Genomics 111 (2019) 1777–1784

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Convex hull principle for classification and phylogeny of eukaryotic proteins

Xin Zhao[a,1], Kun Tian[a,1], Rong L. He[b], Stephen S.-T. Yau[a,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China
[b] Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA.

ARTICLE INFO

Keywords:
Convex hull principle
Classification
Protein kinases
Human proteins
Natural vector
Phylogenetic analysis

ABSTRACT

This study quantitatively validates the principle that the biological properties associated with a given genotype are determined by the distribution of amino acids. In order to visualize this central law of molecular biology, each protein was represented by a point in 250-dimensional space based on its amino acid distribution. Proteins from the same family are found to cluster together, leading to the principle that the convex hull surrounding protein points from the same family do not intersect with the convex hulls of other protein families. This principle was verified computationally for all available and reliable protein kinases and human proteins. In addition, we generated 2,328,761 figures to show that the convex hulls of different families were disjoint from each other. The classification performs well with high and robust accuracy (95.75% and 97.5%) together with reasonable phylogenetic trees validate our methods further.

1. Introduction

The family of eukaryotic protein kinases comprises one of the largest super families of homologous proteins and genes [1,2]. The proteins in this group play key roles in biology and disease [3,4]. A protein kinase (PK) is an enzyme that phosphorylates proteins by chemically adding phosphate groups to specific amino acid residues. The identification and classification of eukaryotic protein kinases are fundamental to a proper understanding of phosphorylation events and will lead to a better description of the biochemical circuitry of cells. This may guide the development of more effective drugs [5,6]. The study of Hanks and Hunter in 1995 classified eukaryotic protein kinases into a four-level hierarchical structure [7], including group, family, subfamily and individual PKs based on the conserved sequence and structural profile of the kinase domain. Eukaryotic protein kinases can also be split into two broad groups: conventional protein kinases (ePKs) and atypical protein kinases (aPKs) [8]. The phylogenetic tree of ePKs contains eight major clusters. A ninth group called the “Other” group consists of a mixed collection of kinases that do not fit into the previous groups [9,10]. The aPKs are a small set of protein kinases that do not share clear sequence similarity with ePKs.

A central problem in protein classification is how proteins are clustered in relation to each other [11,12]. In previous study, the techniques used to cluster or classify protein sequences usually required long computation time to obtain the results, such as multiple sequence alignment, CD-HIT method [13], UCLUST method [14] and so on. On the other hand, the information of protein sequences was not totally reflected by many existing methods, for example, the moment vector method [15]. In order to get a global view of evolutionary distances among multiple proteins, the concept of protein space was introduced [15–18]. A protein space representation using natural vectors was proposed by Yau [19–22]. Each protein sequence is represented by a 60-dimensional natural vector. The biological distance between any two proteins can be measured by the Euclidean distance between the corresponding points in 60-dimensional space. This simplified representation maintains most of the inherent biological information of protein sequences in the study of phylogenetic clustering [19]. Using this method, similar proteins cluster together and arbitrary amino acids sequences are distinguished from proteins [16,23,24]. However, the 60-dimensional natural vector representation has limitations. For example, the 60-dimensional convex hulls of the MAPK family and the STE20 family in our animal protein kinase intersect (Fig. 1A), suggesting that different protein families cannot be distinctly separated. There is still room for improvement in the definition of the natural vector representation.

To improve the 60-dimensional natural vector representation, we incorporated covariance into our model, a concept widely used in statistics. We define the correlation between each pair of the amino acids as their covariance. This quantity indicates the relation between the distribution of two amino acids in a protein sequence. The detailed

* Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ygeno.2018.11.033
Received 2 September 2018; Received in revised form 25 November 2018; Accepted 30 November 2018
Available online 05 December 2018
0888-7543/

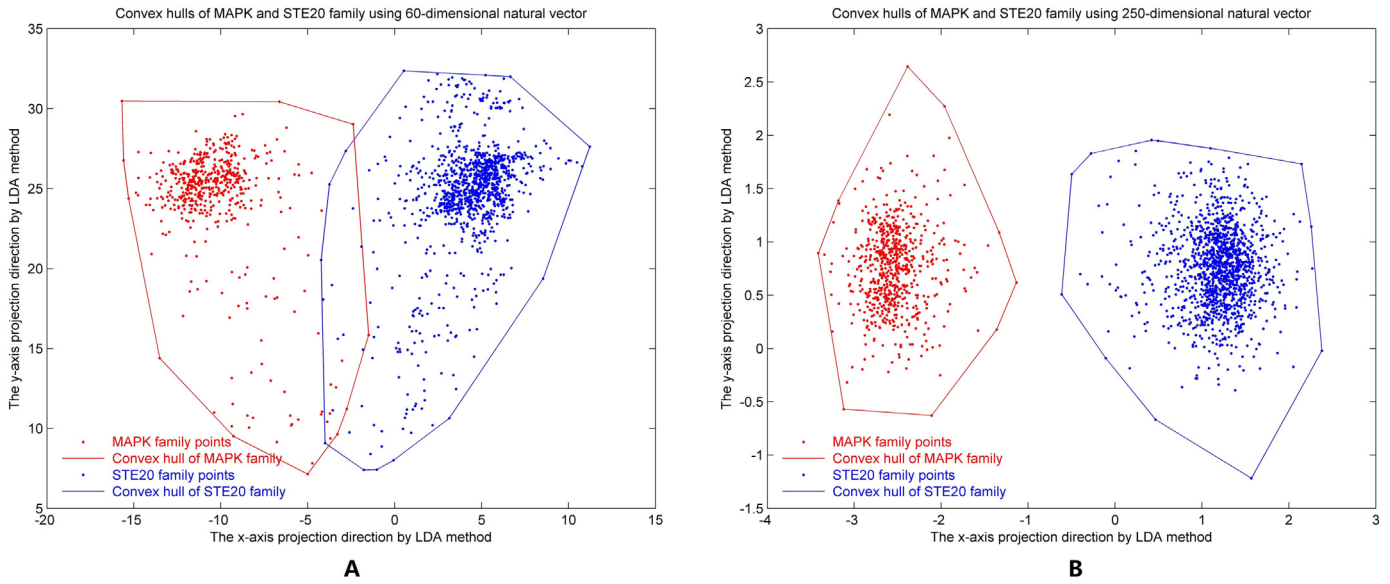

Fig. 1. Convex hulls of MAPK and STE20 family after dimension reduction by LDA method. The red points represent the MAPK family (666 points) and the blue points represent the STE20 family (1396 points) in the animal protein kinase dataset. Fig. 1A shows that the convex hulls of these two families intersect if we use the 60-dimensional natural vector. In Fig. 1B, the two convex hulls have no intersection using the 250-dimensional natural vector. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

