Journal of Theoretical Biology 456 (2018) 34–40

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Convex hull analysis of evolutionary and phylogenetic relationships between biological groups

Kun Tian[†], Xin Zhao[†], Stephen S.-T. Yau[*]

Department of Mathematical Sciences, Tsinghua University, Beijing 100084, P.R. China

ARTICLE INFO

Article history:
Received 26 June 2018
Revised 23 July 2018
Accepted 25 July 2018
Available online 27 July 2018

Keywords:
Convex hull
Group comparison
Phylogenetic analysis
Center point
Disjoint

ABSTRACT

Comparing DNA and protein sequence groups plays an important role in biological evolutionary relationship research. Despite many methods available for sequence comparison, only a few can be used for group comparison. In this study, we propose a novel approach using convex hulls. We use statistical information contained within the sequences to represent each sequence as a point in high dimensional space. We find that the points belonging to one biological group are located in a different region of space than points belonging to other biological groups. To be more precise, the convex hull of the points from one group are disjoint from the convex hulls of points from other groups. This finding allows us to do phylogenetic analysis for groups in an efficient way. Five different theorems are presented for checking whether two convex hulls intersect or are disjoint. Test results for datasets related to HRV, HPV, Ebolavirus, PKC and protein phosphatase domains demonstrate that our method performs well and provides a new tool for studying group phylogeny. More significantly, the convex analysis presents a new way to search for sequences belonging to a biological group by examining points within the group's convex hull.

1. Introduction

Evolutionary and phylogenetic analysis of DNA and protein groups is a basic task that has been studied in biology for years. It is important to understand the natural relationships between groups, such as families, species, or different biological types. Many approaches have been proposed for sequence comparison in the past few decades (Elloumi, 1998; Kantorovitz et al., 2007; Campello and Hruschka, 2009; Sims et al., 2009; Povolotskaya and Kondrashov, 2010), but only a few can be applied to the phylogenetic analysis of groups. Traditionally, most comparison methods are based on multiple alignment, by using dynamic programming techniques to identify the globally optimal alignment solution (Altschul et al., 1997). Unfortunately, multiple alignment is an NP-hard problem, which means in practice that the implementations of these algorithms run slowly and use large amounts of memory. Furthermore, it can't be used to compare groups. Recently, alignment-free approaches based on features descriptor or statistical properties of the sequences have attracted more and more attention. For example, to avoid complete loss of sequence pattern, the PseKNC and PseAAC methods are developed to reflect the core and essential features that are deeply hidden in sequences (Lin et al., 2014; Jia et al., 2016). These methods are used to cluster sequences and predict their various attributes. The graphical representation (Yau et al., 2003, 2008; Yu et al., 2010), the k-mer methods (Vinga and Almeida, 2003) and the natural vector methods (Deng et al., 2011; Yu et al., 2013; Zhao et al., 2016) provide different ways to represent sequences as points in high dimensional space according to their statistical characteristics. Metrics such as the Hausdorff distance (Huttenlocher et al., 1993; Chew et al., 1997; Yu et al., 2014; Tian et al., 2015; Zhao et al., 2017) are used for measuring the similarity between point sets representing the corresponding sequence groups. Note that calculating the Hausdorff distance matrix requires considerable CPU time and memory as the size of the groups increases.

In this study, we establish a new approach for performing evolutionary and phylogenetic analysis of biological sequence groups using convex hulls. Based on the natural vector method originated by Deng et al. (2011), each sequence is converted into a vector. The vector contains the occurrence frequencies, the average positions and the central moments of the four nucleotides or twenty amino acids. If the convex hulls of any two groups do not intersect, we know that the two groups are located in different regions of high dimensional space. A central vector in each group is chosen to represent the spatial position of the group.

Then the question remains how to determine whether two convex hulls constructed by two finite point sets intersect or not. Let $A = \{a_1, a_2, \cdots, a_n\}$, $B = \{b_1, b_2, \cdots, b_m\}$ be two finite point sets in

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

https://doi.org/10.1016/j.jtbi.2018.07.035
0022-5193/

R^k. Assume S is the convex hull function. The problem is to determine whether the two convex hulls S(A) and S(B) have intersection. Although researchers have focused on this problem for years, the complexity of the known algorithms is high when the dimension k of the space is large. In the Materials and Methods section, we present five theorems for solving this problem. The proofs of all these methods could be found in the Supplement materials.

To validate the advantage of approach, in the Results and discussion section, we test it on several biological sequence datasets and compare it with the Hausdorff method. The phylogenetic trees show that our method give results that conform better to accepted evolutionary and phylogenetic analysis. The high bootstrap values and high accuracy indicate the efficiency of our new convex analysis approach. We also present several graphs generated using our method to easily visualize the convex hulls of different group datasets.

