Featureless Pattern Recognition in an Imaginary Hilbert Space and its Application to Protein Fold Classification

Vadim Mottl[1], Sergey Dvoenko[1], Oleg Seredin[1],
Casimir Kulikowski[2], and Ilya Muchnik[2]

[1] Tula State University, Lenin Ave. 92, 300600 Tula, Russia
mottl@atm.tsu.tula.ru
[2] Rutgers University, P.O. Box 8018, Piscataway, NJ 08855, USA
kulikows@cs.rutgers.edu

Abstract. The featureless pattern recognition methodology based on measuring some numerical characteristics of similarity between pairs of entities is applied to the problem of protein fold classification. In computational biology, a commonly adopted way of measuring the likelihood that two proteins have the same evolutionary origin is calculating the so-called alignment score between two amino acid sequences that shows properties of inner product rather than those of a similarity measure. Therefore, in solving the problem of determining the membership of a protein given by its amino acid sequence (primary structure) in one of preset fold classes (spatial structure), we treat the set of all feasible amino acid sequences as a subset of isolated points in an imaginary space in which the linear operations and inner product are defined in an arbitrary unknown manner, but without any conjecture on the dimension, i.e. as a Hilbert space.

1 Introduction

The classical pattern recognition theory deals with objects represented in a finite-dimensional space of their features that are assumed to be defined in advance, before real objects subject to classification are observed. The emphasis on the feature-based representation of objects is reflected in the name of the most popular method of machine learning for pattern recognition called the support vector method [1,2].

At the same time, there exists a wide class of applications in which it is easy to evaluate some numerical characteristics of pairwise relationship between any two objects, but it is hard to indicate a set of rational individual attributes of objects that could form the axis of a feature space.

As an alternative to the feature-based methodology, R. Duin and his colleagues [3,4,5] proposed a featureless approach to pattern recognition, in which objects are assumed to be represented by appropriate measures of their pairwise similarity or dissimilarity. It is just this idea we use here as a basis for creating techniques of protein fold class recognition, i.e. allocating a protein, given by the primary chemical structure of its polymerous molecule as a sequence of amino acids (to be exact, their residues) from the alphabet of 20 amino acids existing in nature, over a finite set of typical spatial structures, each associated with a specific manner in which the primary amino acid chains fold in space under a highly complicated combination of numerous

This work is supported, in part, by the Russian Foundation of Basic Research, Grant No. 99-01-00372, and the State Scientific Program of the Russian Federation "Promising Information Technologies".

physical forces [6,7]. We lean here upon the compactness hypothesis that is understood as the tendency of proteins with "similar" amino acid chains to belong to the same fold class [8].

It is common practice in computational biology to measure the proximity between two amino acid chains as the logarithmic likelihood ratio of two hypotheses, the main hypothesis that both of them originate from the same unknown protein as result of independent successions of local evolutionary mutations versus the null hypothesis that the chains are completely occasional combinations over the alphabet of 20 amino acids [9]. The generally accepted way of measuring such a likelihood ratio is calculating the so-called alignment score between two amino acid sequences, which is based on finding an appropriate consensus sequence from which both sequences might be obtained as result of as a small number of local corrections as possible, namely, deletions, insertions and substitutions of single amino acids [10,11].

By its nature, the logarithmic likelihood ratio may take as positive as well negative values. In addition, such a ratio calculated for an amino acid sequence with itself gives different values for different proteins. As a result, it is hard to interpret the pairwise alignment score as a similarity measure. In this work, we pose the heuristic hypothesis that the set of all feasible amino acid sequences may be considered as a subset of isolated points in an imaginary Hilbert space in which the linear operations are defined in an arbitrary unknown manner, and the role of inner product is played by the alignment score between the respective pair of amino acid chains.

Such an assumption allows for treating the sought-for decision rule of pattern recognition by the principle "one class against another one" as a discriminant hyperplane immediately in the Hilbert space of objects. However, the absence of coordinate axes prevents from finding the "direction element" of the hyperplane, i.e. an element of the Hilbert space that splits all the space points into two nonintersecting regions by values of scalar products with it.

Therefore, we propose to use an assembly of selected "representative" objects as a basis in the Hilbert space of all the feasible objects. The elements of the basic assembly are not assumed to be classified, their mission is to serve as coordinate axes of a finite-dimensional subspace, onto which any new object, including those forming the classified training sample, could be projected by calculating inner products with the basic elements.

The idea of making distinction between the unclassified basic assembly and classified training sample appears to be quite reasonable for the problem of protein fold class recognition, because the number of proteins whose spatial structure is known is much less than the number of proteins with known amino acid chains.

2 The problem of protein fold class recognition

The problem of finding the spatial structure of a protein represented by its primary amino acid sequence is a challenge posed by the nature. On the one hand, the necessity of such algorithms is dictated by the fact that application of usual physical techniques of magnetic resonance and X-ray analysis is problematic in most cases. Although the number of proteins whose spatial structure is known ever grows, the gap between the number of known amino acid sequences and that of known spatial structures is increasing dramatically. On the other hand, the "existence theorem" is proved by nature itself, because it has been never observed that an amino acid chain had more than one spatial structure.

Each protein has its specific spatial organization which does not coincide with that of any other protein. The main principle of establishing the spatial structure of a given

protein from its amino acid chain consists in finding, for the given chain, the most appropriate structure from a bank of known structures and their fragments. For each amino acid residue in the chain forming a protein of a known structure, the vector of some quantitative features is evaluated which are assumed to be responsible for the spatial position of this residue in the three-dimensional structure. The succession of such features along the amino acid chain is called the profile of this structure. The same features are evaluated for the amino acid chain of the new protein, whereupon the succession obtained is compared with profiles of known structures by alignment of positions in this succession and in the respective profile with respect to eventual insertions and deletions. Such a principle named threading [6] is fraught with enumeration of a large number of known structures.