algorithm to compute the covariance appears in the Materials and methods section. For each pair of amino acids in a protein sequence, we get one covariance. When we consider all distinct pairs of amino acids, there are a total of 190 covariance values. Appending these 190 values to the original 60-dimensional natural vector, we get a new 250-dimensional natural vector that contains more information about the amino acid distribution than that generated from the original natural vector. The disjointness of the MAPK and STE20 animal protein kinase families is shown in Fig. 1B. Latter result indicates that 250-dimensional natural vectors represent protein sequences more optimally than 60-dimensional natural vectors. For the above reasons, we employed 250-dimensional natural vectors to represent proteins of interest. This strategy is alignment-free and is capable of representing more biological information than alignment models.

This current study is underpinned by Anfinsen's central dogma [25] of molecular biology: all of the biological properties that are elicited as result of the genotype of a protein are determined by the distribution of the 20 amino acids within the protein. We propose a new 250-dimensional natural vector to describe the distribution of the 20 amino acids within a protein. In order to observe this principle of molecular biology, we performed the convex analysis which states that the convex hull formed from the natural vectors of proteins from the same family do not intersect with convex hulls of natural vectors from other families. This principle indicates that proteins with similar distributions of the 20 amino acids should be in the same family. We verified this principle and presented the results for large reliable datasets on protein kinase domains and human protein sequences. We subsequently demonstrated the effectiveness of method on classification and phylogenetic analysis. Cross-validation and bootstrapping method are also used in this study. The high and robust accuracy indicating the efficiency of both new natural vector and convex analysis.

2. Results

2.1. Protein kinase dataset

The protein kinase dataset used in this study consisted of 31,355 protein kinase domains. Please see the Materials and methods section for further details. Protein kinases mentioned in this study are referring to protein kinase domains. The animal protein kinase dataset contains nine groups divided into 87 families with a total of 19,095 sequences. The plant protein kinase dataset used here contains 12,260 sequences that were divided into seven groups and 20 families. Fig. 2 displays a heatmap of the classifications and identification patterns for several of the major animal protein kinase groups. This heatmap was generated using the gplots program in the R package (http://www.r-project.org/). The results show that the numbers of animal protein kinases in the same group or family can differ greatly across species.

2.2. Convex hull analysis of protein kinase dataset

For each protein kinase, we first calculated the 250-dimensional natural vector and then constructed the convex hull for each protein kinase family in 250-dimensional space. From the results of a linear programming analysis, no intersection was observed between any pair of the convex hulls for the animal protein kinase families. The same conclusion was drawn for plant protein kinase families. Our results indicate that proteins with a similar distribution profile for the 20 amino acids should be in the same family. The results are also consistent with the central law of molecular biology. In order to visualize the results, we applied the linear discriminant analysis (LDA) method to facilitate dimension reduction. LDA is a method used to determine whether two groups are linearly separable. The detailed descriptions can be found in the Materials and methods section. The dimension of the natural vectors was reduced from 250 to 2. We have put the completed results on our website http://yaulab.math.tsinghua.edu.cn/Lda/. Projections of the convex hulls for several animal protein kinase families are shown in Fig. S1. We can clearly see that the points in protein space are clustered, rather than being broadly distributed. This suggests that as new protein kinase sequences are included, their points will lie approximately within the convex hull of the points corresponding to known protein kinase families.

2.3. Classification of animal protein kinases

We examined the classification performance of the 250-dimensional natural vectors of the animal protein kinases. The 1-nearest neighbor algorithm was used to classify the sequences into 87 families from 62