2. Materials and methods

2.1. Natural vector method

Let $S = (s_1, s_2, s_3, \cdots, s_n)$ be a DNA sequence of length n, that is, $s_i \in \{A,\ C,\ G,\ T\},\ i = 1, 2, 3,\ \cdots n$. For each of the 4 nucleotides k, define

$$w_k(\cdot) : \{A, C, G, T\} \to \{0, 1\}$$

such that $w_k(s_i) = 1$ if $s_i = k$ and $w_k(s_i) = 0$ otherwise.

(1) Let $n_k = \sum_{i=1}^{n} w_k(s_i)$ be the number of nucleotide k in the DNA sequence S.
(2) Let $s_{[k][i]} = i \cdot w_k(s_i)$ be the distance from the first nucleotide (regarded as origin) to the ith nucleotide k in the DNA sequence.
(3) Let $T_k = \sum_{i=1}^{n_k} s_{[k][i]}$ be the total distance of each set of the 4 nucleotides.
(4) We then take $\mu_k = T_k/n_k$ as the mean position of the nucleotide k.
(5) Finally, we define the second-order normalized central moments as follows:

$$D_2^k = \sum_{i=1}^{n_k} \frac{\left(s_{[k][i]} - \mu_k\right)^2}{n_k n}.$$

Then the natural vector of the DNA sequence S is given as follows:

$$(n_A,\ \mu_A,\ D_2^A,\ n_C,\ \mu_C,\ D_2^C,\ n_G,\ \mu_G,\ D_2^G,\ n_T,\ \mu_T,\ D_2^T).$$

Similarly, protein sequence could be represented by 60-dimension natural vector using the same definition.

Given a biological group G with N sequences, we can obtain a set containing N points $A = \{a_1, a_2, \cdots, a_N\}$ corresponding to these sequences based on the above natural vector method. Let $a_0 = \sum_{i=1}^{N} a_i/N$ be the center point of group G. Then the difference between two groups is defined as the Euclidean distance of their center points. The phylogenetic tree is constructed by the distance matrix using UPGMA algorithm.

In the next part of this section, we introduce five different methods to check whether two convex hulls intersect or not in high dimensional space. The details of the proofs could be found in the Supplement materials.

2.2. Projection-line method

Let $A = \{a_1, a_2, \cdots, a_n\}$, $B = \{b_1, b_2, \cdots, b_m\}$ be two finite point sets in R^k. Assume S is the convex hull function. Then $S(A) \cap S(B) = \emptyset$ is equivalent with that there is a line $l \subset R^k$, for the projection sets $P(A)$, $P(B)$ of A, B in l, s.t. $S(P(A)) \cap S(P(B)) = \emptyset$.

This means that if we can find any line such that the two segments of the projection sets $P(A)$ and $P(B)$ are disjoint, then the convex hulls of the original point sets A and B have no intersection. The computation is greatly reduced since we transform the problem from k-dimensional to one-dimensional.

2.3. Normal vector method

Let $A = \{a_1, a_2, \cdots, a_n\}$, $B = \{b_1, b_2, \cdots, b_m\}$ be two finite point sets in R^k. Assume S is the convex hull function. Then the necessary and sufficient condition of $S(A) \cap S(B) = \emptyset$ is that there is a normal vector N of one hyperplane of $S(A)$ and $S(B)$, for the projection sets $P(A)$, $P(B)$ of A, B in line N, s.t. $S(P(A)) \cap S(P(B)) = \emptyset$.

This theorem could give confirmatory result after checking all the possible normal vectors since the number of normal vectors for any convex hull is finite. One can treat this method as a special case of the first theorem with given position of projection-line.

2.4. Subset determination method

Let $A = \{a_1, a_2, \cdots, a_n\}$, $B = \{b_1, b_2, \cdots, b_m\}$ be two finite point sets in R^k. Assume S is the convex hull function. Then the necessary and sufficient condition of $S(A) \cap S(B) = \emptyset$ is that for all the possible integers $i_1,\ i_2, \ldots,\ i_{k+1} \in [1, n]$ and $j_1,\ j_2, \ldots,\ j_{k+1} \in [1, m]$, $S(\{a_{i_1}, a_{i_2}, \ldots, a_{i_{k+1}}\}) \cap S(\{b_{j_1}, b_{j_2}, \ldots, b_{j_{k+1}}\}) = \emptyset$.

According to this method, we can divide each convex hull into several convex blocks constructed by $k+1$ points and check whether these small blocks have intersection. In k-dimensional space, each of the convex block is composed of $k+1$ vertices and $k+1$ faces with any possible k vertices. The equations of each $k+1$ faces and corresponding normal vectors of the convex block can be easily computed. It helps us to determine whether each pair of this kind of small blocks are disjoint or not based on the normal vector method in a simple way. Therefore, the computation is also significantly reduced.