Despite the uniqueness of the spatial structure of each protein, it is the usual case that large groups of evolutionary allied proteins have very similar spatial structures. In this sense, there exist "much less" spatial structures than primary ones. Of course, the classification of spatial structures is a problem which is not simple, but once a version of classification is accepted, the problem of assigning an amino acid chain to a class of spatial structures falls into the competence area of pattern recognition.

In an earlier series of experiments [7], an attempt was made to describe the primary amino acid sequence of a protein by vector of its numerical features and consider it as a point in the respective linear vector space. In particular, the primary structure of a protein was represented by frequencies with which amino acids of the polar, neutral and hydrophobic type and their pairs occur in it.

The results of those experiments cannot be assessed as quite successful, to all appearance, because of an immensely rich actual diversity of amino acid properties that may play an important part in forming the spatial structure of a protein. Therefore, we turn here to the featureless formulation of the fold class recognition problem.

When studying the structure and properties of proteins, one of commonly used instruments is the characteristic of mutual similarity of two amino acid sequences $\omega' = (a_1,\ldots,a_N)$ and $\omega'' = (b_1,\ldots,b_K)$ given by an appropriate pair-wise alignment procedure (Fig. 1). Procedures of such a kind lean upon a preset similarity matrix for all 210 pairs of 20 amino acids. Such matrices are called substitution matrices and characterize each amino acid pair (a,b) by logarithmic ratio of, first, the probability of their independent occurrence in two amino acid chains p_{ab} as result of evolutionary substituting the same unknown amino acid c in a common ancestor chain, and, second, the product of general probabilities q_a and q_b of their occurrence in arbitrary sequences [9]:

$$s(a,b) = \log(p_{ab}/q_a q_b)\,. \tag{1}$$

The log likelihood ratio $s(a,b)$ is positive if the probability that these two amino acids have a common ancestor is greater than the product of their general probabilities, equals zero in the indifferent case, and is negative if the hypothesis of their common origin is less likely than that of the null hypothesis of their independent occasional appearance.

There are several versions of substitution matrices [9,12,13], but each of them is result of observations in large sets of proteins aligned in that or other manner by experienced biologists in accordance with their intuition based, in its turn, on that or other model of evolution.

The numerical measure of the proximity of two proteins represented by their amino acid chains is determined as the greatest possible sum of $s(a_{j_i}, b_{k_i})$ over all related

pairs of amino acids (j_i, k_i), $i = 1,2,3,\ldots$, in a pair-wise alignment with respect to some penalties posed on the presence and length of gaps (Fig. 1):

$$\mu(\omega', \omega'') = \sum_i s(a_{j_i}, b_{k_i}) - (\text{gap length penalties}). \quad (2)$$

In our experiments we used this similarity measure of amino acid chains measured by the commonly adopted alignment procedure Fasta 3 [10,11] with substitution matrix Blossum 50 [9].

As the set of experimental data, we took the collection of proteins selected by Dr. Sun-Ho Kim from Lawrence Berkley National Laboratory in the USA. The collection contains 396 protein domains, i.e. relatively isolated fragments of amino acid chains, chosen from the SCOP Database (Structural Classification of Proteins). The protein domains forming the collection belong to 51 fold classes listed in Table 1. The principle of selection was to provide a low similarity of amino acid sequences within each family, with which purpose only those protein domains were chosen whose similarity (2) to other selected domains did not exceed a preset threshold. Such a principle of selection resulted in protein domain families of different size.

3 The pair-wise alignment score of two amino acid chains as their inner product in an imaginary Hilbert space

It appears natural to interpret the log likelihood ratio for two amino acids $s(a,b)$ (1) as experimentally registered outward exhibition of the actual proximity of their hidden properties. Let these properties be expressed by some hidden vectors $\mathbf{y}_a$ and $\mathbf{y}_b$ for which the notion of inner product is defined $(\mathbf{y}_a, \mathbf{y}_b)$, then the structure of (1) suggests the idea to consider $s(a,b)$ as a rough measure of it: $s(a,b) \cong (\mathbf{y}_a, \mathbf{y}_b)$.

By analogy to a single summand, the score of the alignment as a whole (2) may also be interpreted as inner product of the respective combined feature vectors of two proteins $\mu(\omega', \omega'') \cong (\mathbf{x}_{\omega'}, \mathbf{x}_{\omega''})$ in an imaginary linear feature space. The greater the positive value of the similarity, the more "synchronous" are some essential properties of amino acids along the polypeptide chain, the zero value says about full lack of agreement what corresponds to the notion of orthogonality, and a negative value should be interpreted as "opposite phases" of amino acid properties along the chains.

However, this is not more than a cursory analogy. For an accurate justification of the hypothesis that there exists a Hilbert space in which the set of proteins could be embedded, we should show that the score matrix of any finite assembly of proteins tends to be nonnegative definite or, at least, can be approximated by such a matrix.

Table 1. Dr. Kim's collection of proteins.