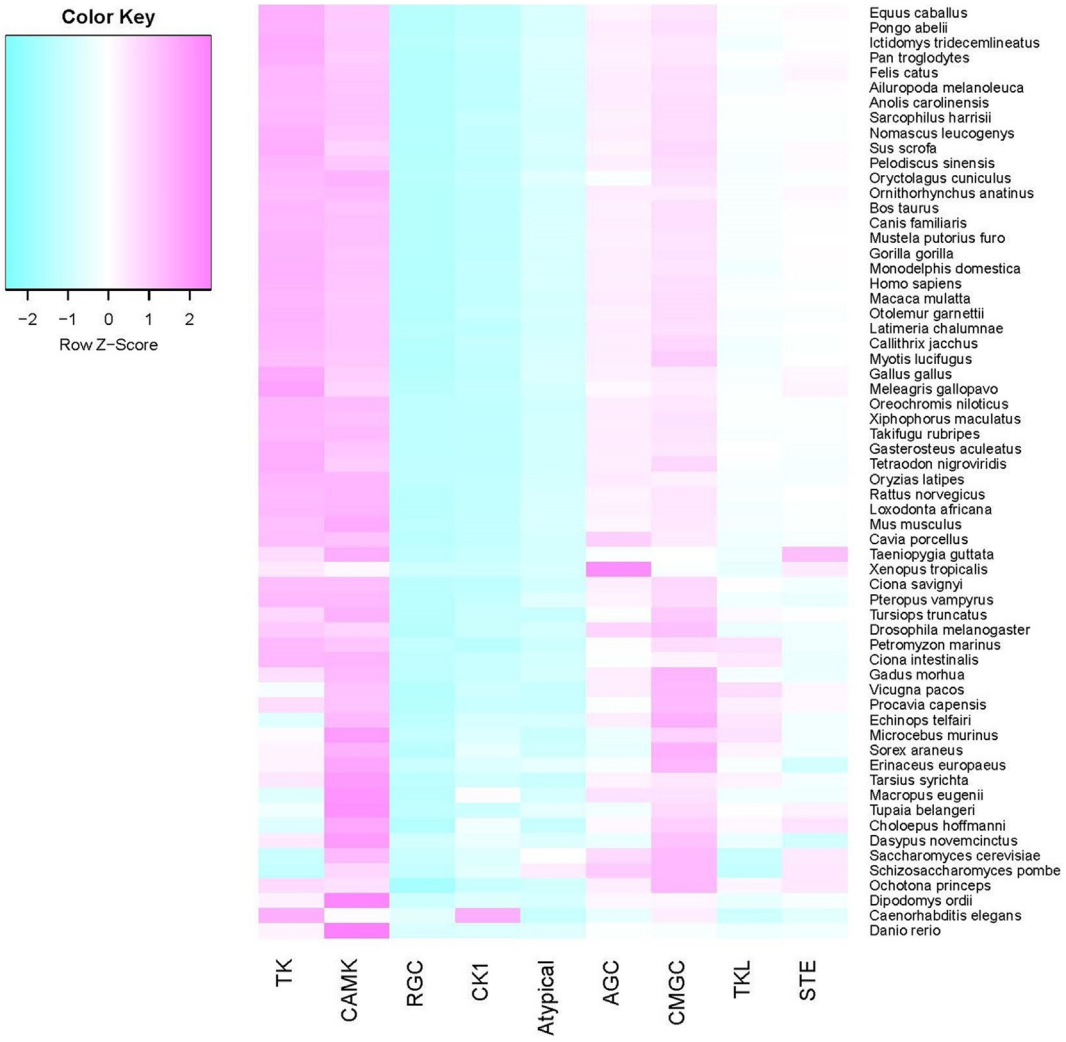

Fig. 2. The heatmap for several major groups of animal protein kinases. Nine major groups of the protein kinases are shown. The pink color grid represents a large number of proteins for the corresponding species on the right hand in major group below. The blue color gives the opposite meaning. For example, CAMK kinases and RGC kinases have been both widely detected in animal protein kinases, but the number of CAMK kinases is larger than RGC kinases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

eukaryotic species. The 62 eukaryotic species appeared in both the training dataset and the test dataset. The accuracy of our classification was tested by predicting the protein kinases in the test dataset and comparing against their existing classification. Fig. 3 shows our accuracy for the nine protein kinase groups. The total accuracy was 18,284/19,095 (95.75%) for all of the animal protein kinase dataset.

Good performance was obtained for the classification of the AGC group (2508/2618). The result of the AGC group classification is shown in Fig. S2. However, the 1-nearest neighbor algorithm performed less well on the PDK1 subfamily, where only 48/58 (82.8%) sequences were classified correctly. In contrast, protein kinases in other families were classified with an accuracy rate of > 90%. This reflected the high degree of sequence conservation in the AGC group across a large evolutionary distance.

The 1-nearest neighbor algorithm also performed well in relation to the classification of another large ePKs group, CAMK (3757/3958). The result of this classification is shown in Table S1. The CAMKL family, which has the most protein kinases, had a relatively high accuracy rate (95.6%). Other families apart from the RAD53 family (accuracy rate of 35/43, 81.4%) also exhibited high classification accuracies. This indicates that the greater the number of sequences in the family, the more features the model can generate which in turn increases the accuracy. The classification accuracy for other ePKs groups was assessed in a similar manner. Tables S2–S6 reveal the results of this classification. Most families from these groups were accurately classified. The families with the largest number of kinases in these groups have a very high accuracy rate (96% to 99%). The 1-nearest neighbor algorithm correctly classified all of the ePKs domains in the TTBK family of the CK1 group.

Classification of atypical protein kinases was extremely accurate with 1034/1062 (97.4%) of these proteins correctly classified (shown in Table S7). The sequences in the PDHK family were all correctly classified and the accuracy rates of the other families were approximately 97%. This high accuracy reflects a high degree of conservation among aPKs in many species over a large evolutionary distance.

2.4. Classification of plant protein kinases

We also performed a classification analysis on plant protein kinase dataset in a similar manner. The dataset used in this study contains seven groups and 20 families with 12,260 sequences in total. Table 1 shows the accuracy rates for the seven protein kinase groups. The total accuracy is 11,959/12,260 (97.54%) for all of the plant protein kinase domains. The 1-nearest neighbor algorithm also performed well on the classification of plant protein kinases. The protein kinases from the PDHK family were all correctly classified and the accuracy rates of the