2.5. Linear programming method

Let $A = \{a_1, a_2, \cdots, a_n\}$, $B = \{b_1, b_2, \cdots, b_m\}$ be two finite point sets in R^k. Assume S is the convex hull function. Then $S(A) \cap S(B) = \emptyset$ is equivalent with that there are no nonnegative real numbers $\lambda_1,\ \lambda_2, \ldots,\ \lambda_n,\ \mu_1,\ \mu_2, \ldots,\ \mu_m$ s.t. $\sum_{i=1}^{n} \lambda_i a_i = \sum_{j=1}^{m} \mu_j b_j$ and $\sum_{i=1}^{n} \lambda_i = \sum_{j=1}^{m} \mu_j = 1$.

We can transform the original problem into an algebra problem by this theorem. If any convex combination of the points in one set equals to that of points in the other set, we then confirm that the two hulls have intersection. No matter how large the dimension of the space is and how many the points are, we can always solve this problem easily by the linear programming function in many kinds of software. It is a very timesaving and effective method.

2.6. Minimum distance method

Let $A = \{a_1, a_2, \cdots, a_n\}$, $B = \{b_1, b_2, \cdots, b_m\}$ be two finite point sets in R^k. Assume S is the convex hull function. For nonnegative real numbers $\lambda_1,\ \lambda_2, \ldots,\ \lambda_n,\ \mu_1,\ \mu_2, \ldots,\ \mu_m$ satisfy $\sum_{i=1}^{n} \lambda_i = \sum_{j=1}^{m} \mu_j = 1$, and Let $D = inf|\sum_{i=1}^{n} \lambda_i a_i - \sum_{j=1}^{m} \mu_j b_j|$. Then the necessary and sufficient condition of $S(A) \cap S(B) = \emptyset$ is that $D > 0$.

Here we translate the problem to another algebra question about calculating the minimum distance of the two convex hulls. They are disjoint if and only if the minimum distance is positive. Many mathematical software could easily solve this minimization problem with quadratic programming functions.

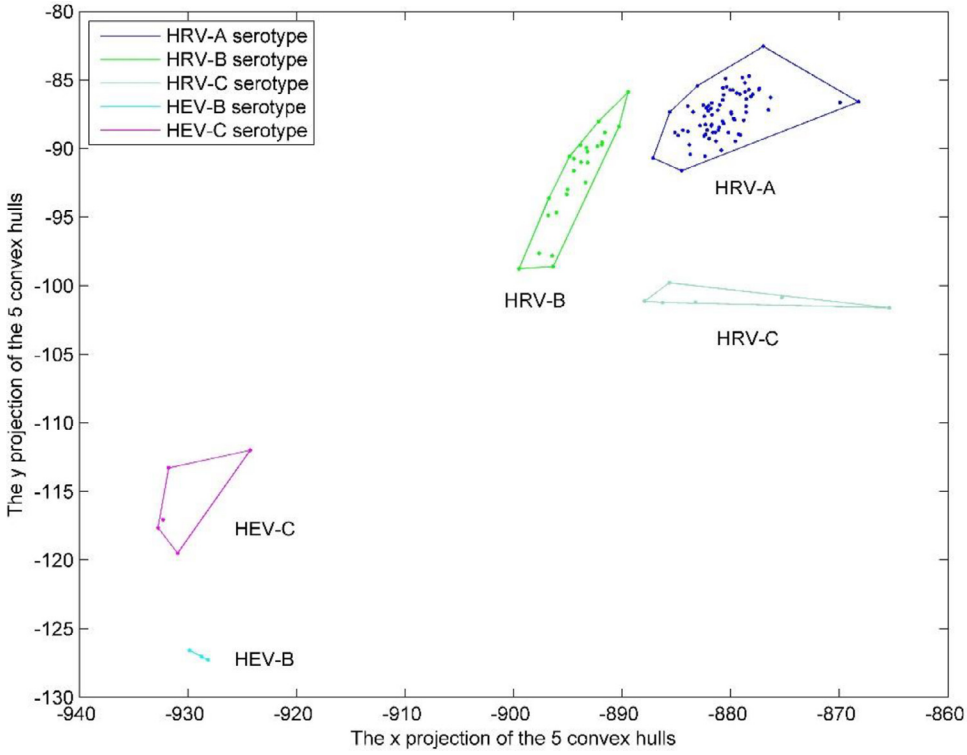

Fig. 1. The convex hulls of the 5 HRV and HEV serotypes are mutually disjoint.

3. Results and discussion

3.1. Human rhinovirus

Firstly, we use the 113 human rhinovirus genomes dataset (HRVs) with about 7165 nucleotides to verify the disjointness for convex hulls of different groups. These 113 genomes could be divided into 5 distinct serotypes: HRV-A, HRV-B, HRV-C, HEV-B and HEV-C. Here the HEV-B and HEV-C serotypes are treated as outgroups (Deng et al., 2011). After calculating the 12-dimensional natural vectors of all these whole genomes, we obtain 5 convex hulls based on the points of each serotype. All the five theorems in the Materials and Methods section can be used to prove that each pair of the five convex hulls is disjoint. In order to view the disjointness of these five groups, we project the 5 convex hulls into two-dimensional space. Each time we choose a two-dimensional plane randomly until any pair of the 5 projection polygons is disjoint in this plane. Since the 5 convex hulls are disjoint mutually in 12-dimensional space, the two-dimensional plane must exist. In general, we can first find such plane which satisfies that the projection of two biggest convex hulls have no intersection, then adjust the spatial position of the plane to make all the projections are disjoint mutually. This is shown in Fig. 1. In this figure, these 5 serotypes are separated from each other clearly. This means that the natural vectors of different groups lie in different areas of high dimensional Euclidean space, which provides a new useful convex analysis tool in evolutionary relationship and phylogenetic analysis.