	Fold class	Size		Fold class	Size
1	Globin	12	27	Flavodoxin	9
2	Cytochrome C	7	28	Adenine nucleotide alpha hydroclase	4
3	Four-helical bundle	8	29	Rossmann-fold domains	14
4	Ferritin	8	30	Thiamin-binding	3
5	4-gelical cytokines	11	31	P-loop containing NTP hydrolases	9
6	EF Hand	13	32	Thioredoxin fold	9
7	Cyclin	4	33	Restriction endonucleases	5
8	Cytochrome P450	5	34	Ribonuclease H motif	9
9	Immunoglobin beta – sandwich	31	35	Phosphoribosyltransferases (PRTases)	3
10	Common fold of difteria toxin / transcription factors / cytochrome	5	36	S-adenosyl-L-methionine-dependent methyltransfe-rases	5
11	Cupredoxins	9	37	Alpha / beta-Hydrolases	12
12	C2 domain	3	38	Phosphorylase / hydrolase	5
13	Viral coat and capsid proteins	15	39	Periplastic binding protein I	7
14	Crystallins / protein S / yeast killer toxin	5	40	Periplastic binding protein II	7
15	Galastose-binding domain	4	41	Lysozyme	4
16	ConA lectins / glucanases	8	42	Cysteine proteinases	4
17	OB-fold	17	43	Beta-Grasp	8
18	Beta-Trefoil	5	44	Cystatin	7
19	Reductase / isomerase / elongation factor	4	45	Ferredoxin	20
20	Trypsin serine proteases	6	46	Zincin	7
21	Acid proteases	5	47	N-terminal nucleophile aminohydrolases	4
22	PH domain	7	48	ADP-ribosylation	4
23	Lipocalings	6	49	C-type lectin	6
24	Double-stranded beta-helix	6	50	Protein kinases (PK), catalytic core	4
25	Barrel-sandwich hybrid	6	51	Beta-Lactamase / D-ala carboxypeptidase	3
26	TIM-barrel	28			

```
ω′ : TNPGNASSTTTTKPTTTS-----RGLKTINETDPCIKNDSCTG
ω″ : GS----ATSTPATSTTAGTKLPCVRNKTDSNLQSCNDTIIEKE
 i = 12    34567  ···
```

Fig. 1. Fragment of an aligned pair of amino acid chains from the protein family *Envelope glycoprotein GP120* in the database Pfam.

We checked this hypothesis for an assembly of 396 proteins (Table 1) by way of calculating all the eigenvalues of the score matrix obtained by the pair-wise alignment procedure Fasta 3 [10,11] with the substitution matrix Blossum 50 [9]. All the eigenvalues turned out to be positive (Fig. 2).

The conclusion suggests itself that the pair-wise similarity measure determined by the procedure Fasta 3 possesses properties having much in common with those of inner product. This circumstance should be considered as a reason in favor of the theoretical applicability of the principle of featureless pattern recognition in a Hilbert space to the problem of protein fold class recognition.

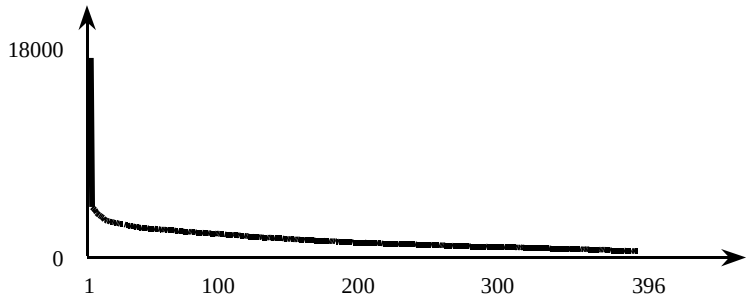

Fig. 2. Eigenvalues of Dr.Kim's collection of proteins: $\lambda_{\max} = 16621$, $\lambda_{\min} = 304$; all eigenvalues are positive.

4 Hilbert space of classified objects and optimal discriminant hyperplane

Let the set Ω of all feasible objects under consideration $\omega \in \Omega$ is partitioned into two classes $\Omega_1 = \{\omega \in \Omega : g(\omega) = 1\}$ and $\Omega_{-1} = \{\omega \in \Omega : g(\omega) = -1\}$ by an unknown indicator function $g(\omega) = \pm 1$. The main idea of the featureless approach to pattern recognition consists in treating the set Ω as a Hilbert space in which the linear operations and inner product are defined in an arbitrary manner under the usual constraints:

(1) addition is symmetric and associative $\omega' + \omega'' = \omega'' + \omega' \in \Omega$, $\omega' + (\omega'' + \omega''') = (\omega' + \omega'') + \omega'''$;

(2) there exists an origin $\phi \in \Omega$ such that $\omega + \phi = \omega$ for any element $\omega \in \Omega$;

(3) there exists the inverse elements $(-\omega) + \omega = \phi$ for any $\omega \in \Omega$;

(4) multiplication by a real coefficient $c\omega \in \Omega$, $c \in \mathsf{R}$, is associative $(cd)\omega = c(d\omega)$ and $1\omega = \omega$ for any $\omega \in \Omega$;

(5) addition and multiplication are distributive $c(\omega' + \omega'') = c\,\omega' + c\,\omega''$, $(c + d)\omega = c\omega + d\,\omega$;

(6) inner product of elements is symmetric $(\omega', \omega'') = (\omega'', \omega') \in \mathsf{R}$ and linear $(\omega, \omega' + \omega'') = (\omega, \omega') + (\omega, \omega'')$, $(\omega, c\,\omega') = c(\omega, \omega')$;

(7) inner product of an element with itself possesses the properties $(\omega, \omega) \geq 0$, $(\omega, \omega) = 0$ if and only if $\omega = \phi$ and gives the norm $||\omega|| = (\omega, \omega)^{1/2} \geq 0$.