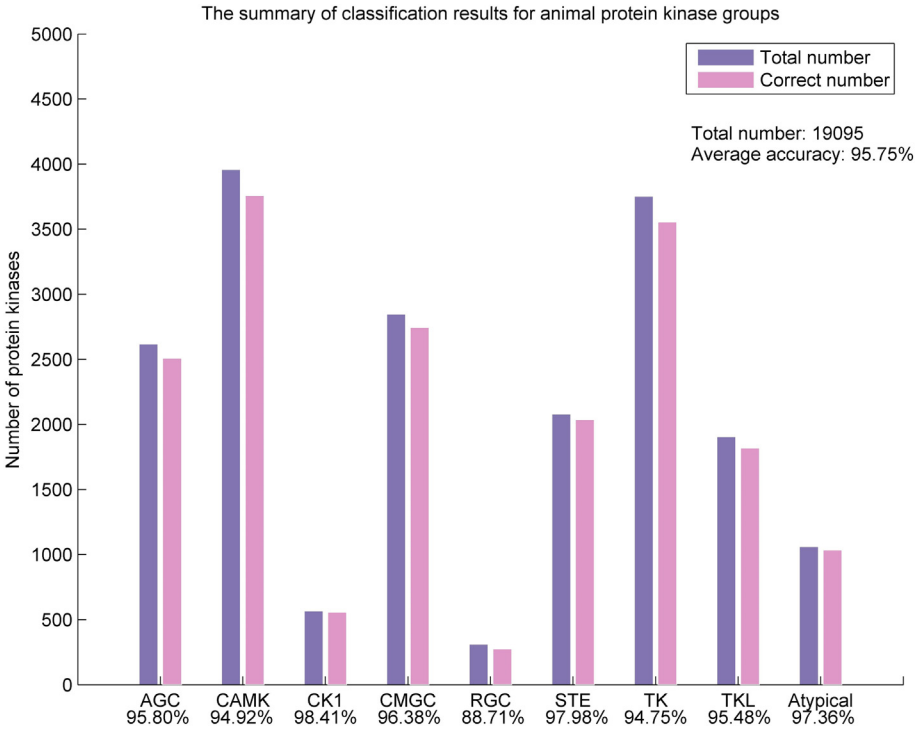

Fig. 3. The summary of classification results for animal protein kinase groups.

Table 1
The summary of classification results for plant protein kinase groups.

Group name	Family name	Total number	Correct number	Accuracy
AGC	MAST	62	59	0.951613
	NDR	137	129	0.941606
	PDK1	31	29	0.935484
	PKA	33	28	0.848485
Atypical	ABC1	395	365	0.924051
	PDHK	45	45	1
	PIKK	118	109	0.923729
	RIO	56	50	0.892857
CAMK	CAMK1	62	48	0.774194
CK1	CK1	306	291	0.95098
CMGC	CDK	686	644	0.938776
	CLK	83	80	0.963855
	DYRK	162	150	0.925926
	GSK	205	201	0.980488
	MAPK	372	358	0.962366
	RCK	76	68	0.894737
STE	STE11	506	483	0.954545
	STE20	193	179	0.927461
TKL	IRAK	8136	8083	0.993486
	MLK	596	560	0.939597
Total	/	12,260	11,959	0.975449

Table 2
The summary of the ten-fold cross validation results on plant protein kinase dataset.

Group name	Family name	Accuracy
AGC	MAST	0.916667
	NDR	0.875
	PDK1	1
	PKA	1
Atypical	ABC1	0.952941
	PDHK	1
	PIKK	1
	RIO	1
CAMK	CAMK1	0.75
CK1	CK1	0.98
CMGC	CDK	0.922078
	CLK	0.888889
	DYRK	0.909091
	GSK	0.953488
	MAPK	0.970588
	RCK	0.909091
STE	STE11	0.938053
	STE20	0.926829
TKL	IRAK	0.991437
	MLK	0.94
Total	/	0.973909

other families were almost all above 90%. This high accuracy rate also reflects that natural vectors provide a good representation of protein space.

2.5. Statistical analysis

The 1-nearest neighbor algorithm performed well in the classification of protein kinase families. The low complexity of this algorithm makes it timesaving and easy to manipulate. To ensure that the model had high prediction accuracy with low bias and low variance, we applied k-fold cross-validation to identify the model [26]. We chose ten-fold validation to generate the results that are listed in Fig. S3 and Table 2. The complete results on animal protein kinase can be found in Supplementary material Tables S8–S15. Performing ten-fold cross-validation yielded high accuracy on animal protein kinase dataset (94.40%) and plant protein kinase dataset (97.30%). This indicates that the 1-nearest neighbor algorithm performed robustly on the classification of protein kinases and the model is valid.

2.6. Phylogenetic analysis on human protein kinases

In addition, we performed phylogenetic analysis on human protein kinases to demonstrate the validation of our method. For each group of ePKs, we chose the center of convex hull to represent this group. After calculating the 250-dimensional natural vector, we then computed Manhattan Distance between groups. Manhattan Distance is a widely-used metric for measuring the difference of high dimensional points. It

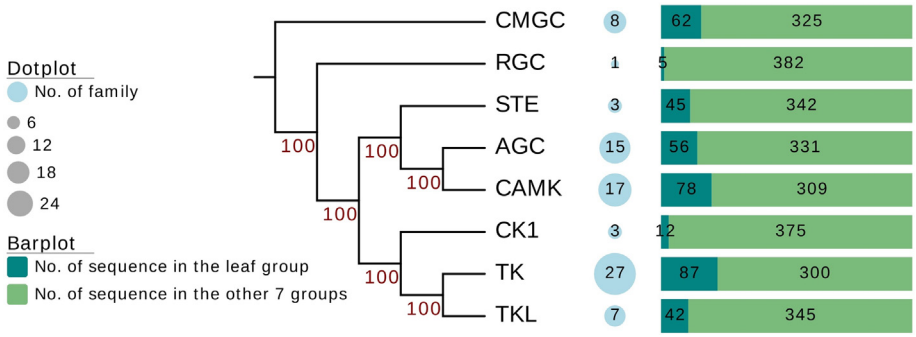

Fig. 4. Phylogenetic tree for eight ePKs groups of human protein kinases. The tree was constructed using UPGMA algorithm based on center of convex hull. The number of families and sequences for each group was presented beside the tree. The bootstrap confidence values were generated using 500 permutations.