3.2. Human papillomavirus

To test the performance of our convex analysis method, we apply it on the 400 human papillomavirus genomes dataset (HPVs) containing about 7914 nucleotides in length which are divided into 12 genotypes: HPV type 11, 16, 18, 31, 33, 35, 45, 52, 53, 58, 6, 66. Human papillomavirus plays an important role in the research of the second most common cancer––cervical cancer (Arbyn et al., 2011). Different types of HPVs can lead to different levels of risk. Therefore, it has crucial significance to cluster HPV types into low and high risk types. Although many methods have been provided to classify the HPV risk types, these genotypes are hardly explored for many reasons, for example, some genotypes have low amplification signals. Here we use the convex analysis approach to cluster these 12 HPV types. The disjointness of these 12 convex hulls constructed by the 12-dimensional natural vector points for each genotype is shown in Fig. 2. In this work, the center point in each convex hull is chosen to represent the corresponding HPV type. The matrix is computed based on the Euclidean distances between the 12 center points. The constructed phylogenetic tree using UPGMA method is shown in Fig. 3. For comparison, we also use the Hausdorff distance method to construct the tree which is shown in Fig. 4. In the previous work, Smith states that HPV types 16, 18, 45, 31, 33, 52, 58, 35 are regarded as high risk genotypes, and the other types 6, 11, 53, 66 have low risk (Smith et al., 2007). Our results are in accordance with that of Smith, while the low risk genotypes can not be distinguished by the Hausdorff method. For example, the low risk HPV types 6 and 11 are correctly clustered into one group by our method rather than far from each other using the Hausdorff approach. This shows that our method performs better in the HPV type dataset. On the other hand, the bootstrapping method (Hillis and Bull, 1993) is used for computing the confidence probabilities on phylogenetic tree. Bootstrapping is a common test used in phylogenetics to estimate the significance of the branches in a tree. The bootstrapping sequences are taken from the original sequences by using sampling with replacement, which can be called bootstrap replicate. In this paper, 100 bootstrap replicates for each sequence are created. We then compared the new subtrees with the original subtree and obtained the confidence probability of the original tree. In this example, the bootstrap values in our tree are all 100%, by comparison, the average value of Hausdorff tree is only 41.9%. These results verify that our

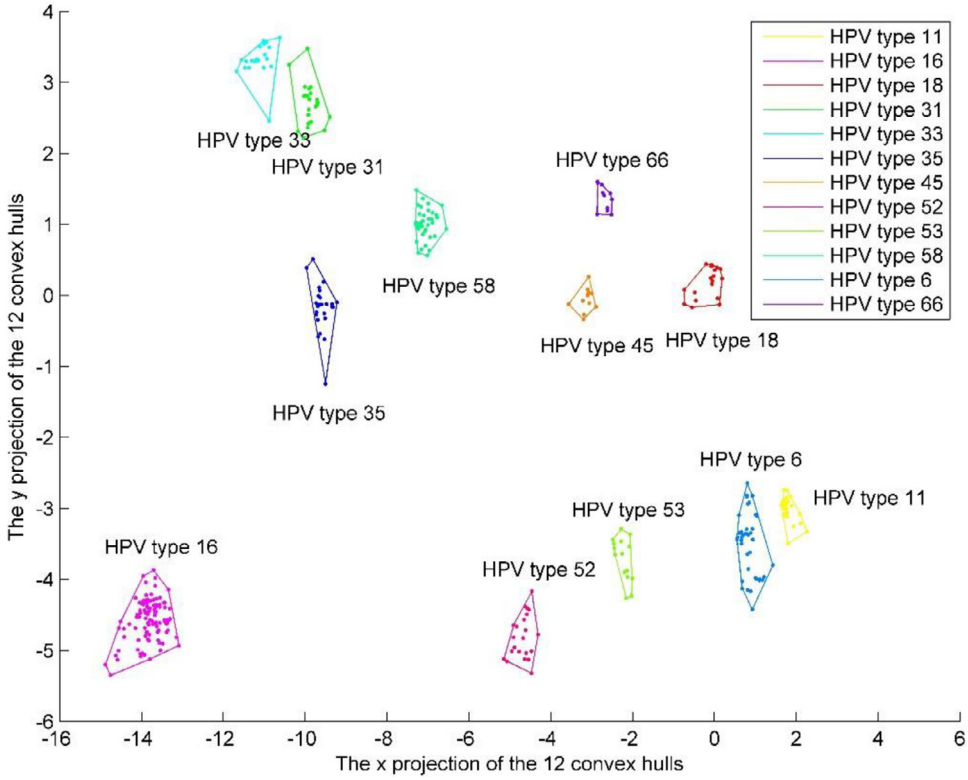

Fig. 2. The convex hulls of the 12 HPV genotypes are mutually disjoint.