It is not meant that all the elements of the Hilbert space Ω do exist in reality. We consider really existing objects as making a subset $\widetilde{\Omega}$ of isolated points in Ω, whereas all the remaining elements are nothing else than products of our imagination. It is just the extension of $\widetilde{\Omega}$ to Ω what allows speaking about "sums" of really existing objects and their "products" with real-valued coefficients.

It is assumed that even if an element of the Hilbert space $\omega \in \Omega$ really exists $\omega \in \widetilde{\Omega} \subset \Omega$, it cannot be perceived by the observer in any other way than through its inner products (ω, ω') with other really existing elements $\omega' \in \widetilde{\Omega} \subset \Omega$. If $\vartheta \in \Omega$ is a fixed element of the Hilbert space, an imaginary one in the general case, the real-valued linear discriminant function $d(\omega \,|\, \vartheta, b) = (\vartheta, \omega) + b$, where $b \in \mathsf{R}$ is a constant, may be used as decision rule $\hat{g}(\omega) : \Omega \to \{1, -1\}$ of judging on the hidden class-membership of an arbitrary object $\omega \in \Omega$, might it really exist or not:

$$d(\omega \,|\, \vartheta, b) = (\vartheta, \omega) + b \begin{cases} > 0 \to \hat{g}(\omega) = 1, \\ < 0 \to \hat{g}(\omega) = -1. \end{cases} \tag{3}$$

Here the element $\vartheta \in \Omega$ plays the role of the direction element of the respective discriminant hyperplane in the Hilbert space $(\vartheta, \omega) + b = 0$.

However, we have, so far, no constructive instrument of choosing the direction element $\vartheta \in \Omega$ and, hence, the decision rule of recognition, because, just as any element of Ω, it can be defined only by its inner products with some other fixed elements that exist in reality.

Let the observer have chosen an assembly of really existing objects $\Omega^0 = \{\omega_1^0, \ldots, \omega_n^0\} \subset \Omega$, called the basic assembly, which is not assumed to be classified, in the general case, and, therefore, it is not yet a training sample. The basic assembly will play the role of a finite basis in the Hilbert space that defines an n-dimensional subspace

$$\Omega_n(\omega_1^0, \ldots, \omega_n^0) = \left\{ \omega \in \Omega : \omega = \sum\nolimits_{i=1}^{n} a_i \omega_i^0 \right\} \subset \Omega. \tag{4}$$

We restrict our consideration to only those discriminant hyperplanes whose direction elements belong to $\Omega_n(\omega_1^0, \ldots, \omega_n^0)$, i.e. can be expressed as linear combinations

$$\vartheta(\mathbf{a}) = \sum_{i=1}^{n} a_i \omega_i^0, \quad \mathbf{a} \in \mathsf{R}^n. \tag{5}$$

The respective parametric family of discriminant hyperplanes $\big(\vartheta(\mathbf{a}), \omega\big) + b = \sum_{i=1}^{n} a_i (\omega_i^0, \omega) + b = 0$ and, so, linear decision rules

$$d\big(\omega \,|\, \vartheta(\mathbf{a}), b\big) = \sum_{i=1}^{n} a_i (\omega_i^0, \omega) + b \begin{cases} > 0 \to \hat{g}(\omega) = 1, \\ < 0 \to \hat{g}(\omega) = -1, \end{cases} \quad \omega \in \Omega, \tag{6}$$

will be completely defined by inner products of elements of the Hilbert space with elements of the basic assembly (ω_i^0,ω), $i=1,\ldots,n$. We shall consider the totality of these values for an arbitrary element $\omega\in\Omega$ as its real-valued "feature vector"

$$\mathbf{x}(\omega)=\left(x_1(\omega)\cdots x_n(\omega)\right)^T\in\mathsf{R}^n,\quad x_i(\omega)=(\omega_i^0,\omega). \tag{7}$$

Mark that if $\left(\vartheta(\mathbf{a}),\omega\right)=0$ then $(\omega_i^0,\omega)=0$ for all $\omega_i^0\in\Omega^0$. This means that by choosing the direction elements in accordance with (5) we restrict our consideration to only those discriminant hyperplanes which are orthogonal to the subspace spanned over the basic assembly of objects. As a result, all elements of the Hilbert space that have the same inner products with basic elements $\mathbf{x}=\left((\omega_1^0,\omega)\cdots(\omega_n^0,\omega)\right)^T$, or, in other words, the same projection on the basic subspace $\Omega_n(\omega_1^0,\ldots,\omega_n^0)$ (4), will be assigned the same class $\hat{g}(\omega)=\pm1$ by linear decision rules (6). Therefore, we call the features (7) projectional features of Hilbert space elements.

We have come to a parametric family of decision rules of pattern recognition in a Hilbert space (6) that lean upon projectional features of objects:

$$d\left(\mathbf{x}(\omega)\,|\,\mathbf{a},b\right)=\mathbf{a}^T\mathbf{x}(\omega)+b\begin{cases}>0\rightarrow\hat{g}(\omega)=1,\\<0\rightarrow\hat{g}(\omega)=-1,\end{cases}\qquad\omega\in\Omega. \tag{8}$$

Thus, the notion of projectional features reduces, at least, superficially, the problem of featureless pattern recognition in a Hilbert space to the classical problem of pattern recognition in a usual linear space of real-valued features.

