helical content (as opposed to all-α folds such as fold 6, which has an α-helical content below 50%, but also a rather large loop content that is ignored in making the all-α SCOP classification), we have found no clear connection between FoldID performance and top-level SCOP classification.

The sequence degeneracy of α-helical regions becomes apparent when multiple profiles from a given class are plotted together. The smoothed (by threefold convolution with a box of length 3 centered at each position) sequence profiles for all 67 members of fold 6 are shown in Fig. 5, along with the smoothed centroid. Regions of greatest variability and decoherence in the profile ''waveform'' typically correspond to α-helical secondary structure, whereas there is a much greater degree of coherence and consistent positioning of features such as minima and maxima in the nonhelical regions. Similarly, Fig. 6 shows the almost completely decoherent profiles for the first 50 smoothed fold 3 sequences, which are entirely α-helical with the exception of the turn in the center. The profile centroid for fold 3 is nearly featureless and uniformly close to zero, whereas that for fold 6 has a much greater norm and richer feature set.

DISCUSSION

The key idea of FoldID is the application of a mathematical optimization procedure to find an optimal index r_{opt} in the vector space R^{20} of all possible indices. The optimal index maps sequences to numerical profile vectors, and fold assignments are based on a nearest centroid fold classification rule in the numerical profile vector space. The merit function $J(r) = r^{\mathrm{T}} S_{\mathrm{B}}\, r / r^{\mathrm{T}} S_{\mathrm{W}}\, r$ represents the ratio of quadratic measures of between-class to within-class variation, and is optimized by the maximal eigenvector of the generalized eigensystem $S_{\mathrm{B}} r = \lambda S_{\mathrm{W}} r$. This is superficially similar to the approach used in Multiple Discriminant Analysis (MDA) as described in Duda and Hart (1973) for multiclass pattern classification, and we have adopted similar notation where appropriate. However, the context, meaning, and underlying mathematical structure of the corresponding optimization problems are fundamentally different. In the MDA case, the original patterns are already

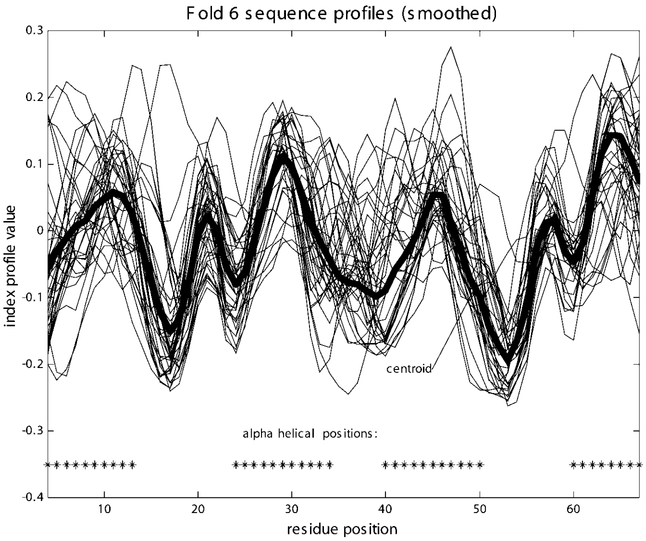


FIGURE 5 Smoothed fold 6 profiles are most coherent in non-α-helical regions, where features such as minima and maxima in the centroid profile (*thick curve*) are conserved in many individual profiles.

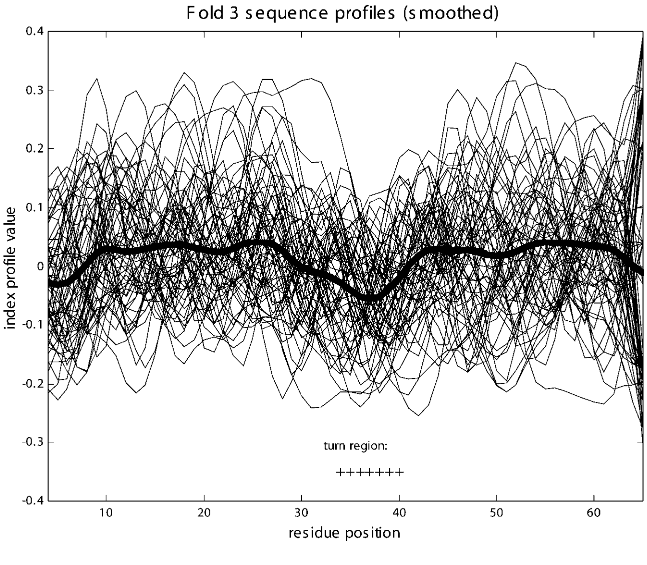


FIGURE 6 Smoothed fold 3 profiles show decoherence typical of predominantly α-helical secondary structure. Centroid profile (*thick curve*) is essentially featureless except for shallow minimum in turn region.

represented as numerical vectors, and the solution to the optimization problem represents a linear function of the vector components that is useful in formulating a classifier. In particular in the two-class case, there is only one meaningful eigenvector because the matrix S_B has rank one, and the corresponding function is a classifier, namely the Fisher linear discriminant. In the FoldID case, the solution represents not a classifier but an optimal mapping from sequences to numerical profile vectors before classification. The matrix S_B is of nearly full rank, leading to many meaningful eigenvectors even in the two-class case, and the necessity for using a more general eigenproblem solution methodology.