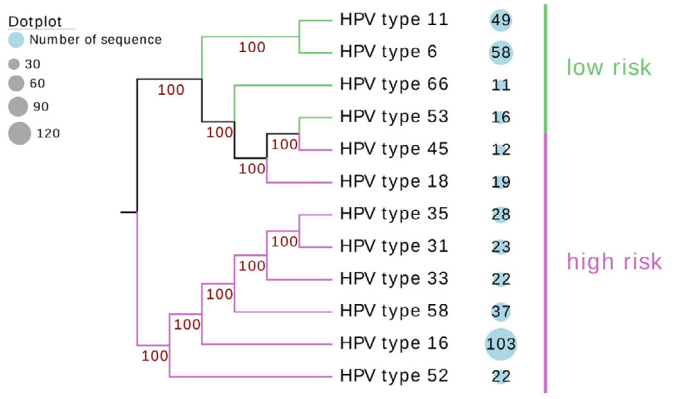

Fig. 3. The phylogenetic tree of the 12 HPV genotypes by our method. The tree is constructed using UPGMA algorithm based on the Euclidean distances between the centers of 12 convex hulls. The number of sequences for each group is presented besides the tree. The bootstrap confidence values are generated using 100 permutations.

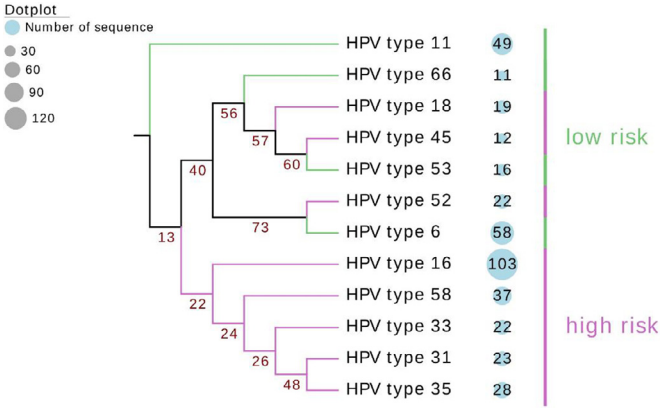

Fig. 4. The phylogenetic tree of the 12 HPV genotypes by Hausdorff method. The tree is constructed using UPGMA algorithm based on the Hausdorff distances of these 12 genotypes. The number of sequences for each group is presented besides the tree. The bootstrap confidence values are generated using 100 permutations.

methods applied on this dataset are stable and convincing with high accuracy.

3.3. Ebolavirus

We also apply the convex analysis method on the Ebolavirus dataset with about 18,936 nucleotides in length. This dataset contains 68 Ebolaviruses which consists of 34 Ebola virus (EBOV), 11 Sudan virus (SUDV), 9 Reston virus (RESTV), 1 Tai Forest virus (TAFV), 6 Bundibugyo virus (BDBV), 6 Marburg virus (MARV) and 1 Lloviuvirus (LLOV) (Zheng et al., 2015). We prove that the 7 convex hulls of these species are mutually disjoint based on the theorems in the Materials and Methods section. Furthermore, the biggest Ebola virus species is composed of 4 small groups: Zaire Ebola virus strain Mayinga, Zaire ebolavirus isolate EBOV, Zaire ebolavirus isolate H.sapiens-wt, and Zaire ebolavirus isolate Ebola virus. The phylogenetic tree constructed by the 6 species and 4 EBOV groups is displayed in Fig. 5. We could see that the 4 EBOV groups cluster together. The SUDV branch is clustered with the EBOV and RESTV branches. BDBV and TAFV viruses are grouped together. These results are consistent with those in previous research.

3.4. Protein kinase c

To assess our method on phylogenetic analysis of protein sequences, we apply it to classify the 124 protein kinase C (PKC) family dataset. The average length of these 124 sequences is 789. The