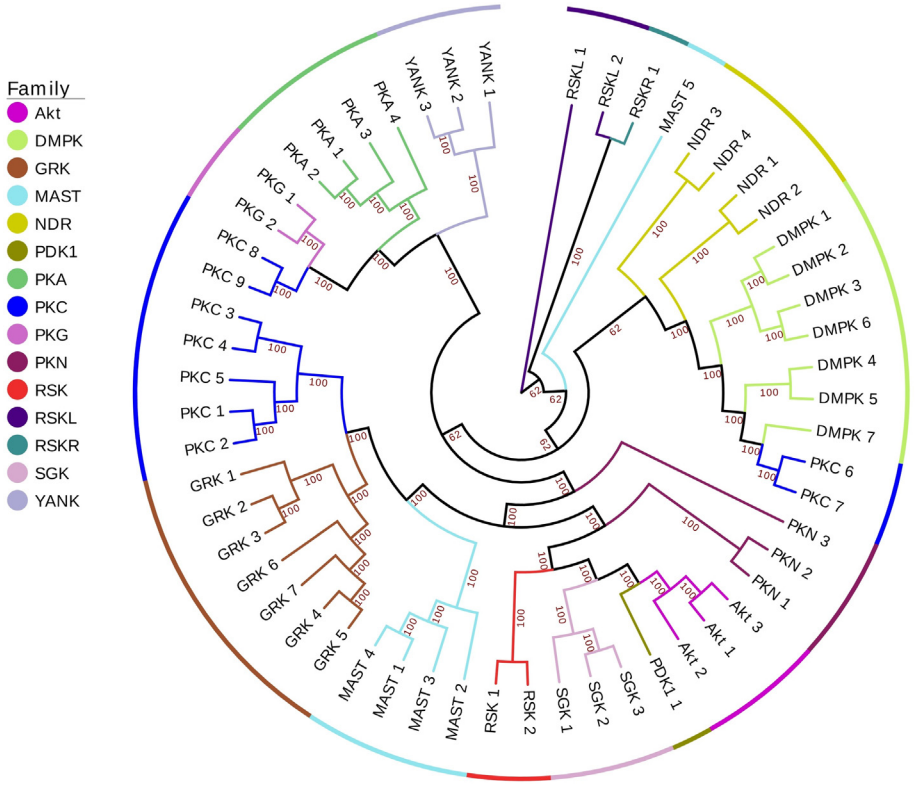

Fig. 5. Phylogenetic tree for fifteen families in AGC group of human protein kinases. The tree was constructed using UPGMA algorithm with the 250-dimensional natural vector method. Different colors were allocated to represent different families. The bootstrap confidence values were generated using 500 permutations.

has been successfully used for computing the difference between natural vectors and constructing phylogenetic trees [18]. We carried out UPGMA algorithm to reconstruct the phylogeny shown in Fig. 4. Furthermore, the phylogeny of families within each group is also established. Fig. 5 shows the phylogenetic relationship for fifteen families of AGC group. From the phylogenetic tree, we could see that sequences from the same families clustered well. The close phylogenetic relationship such as AKT and SGK, RSKR and RSKL clustered together in this tree. To better understand and visualize the disjointness of convex hulls between different families, we also plotted the two-dimensional projection using LDA method shown in Fig. S4. For atypical kinase group, we have also reconstructed the evolutionary relationship of five families as shown in Fig. S5. Fig. S6-S12 shown in Supplementary material present the phylogenies of other seven groups. We performed bootstrapping method to compute the confidence probabilities on phylogenetic trees. Bootstrapping is a common test to estimate the significance of the branches in phylogenetic trees and detailed steps can be found in Materials and methods section. The bootstrap values also confirm that our methods applied on this dataset are reasonable and convincing. Moreover, we performed phylogenetic analysis using the 60-dimensional natural vector and moment vector methods on the same AGC group. The related phylogenetic trees are shown in Fig. S13 and S14 in Supplementary material. Comparing these two trees with Fig. 5, we can see that the current 250-dimensional natural vector method performs better than the other methods. The 250-dimensional natural vector contains more useful statistics information of sequences including the correlations between different nucleotides, which could make results more precisely in phylogenetic analysis.

2.7. Convex hull analysis of human protein family

Moreover, we applied convex hull method on the human protein family dataset to illustrate its effectiveness for classification analysis on anther data platform. This dataset used in our study was downloaded from Uniprot. The dataset contained a total of 19,593 protein sequences, 5912 of which were not annotated with family information.

We removed the sequences without annotation, and applied our new natural vector method to the 13,681 sequences which were annotated with family information. There are 4854 families represented in this dataset. The distribution of sequences into different families is uneven since the biggest family has 670 sequences while the smallest family has only one sequence. Among these 4854 families, 2698 have only one protein member. Except for these 2698 sequences, the other 10,983 sequences belong to 2156 different families. Further detailed information can be found in Supplementary material.

For each protein, we first computed the associated 250-dimensional natural vector based on amino acid sequence and we then constructed the convex hull for each protein family in 250-dimensional space. Using linear programming analysis and the LDA method, no intersection was found between any pair of convex hulls when we used the 250-dimensional natural vector representation. This result suggests that based on the 250-dimensional natural vector representations, the convex hulls formed by different families were disjoint from each other. We believe that the 250-dimensional natural vector representation can be applied to the study of protein sequence clustering. For example, in this study we illustrate that the 250-dimensional convex hulls of the two biggest families do not intersect using the LDA method (Fig. S15). We have produced similar illustrations for each pair of human protein families, so the interested reader can verify that their convex hulls are disjoint. The complete graphical visualization results can be found on our website http://yaulab.math.tsinghua.edu.cn/Lda/.

3. Discussion

3.1. Natural vectors provide a good representation of protein space

In this study, we used 250-dimenstional natural vectors to represent protein sequences as points in 250-dimensional Euclidean space by using distribution information for each amino acid within the protein. Each protein sequence is in one-to-one correspondence with a point in protein space, where proteins with similar properties stay close together. Therefore, the distance between two points in protein space represents the biological distance of the corresponding two proteins. Following a linear programming analysis, no intersection was observed between any pair of the convex hulls of protein families using 250-dimensional natural vectors. This result suggests that when 250-dimensional natural vectors are used, the convex hulls formed by different families are disjoint from each other. This disjointness property indicates that protein sequences from the same family are likely clustered, rather than being broadly distributed. In addition, good performances of classification and phylogenetic results using 250-dimensional natural vectors also suggest that natural vectors appropriately represent protein space.