The availability of secondary eigenvectors opens the possibility of using a more complex classification rule based on two or more profiles corresponding to different eigenvectors. We have investigated a nearer weighted centroid rule that selects the fold with the smallest value of $w_1||Pr_1 - c_J(r_1)|| + w_2||Pr_2 - c_J(r_2)||$. Using weights $w_i = \lambda_i$, the CV correct classification performance improves from 70.7% for r_{opt} alone to 73.0% for the two-index rule. Larger improvements may be possible when there is not a single dominant eigenvector and lower eigenvectors have more discriminatory power than is the case here.

Correlation with indices in AAindex shows that r_{opt} is most similar to the Average Surrounding Hydrophobicity index of Manavalan and Ponnuswammy (1978), but r_{opt} shows considerably better classification performance. The ASH index also has the highest predicted CV performance of all indices in the AAaindex database based on its merit function value, and we have verified that the CV classification performance of r_{opt} is significantly better than all other AAindex indices. Because the eigenvectors of $S_B r = \lambda S_W r$ span all possible indices of interest, and r_{opt} is the global optimum of the merit function over that space of indices, it is reasonably certain that there can be no indices based on known or as yet undiscovered physiochemical or statistical properties of amino acids that can significantly outperform r_{opt} in this specific context. This assumes that the observed monotonic relationship between merit function and actual classification performance remains valid over the entire space, and, of course, does not preclude the possibility of another index outperforming r_{opt} in a different algorithm based on index profiles.

The ASH index is closely related to packing density as measured by the ‘‘8-Å contact number’’ index constructed in Nishikawa and Ooi (1980) as a database average of the number of residues within an 8-Å radius of a given residue type. Thus ASH, which sums Jones hydrophobicities (Jones, 1975) over the same sphere, can be regarded as a hybrid of the underlying Jones index and the Nishikawa-Ooi contact number. The r_{opt} index is much more highly correlated with Nishikawa-Ooi contact number (0.903) than Jones hydrophobicity (0.628) or the other hydrophobicity indices considered here. Thus it is probably more appropriate to associate r_{opt} with packing density than hydrophobicity. We note that Bagci et al. (2002) have recently pointed out possible correlations among packing density, sequence conservation, and folding nucleation.

Unlike ASH, r_{opt} is completely independent of any experimental determinations of physiochemical amino acid properties because it is derived directly from the structural classification information inherent in the training library. Moreover, there is a compelling universality to r_{opt} in that it does not appear to depend strongly on the specific fold classes used to compute it. For example, optimal indices based on randomly selected sets of 10 fold classes typically correlate at the 0.98 or higher level with r_{opt}. Similarly, the 45 optimal index vectors generated from application of FoldID to all possible fold pairs selected from folds 1 through 10 (the ten most highly populated folds) have an average 0.92 correlation with r_{opt}. Even severely biasing the fold library toward specific secondary structure compositions has relatively little effect. All 174 folds were divided into disjoint α, mixed α-β, and β sublibraries. The optimal index trained exclusively to separate the predominantly α-folds from each other was correlated with the overall r_{opt} trained on all folds at the 0.974 level. Similarly the optimal index trained on the mixed α-β folds correlated with r_{opt} at 0.991. The optimal index based on predominantly β-folds had a somewhat lower, but still quite strong, correlation of 0.902.

Finally, it must be emphasized that FoldID addresses only the fold assignment problem, using training and test sequences that have already been optimally aligned. To develop a procedure to identify likely occurrences of the CE structures in a new, unaligned target sequence, it will be necessary to couple FoldID to an alignment procedure. A natural choice is to align the target sequence profile of r_{opt} to each of the fixed profile centroids $c_J(r_{opt})$ using a standard Smith-Waterman dynamic programming local alignment algorithm (Smith and Waterman, 1981). The alignment score is based on distance of the aligned sequence to the centroid $c_J(r_{opt})$ and appropriate gap penalties in the aligned sequences (the profile centroids are held fixed with no gaps or deletions). For any input target sequence S, the separate alignments of S to $c_J(r_{opt})$ will produce optimal aligned sequences S_J^* and corresponding z-scores relative to random alignments. Scores exceeding a suitable threshold will then identify which, if any, of the S_J^* are likely candidates for assignment to fold J. The method is currently being implemented and tested, and is described in detail by J. B. Rosen, R. H. Leary, P. Jambeck, J. Glick, and J. Chodera (unpublished data, 2003).