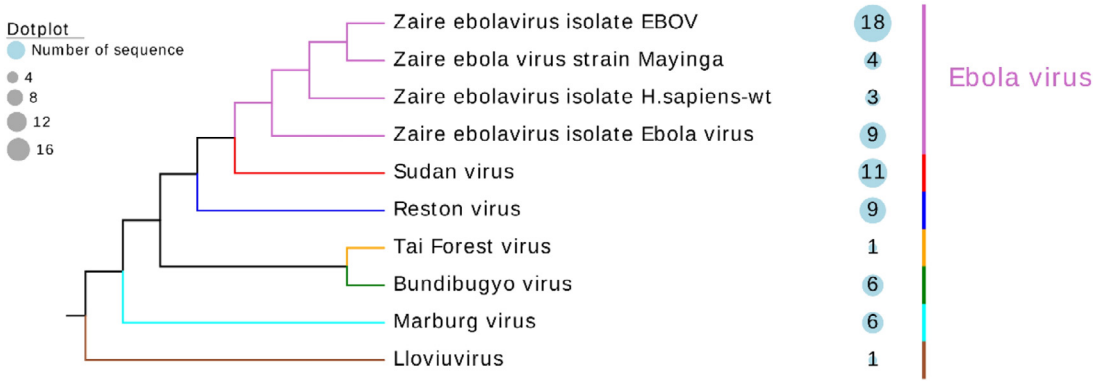

Fig. 5. The phylogenetic tree of the Ebolavirus species. The tree is constructed using UPGMA algorithm based on the Euclidean distances between the centers of convex hulls. The number of sequences for each group is presented besides the tree.

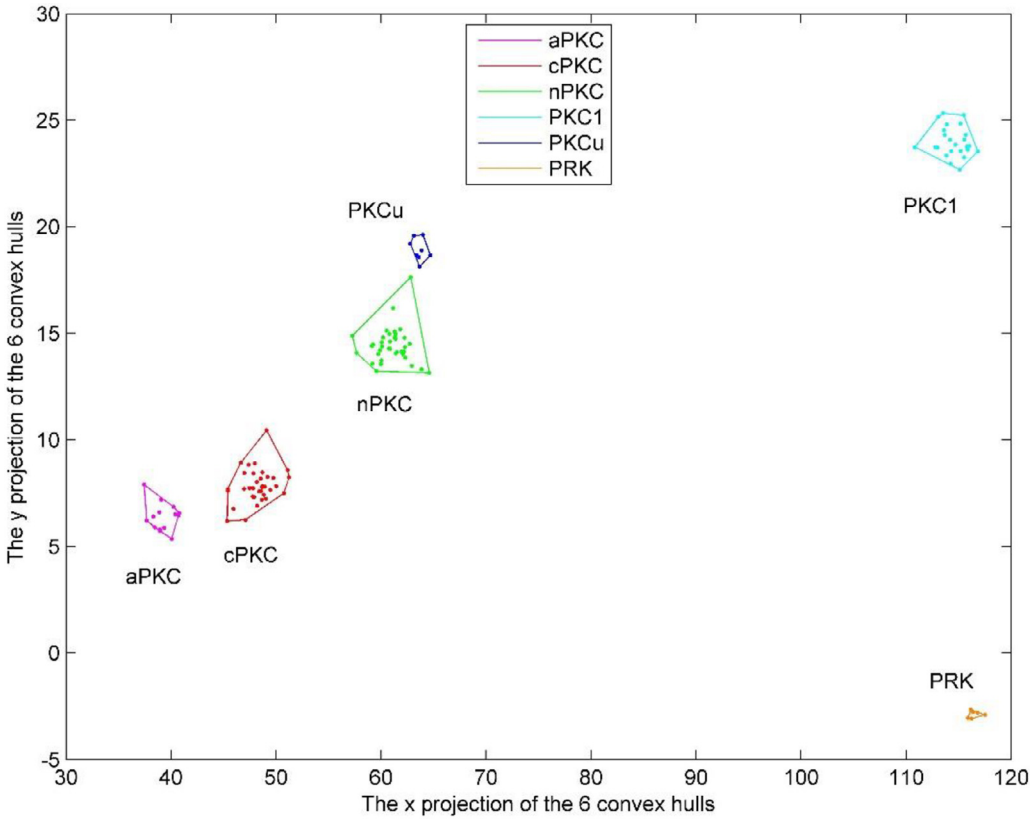

Fig. 6. The convex hulls of the 6 PKC subfamilies are mutually disjoint.

protein kinase C family is a large group of enzymes regulating the Ca^{2+}-dependent pathways in cells (Nishizuka, 1986). PKC is classified into six subfamilies: aPKC, cPKC, nPKC, PKC1, PKCμ and PRK. The 6 convex hulls are constructed by the 60-dimensional natural vectors that represent the protein sequences of these subfamilies in R^{60} space. Each pair of the 6 hulls is demonstrated to be disjoint. The two-dimensional projection for visualizing the disjointness of all the 6 convex hulls is shown in Fig. 6. The center points of these 6 convex hulls are used for representing the corresponding subfamily as well as building the phylogenetic tree. As shown in Fig. 7, the aPKC, cPKC and nPKC are clustered together and positioned away from the PRK subfamily. This shows our convex analysis method characterizes the relationship between proteins in a way that is closer to the actual nature of the proteins.