Let the observer be submitted a classified training sample of objects $\Omega^*=\{\omega_1,\ldots,\omega_N\}\subset\Omega$, $g_1=g(\omega_1),\ldots,g_N=g(\omega_N)$, that does not coincide, in the general case, with the basic assembly $\Omega^0=\{\omega_1^0,\ldots,\omega_n^0\}$. The observer has no other way of perceiving them than to calculate their inner products with objects of the basic assembly, what is equivalent to evaluating their projectional features

$$\mathbf{x}(\omega_j)=\left(x_1(\omega_j)\cdots x_n(\omega_j)\right)^T=\left((\omega_1^0,\omega_j)\cdots(\omega_n^0,\omega_j)\right)^T\in\mathsf{R}^n.$$

Parameters of the discriminant hyperplane $\mathbf{a}\in\mathsf{R}^n$ and $b\in\mathsf{R}$ (8) should be chosen so that the training objects would be classified correctly with a positive margin $\xi>0$:

$$d\left(\omega_j\,|\,\vartheta(\mathbf{a}),b\right)=\sum_{i=1}^{n}a_i(\omega_i^0,\omega_j)+b=\mathbf{a}^T\mathbf{x}(\omega_j)+b\begin{cases}\geq\xi\ \text{ when }\ g(\omega_j)=1,\\\leq-\xi\ \text{ when }\ g(\omega_j)=-1.\end{cases} \tag{9}$$

If the training sample is linearly separable with respect to the basic assembly, there exists a family of hyperplanes that satisfy these conditions. It is clear that the margin ξ remains positive after multiplying the pair $\left(\vartheta(\mathbf{a})\in\Omega,b\in\mathsf{R}\right)$ with a positive coefficient $\left(c\,\vartheta(\mathbf{a})\in\Omega,cb\in\mathsf{R}\right)$, $c>0$, thus, it is sufficient to consider direction elements of a preset norm $||\vartheta(\mathbf{a})||=\left(\vartheta(\mathbf{a}),\vartheta(\mathbf{a})\right)^{1/2}=\text{const}$. One of them, for which $\xi\rightarrow\max$ and the conditions (9) are met, will be called the optimal discriminant hyperplane in the Hilbert space.

Because the direction element of the discriminant hyperplane is determined here by a finite-dimensional parameter vector, such a problem, if considered in the basic subspace $\Omega_n(\omega_1^0,\ldots,\omega_n^0)$ (4), completely coincides with the classical statement of the pattern recognition problem as that of finding the optimal discriminant hyperplane. The same reasoning as in [2] leads to the conclusion that the maximum margin is

provided by choosing the direction element $\vartheta(\mathbf{a}) \in \Omega$ and threshold $b \in \mathsf{R}$ from the condition

$$||\vartheta(\mathbf{a})||^2 \to \min, \quad g_j\left[\left(\vartheta(\mathbf{a}), \omega_j\right) + b\right] \geq 1, \quad j = 1,\ldots,N. \tag{10}$$

However, such an approach becomes senseless in case the classes are inseparable in the basic subspace, and the constraints (9) and, hence, (10) are incompatible. To design an analogous criterion for such training samples, we, just as V. Vapnik, admit nonnegative defects $g_j\left[\left(\vartheta(\mathbf{a}), \omega_j\right) + b\right] \geq 1 - \delta_j$, $\delta_j \geq 0$, and use a compromise criterion $(\vartheta, \vartheta) + C\sum_{j=1}^{N} \delta_j \to \min$ with a sufficiently large positive coefficient C meant to give preference to the minimization of these defects. So, we come to the following formulation of the generalized problem of finding the optimal discriminant hyperplane in the Hilbert space that covers both the separable and inseparable case:

$$\begin{cases} ||\vartheta(\mathbf{a})||^2 + C\sum\limits_{j=1}^{N} \delta_j \to \min, \\ g_j\left(\mathbf{a}^T \mathbf{x}(\omega_j) + b\right) \geq 1 - \delta_j, \quad \delta_j \geq 0, \quad j = 1,\ldots,N. \end{cases} \tag{11}$$

5 Choice of the norm of the direction element

The norm of the direction element of the sought-for hyperplane can be understood, at least, in two ways, namely whether as that of an element of the Hilbert space $\vartheta \in \Omega$ or as the norm of its parameter vector in the basic subspace $\mathbf{a} \in \mathsf{R}^n$. In the former case we have, in accordance with (5),

$$||\vartheta(\mathbf{a})||^2 = \left(\vartheta(\mathbf{a}), \vartheta(\mathbf{a})\right)^2 = \sum_{i=1}^{n}\sum_{l=1}^{n} (\omega_i^0, \omega_l^0) a_i a_l = \mathbf{a}^T \mathbf{M} \mathbf{a}, \tag{12}$$

where $\mathbf{M} = \left((\omega_i^0, \omega_l^0), i, l = 1,\ldots,n\right)$ is matrix $(n \times n)$ formed by inner products of basic elements $\omega_1^0,\ldots,\omega_n^0$, whereas in the latter case

$$||\vartheta(\mathbf{a})||^2 = \sum_{i=1}^{n} a_i^2 = \mathbf{a}^T \mathbf{a}. \tag{13}$$

In the "native" version of norm (12), the training criterion (11) is aimed at finding the shortest direction element $\vartheta \in \Omega$, and, so, all orientations of the discriminant hyperplane in the original Hilbert space are equally preferable. On the contrary, if the norm is measured as that of the vector of coefficients representing the direction element in the space of projectional features (13), the criterion (11) seeks the shortest vector $\mathbf{a} \in \mathsf{R}^n$ (13), so that equally preferable are all orientations of the hyperplane in R^n but not in Ω.