3.2. Prediction of unannotated proteins and discovery of new proteins

Using 250-dimensional natural vectors, we can predict families of unannotated proteins. The convex hulls of protein families where the associated sequences are represented by 250-dimensional natural vectors can be used as clear boundaries to cluster proteins into families. This principle can be used to predict which families new proteins belong to. If an unknown function protein sequence S is not clustered into any known protein family, we can calculate the distances between this sequence and convex hulls of each known protein family. By finding the minimum value of these distances, we can obtain the protein family F which has the smallest distance to S. It's reasonable to infer that the function of sequence S is close to the protein family F and therefore could be predicted by the properties of the proteins in family F. The prediction strategy is even more accurate when more annotated protein families are available. The most significant potential application of our convex analysis involves searching for new proteins that lie within the convex hull of a protein family. Because the convex hull is composed of natural vectors of proteins from the same family, the natural vectors of undiscovered proteins from a family should also be within the convex hull. We can test natural vectors within the convex hull. Because natural vectors and sequences are in one-to-one correspondence, there should only be one amino acid sequence corresponding to that natural vector. This amino acid sequence can subsequently be synthesized in the laboratory. Accordingly, from the sequence we can predict protein function from the properties of the proteins in the corresponding family.

4. Conclusions

We have shown here that a novel 250-dimensional natural vector has effectiveness for describing the distribution of the 20 amino acids within a protein. Proteins from the same family are found to cluster together, and the convex hull surrounding protein points of natural vectors from the same family do not intersect with convex hulls of other protein families. This convex analysis implies that proteins with similar distributions of the 20 amino acids should be in the same family. We verify our principle computationally by using all available and reliable sequences on protein kinase datasets and human proteins. In addition, we also provide graphical figures for each of these 2,328,761 pairs of convex hulls so that interested readers can visually verify that the convex hulls of different human protein families or different protein kinase families are disjoint from each other. The complete visualization results can be found on our website http://yaulab.math.tsinghua.edu.cn/Lda/. Moreover, the classification and phylogenetic analysis are also carried out to validate and demonstrate the method in this study.

There are many applications of convex analysis. It allows us to do phylogenetic analysis of protein families. It also provides a quick way to assign a newly discovered protein to a protein family. Perhaps the most significant potential application of our convex hull principle is to search for new proteins that lie within the convex hull of a protein family. Once a new protein is found, it could be classified based on what convex hull its natural vector belongs to. Its functions could be predicted from the properties of the proteins in the corresponding family. It is hoped that this strategy will open up a new combinatorial approach that encompasses interdisciplinary research in biology, mathematics and computer science. We believe that our convex analysis will become a powerful tool in the study of proteins.

5. Materials and methods

5.1. Datasets

The protein kinase dataset used in this study is from the Eukaryotic Kinase and Phosphatase Database (http://ekpd.biocuckoo.org/) [27], which is managed and sponsored by the researcher Yu Xue. This dataset contains information pertaining to animal protein kinases and plant protein kinases. The current classification scheme for eukaryotic protein kinases is used in this dataset, where eukaryotic protein kinases are split into two broad groups: conventional protein kinases (ePKs) and atypical protein kinases (aPKs). The ePKs contains eight groups: the AGC group, the CAMK group, the CK1 group, the CMGC group, the RGC group, the STE group, the TK group and the TKL group. The full names of these eight groups are as follows: cyclic nucleotide- and calcium-phospholipid-dependent kinases (the AGC group including the PKA, PKG, and PKC families), calmodulin-dependent kinases (the CAMK group), casein kinase 1 (the CK1 group), cyclin-dependent kinases, mitogen-activated protein kinases, CDK-like kinases, and glycogen synthase kinase (the CMGC group), receptor guanylate cyclase kinases (the RGC group), many kinases functioning in MAP kinase cascades (the STE group), tyrosine kinases (the TK group) and tyrosine kinase-like kinases (the TKL group). The aPKs are a small set of protein kinases that do not share clear sequence similarity with ePKs. Only Alpha, PIKK, PHDK, and RIO families have been shown to exhibit kinase activity.

Table 3
The group names and number of protein kinase domains of each group as well as family numbers for each group in this study.

Group	No. sequences for animal	No. families for animal	No. sequences for plant	No. families for plant
AGC	2618	15	263	4
CAMK	3958	17	62	1
CK1	566	3	306	1
CMGC	2847	8	1584	6
RGC	310	1	0	/
STE	2078	3	699	2
TK	3752	28	0	/
TKL	1904	7	8732	2
Atypical	1062	5	614	4
Total	19,095	87	12,260	20

Table 3 displays the nine groups of animal protein kinase domains and seven groups of plant protein kinase domains used in this study with their respective protein kinase sequence numbers. The minimum, maximum and average lengths of sequences in each group for animal and plant protein kinases are shown in Table S16-S17 in Supplementary material. A total of 19,095 animal protein kinases sourced from 62 different eukaryotic species and 12,260 plant protein kinases from 22 eukaryotic species were analyzed in this study. We verified the convex hull principle by analyzing these 31,355 protein kinase domains as well as 20,000 human protein sequences.

5.2. Number, mean position and normalized variation features of natural vector

Assume $S = (s_1, s_2, s_3, \cdots, s_N)$ is a protein sequence of length N, i.e. $s_i \in \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$ for each positive integer i from 1 to N. Firstly, we define the following function of twenty amino acids and sequence positions. For a given amino acid k and position $1 \le i \le N$, we define $f(k,i) = 1$ if k appears at the ith position of the sequence, otherwise, $f(k,i) = 0$. Let $n_k = \Sigma_{i=1}^{N} f(k,i)$ be the number of amino acid k in the sequence, $\mu_k = \Sigma_{i=1}^{N} i \cdot f(k,i)/n_k$ be the mean position of amino acid k in the sequence and $D_2^k = \sum_{i=1}^{N} (i - \mu_k)^2 f(k,i)/(n_k N)$ be the normalized variation of the position for amino acid k.