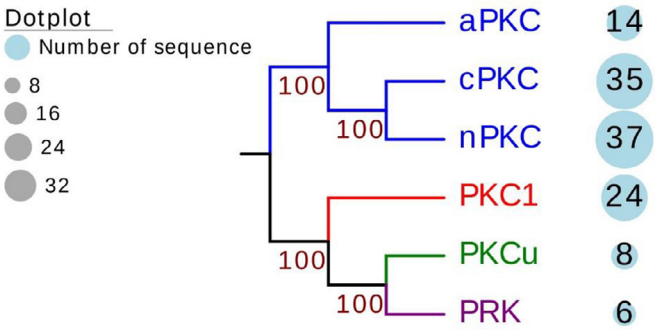

Fig. 7. The phylogenetic tree of the 6 PKC subfamilies. The tree is constructed using UPGMA algorithm based on the Euclidean distances between the centers of convex hulls. The number of sequences for each group is presented besides the tree. The bootstrap confidence values are generated using 100 permutations.

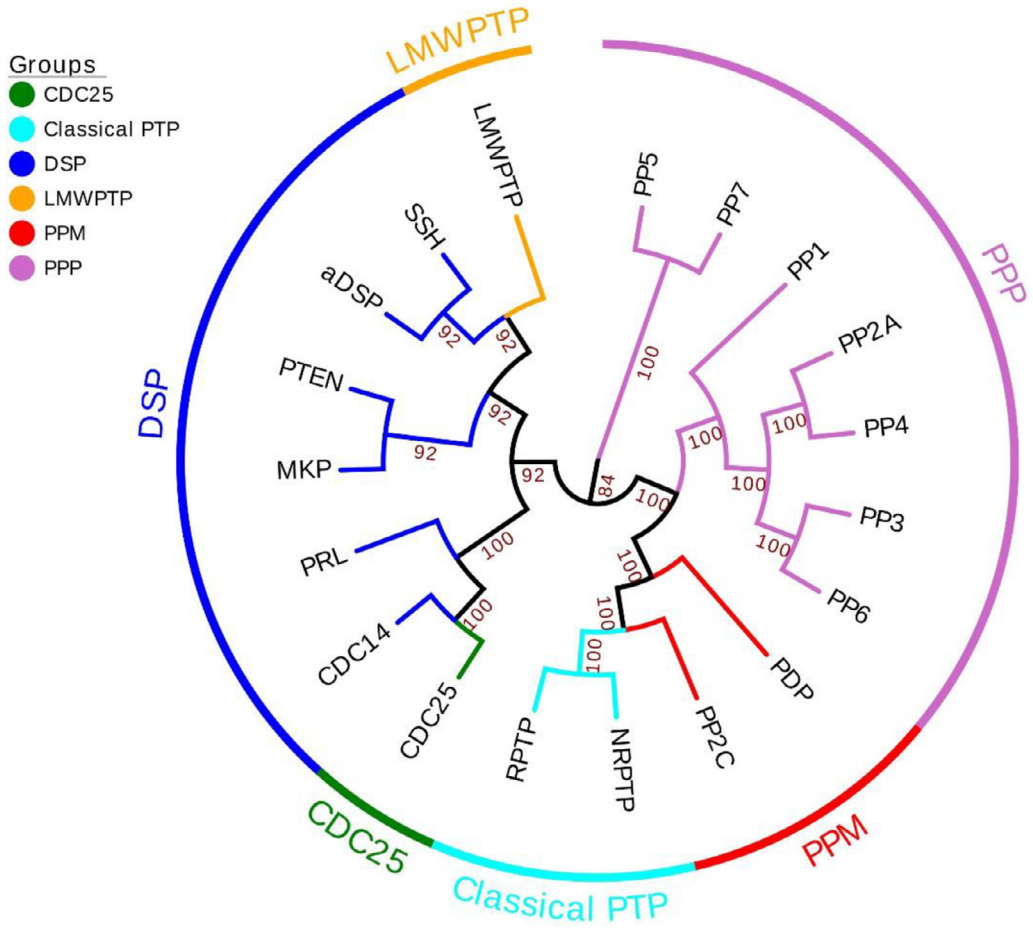

Fig. 8. The phylogenetic tree of the 19 protein phosphatase domain families. The tree is constructed using UPGMA algorithm based on the Euclidean distances between the centers of convex hulls. The bootstrap confidence values are generated using 100 permutations.

3.5. Protein phosphatase domain

Phosphorylation by protein kinases is recognized as an important mechanism playing a role in virtually every activity of eukaryotic cells. The classification of protein phosphatase sequences has gained increasing attention in biological sciences. We analyze phosphatase domain dataset (Wang et al., 2013) containing 5802 protein sequences which are divided into 19 families from 62 species. These 19 families belong to 6 groups: CDC25, Classical PTP, DSP, LMWPTP, PPM and PPP. As displayed in Fig. 8, these 19 families are well clustered using our method. Except for the CDC25 and LMWPTP groups that each of them contains only one family, the Classical PTP, DSP, PPM and PPP groups are all gathered together. For example, the 6 families of the dual specificity phosphatase group with blue color are classified together correctly. The 7 well-known enzyme families PP1, PP2A, PP3, PP4, PP5, PP6 and PP7 with pink color are clustered closed and form the PPP group. Therefore, our method produces accurate and effective classification results on phylogenetic analysis.