CONCLUSION

FoldID, a supervised learning algorithm for the protein fold assignment problem, has been applied to 3876 training sequences drawn from 174 structurally aligned fold classes with low sequence identity. For any given amino acid index, the profile vector of a sequence is obtained by substituting the corresponding index value for each residue type. Each amino acid index defines a simple classification rule based on the closest match of a test sequence profile to the various class profile centroids. An optimal index that maximizes the ratio of between-class to within-class profile variation in the training set was computed as the solution to a generalized eigenvalue problem. The performance of the corresponding classification rule was evaluated by a cross-validation test and compared to that of rules corresponding to known indices in the literature. The classification performance of the optimal index computed as the solution to the eigenproblem was found to be significantly higher than that of any known index, with a 70.7% successful classification rate for all sequences and, in several cases, a nearly 100% success rate for sequences from folds with well-defined, conserved structural features. Among known indices, the optimal index most closely resembles average surrounding hydrophobicity, and exhibits a striking universality in that it is relatively insensitive to variations in the selection the specific structural classes used in its computation.

We are indebted to P. Bourne, K. Dill, and S. Subramaniam for useful discussions and comments on the manuscript.

R. H. Leary acknowledges support from National Science Foundation/Division of Biological Infrastructure (grant 0077130), and J. B. Rosen from National Science Foundation/Information Technology Research (grant 0082146).

REFERENCES

Anderson, E., Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. 1999. LAPACK User's Guide, 3rd ed. SIAM, Philadelphia, PA.

Aurora, R., and G. D. Rose. 1998. Helix capping. *Protein Sci.* 7:21–38.

Bagci, Z., R. L. Jernigan, and I. Bahar. 2002. Residue packing in proteins: uniform distribution on a coarse-grained scale. *J. Chem. Phys.* 116:2269–2276.

Bahar, I., A. R. Atilgan, R. L. Jernigan, and B. Erman. 1997. Understanding the recognition of protein structural classes by amino acid composition. *Proteins.* 29:172–175.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The protein data bank. *Nucleic Acids Res.* 28:235–242.

Bowie, J. U., N. D. Clarke, C. O. Pabo, and R. T. Sauer. 1990. Identification of protein folds—matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins.* 7:257–264.

Cai, Y. D., X. J. Liu, X. B. Xu, and K. C. Chou. 2003. Support vector machines for prediction of protein domain structural class. *J. Theor. Biol.* 221:115–120.

Chou, K. C. 1995. A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. *Proteins.* 21:319–344.

Ding, C. H., and I. Dubchak. 2001. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics.* 17:349–358.

Duda, R. O., and P. E. Hart. 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Eisenberg, D., E. Schwarz, M. Komaromy, and R. Wall. 1984. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* 179:125–142.

Fauchere, J. L., and V. Pliska. 1983. Hydrophobic parameters-pi of amino-acid side chains from the partitioning of n-acetyl-amino-acid amides. *Eur. J. Med. Chem.* 18:369–375.

Fischer, D., and D. Eisenberg. 1996. Protein fold recognition using sequence-derived predictions. *Protein Sci.* 5:947–955.

Gansner, E. R., and S. C. North. 2000. An open graph visualization system and its applications to software engineering. *Software: Pract. And Exp.* 30:1203–1233.

Jones, D. D. 1975. Amino acid properties and side-chain orientation in proteins: a cross correlation approach. *J. Theor. Biol.* 50:167–183.

Kawashima, S., and M. Kanehisa. 2000. AAindex: amino acid index database. *Nucleic Acids Res.* 28:374.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *In* Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95. Morgan Kaufmann, San Francisco, CA. 1137–1145.

Mallick, P., R. Weiss, and D. Eisenberg. 2002. The directional atomic solvation energy: an atom-based potential for the assignment of protein sequences to known folds. *Proc. Natl. Acad. Sci. USA.* 99:16041–16046.

Manavalan, P., and P. K. Ponnuswamy. 1978. Hydrophobic character of amino acid residues in globular proteins. *Nature.* 275:673–674.

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–540.

Nishikawa, K., and T. Ooi. 1980. Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int. J. of Pept. Protein Res.* 16:19–32.

Ponnuswamy, P. K., and M. M. Gromiha. 1993. Prediction of transmembrane helices from hydrophobic characteristics of proteins. *Int. J. of Pept. Protein Res.* 42:326–341.

Rackovsky, S., and H. A. Scheraga. 1982. Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids. *Macromolecules.* 15:1340–1346.

Shindyalov, I. N., and P. E. Bourne. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 11:739–747.

Shindyalov, I. N., and P. E. Bourne. 2000. An alternative view of protein fold space. *Proteins.* 38:247–260.

Smith, T. S., and M. S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195–197.

Thomas, P. D., and K. A. Dill. 1996. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. USA.* 93:11628–11633.

Tomii, K., and M. Kanehisa. 1996. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* 9:27–36.