For example, given a protein sequence $'ACDEAC'$, we could compute $n_A = n_C = 2$, $n_D = n_E = 1$. For the mean positions, $\mu_A = (1 + 5)/2 = 3$, $\mu_C = (2 + 6)/2 = 4$, $\mu_D = 3$, $\mu_E = 4$. For the normalized variations, $D_2^A = [(1-3)^2 + (5-3)^2]/(2 \cdot 6) = 2/3$, $D_2^C = [(2-4)^2 + (6-4)^2]/(2 \cdot 6) = 2/3$, $D_2^D = (3-3)^2/(1 \cdot 6) = 0$, $D_2^E = (4-4)^2/(1 \cdot 6) = 0$. Here $n_k = \mu_k = D_2^k = 0$ for the other 16 amino acids $k \in \{R, N, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$.

5.3. Covariance between different amino acids

For two finite point sets $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_m\}$ in R,where $a_1 < a_2 < \ldots < a_n$ and $b_1 < b_2 < \ldots < b_m$, the covariance between the two sets $Cov(A,B)$ can be calculated in two cases. If $m = n$, we define

$$Cov(A,B) = \sum_{i=1}^{n} (a_i - \mu_A)(b_i - \mu_B)/n \quad (1)$$

here $\mu_A = \Sigma_{i=1}^{n} a_i/n$, $\mu_B = \Sigma_{i=1}^{m} b_i/m$. If $m \neq n$, assume that $m > n$. Then the covariance between A and any n values in B could be computed. We take the average of these C_m^n results as the final covariance $Cov(A,B)$ between the two point sets.

For a sequence S of length N, we could compute the covariance between any pair of amino acids k_1 and k_2. Assume that position of k_1 appeared in the sequence S is $A = \{a_1, a_2, \cdots, a_n\}$, the position of k_2 is $B = \{b_1, b_2, \cdots, b_m\}$. Then the covariance formula between k_1 and k_2 is defined as

$$Cov(k_1, k_2) = Cov(A,B)/N \quad (2)$$

As the example of sequence $'ACDEAC'$, the covariance of nucleotide A and C can be computed as follows. Based on $\mu_A = 3$, $\mu_C = 4$ and the positions of amino acids $A = \{1,5\}$, $C = \{2,6\}$ in the sequence, we could get $Cov(A,C) = [(1-3)(2-4)/2 + (5-3)(6-4)/2]/6 = 2/3$. The other covariances could also be calculated in the same way.

5.4. 250-dimensional natural vector

After getting the covariances of the pairs of amino acids, we add the covariances to the original natural vector of the sequence S. The number of pairs of amino acids is $C_{20}^2 = 190$. Therefore, the dimension of the following natural vector with covariance is 60 + 190 = 250, and the 250-dimensional natural vector is

$$(n_A, n_R, \cdots, n_V, \mu_A, \mu_R, \cdots, \mu_V, D_2^A, \cdots, D_2^V, Cov(A,R)/N, Cov(A,N)/N, \cdots Cov(Y,V)/N) \quad (3)$$

In this study, we used the 250-dimension natural vector with covariance to represent protein sequence which contains more statistic information.

5.5. Convex hull of a given point set

Given a point set $A = \{a_1, a_2, \cdots, a_n\}$ in R^k space, the convex hull of A is defined as

$$Conh(A) = \left\{p \mid p = \sum_{i=1}^{n} \lambda_i a_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \ge 0, 1 \le i \le n\right\} \quad (4)$$

or

$$Conh(A) = \{\cap C \mid C \supset A, C \text{ is convex set}\} \quad (5)$$

Here a convex set C satisfies that

$$\lambda_1 c_1 + \lambda_2 c_2 \in C \text{ for any } c_1 \in C, c_2 \in C, \lambda_1 + \lambda_2 = 1, \lambda_1 \ge 0, \lambda_2 \ge 0 \quad (6)$$

According to this definition, a convex hull is the smallest convex set containing the given points.

5.6. Methods for checking the disjointness of two convex hulls

Given two finite point sets $A = \{a_1, a_2, \cdots, a_n\}$ and $B = \{b_1, b_2, \cdots, b_m\}$ in R^k, we want to check whether the convex hulls of A and B have intersection. Based on the definition of convex hull, if there are two groups of coefficients λ_i and μ_j such that.

$\Sigma_{i=1}^{n} \lambda_i a_i = \Sigma_{j=1}^{m} \mu_j b_j$, $\Sigma_{i=1}^{n} \lambda_i = 1$, $\Sigma_{j=1}^{m} \mu_j = 1$, $0 \le \lambda_i$, $\mu_j \le 1$, $1 \le i \le n$, $1 \le j \le m$, (7).

then the two convex hulls of A and B have interactions, otherwise, the two convex hulls are disjoint [28].

On the other hand, the linear discriminant analysis (LDA) method is also used for determining whether two convex hulls have intersection. LDA is a generalization of Fisher's linear discriminant. It projects the high dimensional point sets into low dimensional space to check whether the two groups are linearly separable. Linearly separable suggests that the groups can be separated by a linear combination of features [29]. If two sets are linearly separable, then the two corresponding convex hulls have no interaction. To better understand and visualize the disjointness of convex hulls, we use the LDA method to plot two point sets in two-dimensional space.

5.7. Phylogenetic analysis and comparisons with other methods

After checking the disjointness of convex hulls between different families, we choose the center of a convex hull to represent this family. The distance between two centers can be measured by Manhattan Distance, which is defined as $d(x,y) = \Sigma_{i=1}^{250} |x_i - y_i|$. Here x and y are center points of convex hulls and could be denoted as

$x = (x_1, x_2, \cdots, x_{250})$ and $y = (y_1, y_2, \cdots, y_{250})$. In this study, we apply the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm for phylogenetic analysis. The bootstrapping method is carried out to validate the phylogenetic trees further. The bootstrapping protein sequences are taken from the original protein sequences by using sampling with replacement, which could be called bootstrap replicate. In this research, 500 bootstrap replicates for each sequence are created. By comparing the new subtrees with the original subtree, we could obtain the confidence probability of the original tree. For comparison with other clustering methods, we also perform 60-dimensional natural vector [19] and moment vector [15] methods on the same dataset.