It is easy to see that if $\vartheta \in \Omega$ and $\omega \in \Omega$ are two arbitrary elements of a Hilbert space Ω, then the squared Euclidean distance from ω to its projection onto the beam formed by element ϑ equals $(\omega, \omega) - (\omega, \vartheta)^2 / (\vartheta, \vartheta)$. In its turn, it can be shown [8] that if $\mathbf{a}^T \mathbf{a} \to \min$ under the constraint $\left(\vartheta(\mathbf{a}), \vartheta(\mathbf{a})\right) = \mathbf{a}^T \mathbf{M} \mathbf{a} = \text{const}$, then $\sum_{j=1}^{n} \left(\omega_j, \vartheta(\mathbf{a})\right)^2 \to \max$, and, so, $\vartheta(\mathbf{a})$ tends to be close to the major inertia axis of the basic assembly.

Thus, training by criterion (11) with $||\vartheta(\mathbf{a})||^2 = \mathbf{a}^T\mathbf{a}$, i.e. without any preferences in the space of projectional features, is equivalent to a pronounced preference in the original Hilbert space in favor of direction elements oriented along the major inertia axis of the basic assembly of object. As a result, the discriminant hyperplane in the Hilbert space tends to be orthogonal to that axis (Fig. 3).

This is out of significance if the region of major concentration of objects in the Hilbert space is equally stretched in all directions. But such indifference is rather an exclusion than a rule. It is natural to expect the distribution of objects be differently extended in different directions, what fact will be reflected by the form of the basic assembly and, then, by the training sample. In this case, a reliable decision rule of recognition exists only if objects of two classes are spaced just in one of the directions where the extension is high. Therefore, it appears reasonable to escape discriminant hyperplanes oriented along the basic assembly even if the gap between the points of the first and the second class in the training sample has such an orientation, and prefer transversal hyperplanes (Fig. 3). It is just this preference that is expressed by the training criterion (11) with $||\vartheta(\mathbf{a})||^2 = \mathbf{a}^T\mathbf{a} \to \min$ in contrast to $||\vartheta(\mathbf{a})||^2 = \mathbf{a}^T\mathbf{M}\mathbf{a} \to \min$.

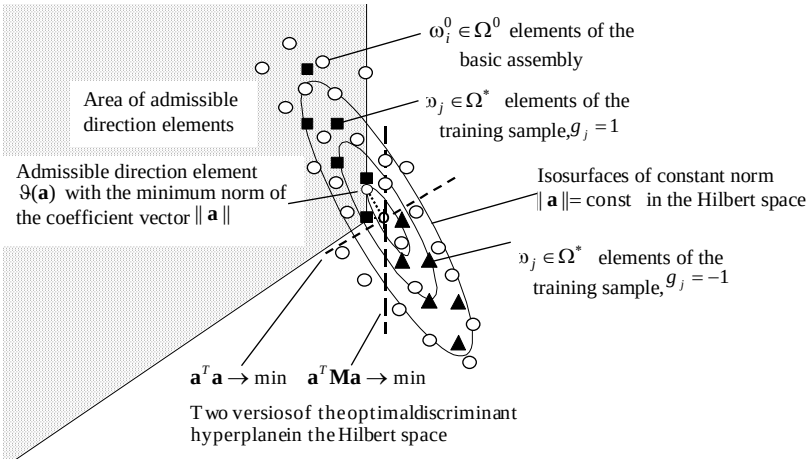

Fig. 3. Minimum norm of the direction vector of the discriminant hyperplane in the space of projectional features as criterion of training. In the original Hilbert space, the discriminant hyperplanes are preferred whose direction elements are oriented along the major inertia axis of the basic assembly.

6 Smoothness principle of regularization in the space of projectional features

Actually, training by criterion $\mathbf{a}^T\mathbf{a} \to \min$ is nothing else than a regularization method that makes use of some information on the distribution of objects in the Hilbert space. This information is taken from the basic assembly and, so, should be considered as a priori one relative to the training sample. In case the distribution is almost degenerate in some directions, it is reasonable to prefer discriminant hyperplanes of

transversal orientation even if the training sample suggests the longitudinal one as it is shown in Fig. 3.

In this Section, we consider another source of a priori information that may be drawn from the basic assembly of objects before processing the training sample. The respective regularization method follows from the very nature of projectional features, namely, from the suggestion that the closer are two objects of the basic assembly, the less should be the difference between the coefficients of their participating in the direction element of the discriminant hyperplane (5).

In the feature space of an arbitrary nature, there are no a priori preferences in favor of that or other mutual arrangement of classes, and the only source of information on the sought-for direction is the training sample. But in the space of projectional features different directions are not equally probable, and it is just this fact that underlies the regularization principle considered here.

The elements of the projectional feature vector of an object $\omega \in \Omega$ are its scalar products with objects of the basic assembly $\mathbf{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega)\right)^T \in \mathsf{R}^n$, $x_k(\omega) = (\omega, \omega_k^0)$, $\omega_k^0 \in \Omega^0 \subset \Omega$. The basic objects, in their turn, are considered as elements of the same linear Hilbert space and, so can be characterized by their mutual proximity. If two basic objects ω_j^0 and ω_k^0 are close to each other, the respective projectional features do not carry essentially different information on objects of recognition $\omega \in \Omega$, and it is reasonable to assume that the coefficients a_j and a_k in the linear decision rule should also take close values. Therein lies the a priori information on the direction vector of the discriminant hyperplane that is to be taken into account in the process of training.

In fact, the coefficients a_j are functions of basic points in the Hilbert space $a_j = a(\omega_j^0)$, and the regularization principle we have accepted consists in the a priori assumption that this function should be smooth enough. It is just this interpretation that impelled us to give such a principle of regularization the name of smoothness principle.