Authors' contributions

SSTY conceived the ideas. XZ, KT, RH and SSTY designed the methodology used; XZ and KT collected and analyzed the data; XZ, KT and SSTY led the writing of the manuscript. All authors contributed critically to the draft and gave final approval for publication.

Conflict of interest

The authors have declared no conflict of interest.

Acknowledgements

The authors wish to thank Dr. Benson from Department of Computer Science, Seattle Pacific University for help with revising the manuscript, and the Department of Mathematical Science at Tsinghua University for providing the work space and library facilities. This study is supported by the National Natural Sciences Foundation of China (91746119), Tsinghua University start up fund. The authors wish to thank Tsinghua Qingfeng Scholarship (THQF2018-13). The funders did not take part in study design; in collection and analysis of data; in the writing of the manuscript; in the decision to publish this manuscript.

Availability of data and material

All the datasets used in this study could be found http://ekpd.biocuckoo.org/ and http://www.uniprot.org/.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.11.033.

References

[1] T. Hunter, Protein kinase classification, Methods Enzymol. 200 (1991) 3–37.
[2] S. Hanks, A. Quinn, Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members, Methods Enzymol. 200 (1991) 38–62.
[3] S. Julien, N. Dube, S. Hardy, M. Tremblay, Inside the human cancer tyrosine phosphatome, Nat. Rev. Cancer 11 (2011) 35–49.
[4] P. Lahiry, A. Torkamani, N. Schork, R. Hegele, Kinase mutations in human disease: interpreting genotypephenotype relationships, Nat. Rev. Genet. 11 (2010) 60–74.
[5] S. Lapenna, A. Giordano, Cell cycle kinases as therapeutic targets for cancer, Nat. Rev. Drug Discov. 8 (2009) 547–566.
[6] Z. Zhang, Protein tyrosine phosphatases: prospects for therapeutics, Curr. Opin. Chem. Biol. 5 (2001) 416–423.
[7] S. Hanks, T. Hunter, Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification, FASEB J. 9 (1995) 576–596.
[8] D. Miranda-Saavedra, G. Barton, Classification and functional annotation of eukaryotic protein kinases, Proteins Struct. Funct. Bioinforma. 68 (2007) 893–914.
[9] G. Manning, D. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The protein kinase complement of the human genome, Science 298 (2002) 1912–1934.
[10] S. Hanks, Genomic analysis of the eukaryotic protein kinase superfamily: a perspective, Genome Biol. 4 (2003) 111.
[11] L. Holm, C. Sander, Mapping the protein universe, Science 273 (1996) 595–603.
[12] S. Nepomnyachiy, N. Ben-Tal, R. Kolodny, Global view of the protein universe, Proc. Natl. Acad. Sci. 111 (2014) 11691–11696.
[13] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics 26 (2010) 680–682.
[14] R.C. Edgar, Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26 (2010) 2460–2461.
[15] S.S.-T. Yau, C. Yu, R. He, A protein map and its application, DNA Cell Biol. 27 (2008) 241–250.
[16] S.S.-T. Yau, W. Mao, M. Benson, M.R. He, Distinguishing proteins from arbitrary amino acid sequences, Sci. Rep. 5 (2015) 1–8.
[17] X. Zhao, X. Wan, R. He, S.S.-T. Yau, A new method for studying the evolutionary origin of the SAR11 clade marine bacteria, Mol. Phylogenet. Evol. 98 (2016) 271–279.
[18] X. Zhao, K. Tian, R. He, S.S.-T. Yau, Establishing the phylogeny of Prochlorococcus with a new alignment-free method, Ecol. Evol. 7 (2017) 11057–11065.
[19] C. Yu, M. Deng, S.Y. Cheng, S.C. Yau, R. He, S.S.-T. Yau, Protein space: a natural method for realizing the nature of protein universe, J. Theor. Biol. 318 (2013) 197–204.
[20] C. Yu, Q. Liang, C. Yin, R. He, S.S.-T. Yau, A novel construction of genome space with biological geometry, DNA Res. 17 (2010) 155–168.
[21] M. Deng, C. Yu, Q. Liang, R. He, S.S.-T. Yau, A novel method of characterizing genetic sequences: genome space with biological distance and applications, PLoS ONE 6 (2011) 1–9.
[22] C. Yu, T. Hernandez, H. Zheng, S.C. Yau, H.H. Huang, R. He, J. Yang, S.S.-T. Yau, Real time classification of viruses in 12 dimensions, PLoS ONE 8 (2013) 1–10.
[23] K. Tian, X. Yang, Q. Kong, C. Yin, R. He, S.S.-T. Yau, Two dimensional Yau-Hausdorff distance with applications on comparison of DNA and protein sequences, PLoS ONE 10 (2015) 1–19.
[24] K. Tian, X. Zhao, Y. Zhang, S.S.-T. Yau, Comparing protein structures and inferring functions with a novel three-dimensional Yau-Hausdorff method, J. Biomol. Struct. Dyn., DOI: https://doi.org/10.1080/07391102.2018.1540359.
[25] C.B. Anfinsen, Principles that govern the folding of protein chains, Science 181 (1973) 223–230.
[26] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, International Joint Conference on Artificial Intelligence, Vol. 14 1995, pp. 1137–1145.
[27] Y. Wang, Z. Liu, H. Cheng, T. Gao, Z. Pan, Q. Yang, A. Guo, Y. Xue, EKPD: a hierarchical database of eukaryotic protein kinases and protein phosphatases, Nucleic Acids Res. 42 (2013) D496–D502.
[28] K. Tian, X. Zhao, S.S.-T. Yau, Convex hull analysis of evolutionary and phylogenetic relationships between biological groups, J. Theor. Biol. 456 (2018) 34–40.
[29] J. Ye, R. Janardan, Q. Li, Two-dimensional linear discriminant analysis, Adv. Neural Inf. Proces. Syst. 5 (2005) 1431–1441.