It remains only to decide how the pair-wise proximity of basic objects should be quantitatively measured. For instance, inner products $\mu_{jk} = (\omega_j, \omega_k)$ might be taken as such a measure. Then, the a priori information on the sought-for direction element can be easily introduced into the training criterion $\mathbf{a}^T\mathbf{a} + C\sum_{j=1}^{N}\delta_j \to \min$ in (11) with $||\vartheta(\mathbf{a})||^2 = \mathbf{a}^T\mathbf{a}$ as an additional quadratic penalty $\mathbf{a}^T(\mathbf{I} + \alpha\mathbf{B})\mathbf{a} + \; C\sum_{j=1}^{N}\delta_j \to \min$ where

$$\mathbf{a}^T\mathbf{B}\mathbf{a} = \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n}\mu_{il}(a_i - a_l)^2, \quad \mathbf{B} = \begin{pmatrix} -\mu_{11} + \sum_{i=1}^{n}\mu_{1i} & \cdots & -\mu_{1n} \\ \vdots & \ddots & \vdots \\ -\mu_{n1} & \cdots & -\mu_{nn} + \sum_{i=1}^{n}\mu_{ni} \end{pmatrix},$$

and parameter $\alpha > 0$ presets the intensity of regularization.

Because the size of the training sample N is, as a rule, less than the dimensionality n of the space of projectional features, the subsamples of the first and the second class will most likely be linearly separable. On the force of this circumstance, when solving the quadratic programming problem (11) without regularizing penalty, the

optimal shifts of objects will equal zero $\delta_j = 0$, $j = 1,...,N$. After introducing the regularization penalty, the errorless hyperplane may turn out to be unfavorable from the viewpoint of a priori preferences expressed by matrix **B** with sufficiently large coefficient α. In this case, the optimal hyperplane will sacrifice, if required, the correct classification of some especially nuisance objects of the training sample, what will result in positive values of their shifts $\delta_j > 0$.

7 Experiments on protein fold class recognition "one against one"

Experiments on fold class recognition were conducted with the collection of amino acid sequences of 396 protein domains grouped into 51 fold classes (Table 1). As the initial data set served the matrix 396×396 of pair-wise alignment scores obtained by alignment procedure Fasta 3 and considered as matrix of inner products of respective protein domains (ω_j, ω_k) in an imaginary Hilbert space.

In the series of experiments described in this Section, we solved the problem of pair-wise fold class recognition by the principle "one against one". There are $m = 51$ classes in the collection and, so, $m(m-1)/2 = 1275$ class pairs, for each of which we found a linear decision rule of recognition.

As the basic assembly $\Omega^0 = \{\omega_1^0,...,\omega_n^0\}$, we took amino acid chains of 51 protein domains, $n = 51$, one from each fold class. As representatives of classes, their "centers" were chosen, i.e. the protein domains that gave the maximum sum of pair-wise scores with other members of the respective class. Thus, each protein domain was represented by a 51-dimensional vector of its projectional features (7).

For each of the 1275 class pairs, the training sample consisted of all protein domains making the respective two classes (Table 1). Thus, the size of the training sample varied from $N = 7$ for pairs of small classes, such as (50) *Protein kinases, catalic core* and (51) *Beta-Lactamase*, to $N = 59$ in two greatest classes (9) *Immunoglobin beta - sandwich* and (26) *TIM-barrel*.

We applied the technique of pattern recognition with preferred orientation of the discriminant hyperplane along the major inertia axis of the basic assembly in the Hilbert space. The quadratic programming problem (11) was solved for each of 1275 class pairs in its dual formulation [2].

A way of empirical estimating the quality of the decision rule immediately from the training sample offers the well-known leave-one-out procedure [2]. One of the objects of the full training sample containing N objects is left out at the stage of training, and the decision rule inferred from the remaining $N-1$ objects is applied to the left-out one. If the result of recognition coincides with the actual class given by the trainer, this fact is registered as success at the stage of examination, otherwise an error is fixed. Then the control object is returned to the training sample, another one is left out, and the experiment is run again. Such a procedure is applied to all the objects of the training sample, and the percentage of errors or correct decisions is calculated, which is considered as an estimate on the quality of the decision rule inferred from the full sample would it be applied to the general population.

In each of 1275 experiments, the separability of the respective two fold classes was estimated by such a procedure. Two rates were calculated for each class pair, namely, the percentage of correctly classified protein domains of the first and the second class. As the final estimate of the separability, the worst, i.e. the least, of these two percentages was taken.

As a result, the separability was found to be not worse than:

100% in 9% of all class pairs (completely separable class pairs),
90% in 14% of all class pairs,
80% in 32% of all class pairs,
70% in 53% of all class pairs.

The separability of 26 classes from more than one half of other classes is not worse than 70%. One class, namely, (50) *Protein kinases (PK), catalytic core*, showed its complete separability from all the classes.

On a data set of a lesser size, we checked how the pair-wise separability of fold classes will change if the number of basic proteins, i.e. the dimensionality of the projectional feature space, increases essentially. For this experiment, we took all the proteins of the collection as basic ones, so that the dimensionality of the projectional feature space became $n = 396$.

The same truncated data set was used for studying how the separability of classes is affected by normalization of the alignment scores between amino acid chains, what is equivalent to projection of respective points of the imaginary Hilbert space onto the unit sphere. If (ω', ω'') is inner product of two original points of the Hilbert space associated with the respective two protein domains ω' and ω'', the inner product of their projections $\overline{\omega}'$ and $\overline{\omega}''$ onto the unit sphere will be $(\overline{\omega}', \overline{\omega}'') = (\omega', \omega'') \Big/ \left(\sqrt{(\omega', \omega')} \sqrt{(\omega'', \omega'')} \right)$. We used these values, instead of (ω', ω''), as similarity measure of protein domain pairs for fold class recognition.

For this series of experiments, we selected 7 fold classes different by their size and averaged separability from other classes. The chosen classes that contain in sum 85 protein domains are shown in Table 2.

The results are presented in Table 3. As we see, the extension of the basic assembly improved the separability of the class pairs that participated in the experiment. As to the normalization of the alignment score, it led to an improvement with the small basic assembly and practically did not change the separability with the enlarged one.

Experimental study of effects of regularization was conducted with the same truncated data set (Table 2). We examined how the smoothness principle of regularization, expressed by the modified quadratic programming problem (4.1, improves the separability of fold classes “one against one” within the selected part of the collection. The separability of each of 21 pairs of classes was estimated by the leave-one-out procedure several times with different values of the regularization coefficient α. Each time, the separability of a class pair was measured by the worst percentage of correct decisions in the first and the second class, whereupon the averaged separability over all 21 class pairs was calculated for the current value of α.

Such a series of experiments was carried out twice, with original and normalized alignment scores. The dependence of the separability on the regularization coefficient in both series is shown in Fig. 4. In both series of experiments, a marked improvement of the separability is gained. The quality of training grows as the regularization coefficient increases, however, the improvement in not monotonic. A slight drop in separability with further increase in the coefficient after the maximum is attained arises from a too deep roughness of the decision rule adjustment.

Table 2. Seven fold classes selected for the additional series of experiments.

	Fold class	Size	Averaged separability from other classes

1	Globin	12	73.4 %
3	Four-helical bundle	8	70.8 %
4	Ferritin	8	60.4 %
5	4-gelical cytokines	11	66.0 %
10	Common fold of difteria toxin / transcription factors / cytochrome	5	65.2 %
12	C2 domain	3	8.2 %
26	TIM-barrel	28	52.6 %

Table 3. Averaged pair-wise separability of seven fold classes in four additional experiments.

Size of the basic assembly n	**Averaged separability**	
	Original score matrix (ω', ω'')	**Normalized score matrix** $(\overline{\omega}', \overline{\omega}'')$
51	63.5 %	69.4 %
396	76.6 %	75.3 %

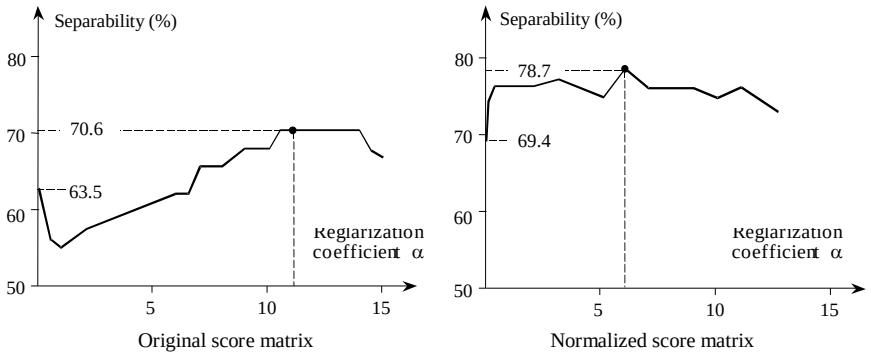

Fig. 4. Dependence of the averaged pair-wise separability over 21 fold class pairs on the regularization coefficient.

8 Conclusions

Within the bounds of the featureless approach to pattern recognition, the main idea of this work is treating the pair-wise similarity measure of objects of recognition as inner product in an imaginary Hilbert space, into which really existing objects may be mentally embedded as a subset of isolated points. Two ways of regularization of the training process follow from this idea, which contribute to overcoming the small size of the training sample. In the practical problem of protein fold class recognition, to embed the discrete set of known proteins into a continuous Hilbert space, we propose to consider as inner product the pair-wise alignment score of amino acid chains, which is

commonly adopted in bioinformatics as their biochemically justified similarity measure.

9 References

1. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning, Vol. 20, No.3, 1995.
2. Vapnik, V. Statistical Learning Theory. John-Wiley & Sons, Inc. 1998.
3. Duin, R.P.W, De Ridder, D., Tax, D.M.J. Featureless classification. Proceedings of the Workshop on Statistical Pattern Recognition, Prague, June 1997.
4. Duin, R.P.W, De Ridder, D., Tax, D.M.J. Experiments with a featureless approach to pattern recognition. Pattern Recognition Letters, vol. 18, no. 11-13, 1997, pp. 1159-1166.
5. Duin, R.P.W, Pekalska, E., De Ridder, D. Relational discriminant analysis. Pattern Recognition Letters, Vol. 20, 1999, No. 11-13, pp. 1175-1181.
6. Fetrow J.S., Bryant S.H. New programs for protein tertiary structure prediction. Biotechnology, Vol. 11, April 1993, pp. 479-484.
7. Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., Kim, S.-H. Recognition of a protein fold in the context of the SCOP classification. Proteins: Structure, Function, and Genetics, 1999, 35, 401-407.
8. Mottl, V., Dvoenko, S., Seredin, O., Kulikowski, C., Muchnik, I. Alignment Scores in a Regularized Support Vector Classification Method for Fold Recognition of Remote Protein Families. DIMACS Technical Report 2001-01, January 2001. Center for Discrete Mathematics and Theoretical Computer Science. Rutgers University, the State University of New Jersey, 33 p.
9. Durbin, R., Eddy, S., Krogh, A., Mitchison, G. Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1988.
10. Pearson, W. R., Lipman, D. J. Improved tools for biological sequence analysis. PNAS, 1988, 85, 2444- 2448.
11. Pearson, W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in Enzymology, 1990, 183, 63-98.