4. Conclusions

This article proposes a new convex analysis approach for comparing and classifying DNA and protein groups. We introduce five different approaches to check whether two convex hulls intersect or not in high dimensional space, which are treated as the basics of our method: projection-line method, normal vector method, subset determination method, linear programming method and minimum distance method. Given some groups containing DNA or protein sequences, we first calculate the natural vectors of sequences in each group. We then use these five approaches to check whether each pair of the convex hulls constructed by the groups are disjoint. If so, that means the natural vector points belonging to different groups are located in different regions of high dimensional Euclidean space, which provides a new useful convex analysis tool in evolutionary research. Then the center point in each group is chosen to represent this group. The Euclidean distances between these center points are computed for obtaining a distance matrix that contains information about these groups.

In this study, we test the performance of our method on five datasets, including the HRV, HPV, Ebolavirus, PKC and protein phosphatase domain datasets. The convex hulls of each group are mutually disjoint and the phylogenetic trees are reconstructed according to the distance matrix. This method produces accurate and reasonable clustering results as well as high reduced computation, suggesting the potential utility of the approach we describe in constructing the phylogeny in an efficient manner. Using convex analysis method, we could study evolutionary relationships between biological groups. It provides us new insights of analyzing evolutionary relationships and phylogeny among groups in molecular biological study.

Availability of data

All the datasets used in this study could be found in the Supplement materials.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

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

Acknowledgments

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 Science 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.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jtbi.2018.07.035.

References

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Arbyn, M., Castellsague, X., Sanjose, S., Bruni, L., Saraiya, M., Bray, F., Ferlay, J., 2011. Worldwide burden of cervical cancer in 2008. Ann. Oncol. 22, 2675–2686.

Campello, R., Hruschka, E., 2009. On comparing two sequences of numbers and its applications to clustering analysis. Inf. Sci. 179, 1025–1039.

Chew, L., Goodrich, M., Huttenlocher, D., Kedem, K., Kleinberg, J., Kravets, D., 1997. Geometric pattern matching under Euclidean motion. Comput. Geometry 7, 113–124.

Deng, M., Yu, C., Liang, Q., He, R., Yau, S., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS ONE 6, 1–9.

Elloumi, M., 1998. Comparison of strings belonging to the same family. Inf. Sci. 111, 49–63.

Hillis, D.M., Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42 (2), 182–192.

Huttenlocher, D., Klanderman, G., Rucklidge, W., 1993. Comparing images using the Hausdorff distance. Pattern Anal. Mach. Intell. 15, 850–863.

Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.C., 2016. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J. Theor. Biol. 394, 223–230.

Kantorovitz, R., Robinson, E., Sinha, S., 2007. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23, 249–255.

Lin, H., Deng, E.Z., Ding, H., Chen, W., Chou, K.C., 2014. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 42, 12961–12972.

Nishizuka, Y., 1986. Studies and perspectives of protein kinase C. Science 233, 305–312.

Povolotskaya, I., Kondrashov, F., 2010. Sequence space and the ongoing expansion of the protein universe. Nature 465, 922–926.

Sims, G., Jun, S., Wu, G., Kim, S., 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. USA 106, 2677–2682.

Smith, J., Lindsay, L., Hoots, B., Keys, J., Franceschi, S., Winer, R., Clifford, G., 2007. Human papillomavirus type distribution in invasive cervical cancer and high–grade cervical lesions: a meta-analysis update. Int. J. Cancer 121, 621–632.

Tian, K., Yang, X., Kong, Q., Yin, C., He, R., Yau, S., 2015. Two dimensional Yau-Hausdorff distance with applications on comparison of DNA and protein sequences. PLoS ONE 10, 1–19.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison–a review. Bioinformatics 19, 513–523.

Wang, Y., Liu, Z., Cheng, H., Gao, T., Pan, Z., Yang, Q., Guo, A., Xue, Y., 2013. EKPD: a hierarchical database of eukaryotic protein kinases and phosphatases. Nucleic Acids Res. 42, D496–D502.

Yau, S., Wang, J., Niknejad, A., Lu, C., Jin, N., Ho, Y., 2003. DNA sequence representation without degeneracy. Nucleic Acids Res. 31, 3078–3080.

Yau, S., Yu, C., He, R., 2008. A protein map and its application. DNA Cell Biol. 27, 241–250.

Yu, C., Liang, Q., Yin, C., He, R., Yau, S., 2010. A novel construction of genome space with biological geometry. DNA Res. 17, 155–168.

Yu, C., Deng, M., Cheng, S., Yau, SC., He, R., Yau, S., 2013. Protein space: a natural method for realizing the nature of protein universe. J. Theor. Biol. 318, 197–204.

Yu, C., He, R., Yau, S., 2014. Viral genome phylogeny based on Lempel-Ziv complexity and Hausdorff distance. J. Theor. Biol. 348, 12–20.

Zhao, X., Wan, X., He, R., Yau, S., 2016. A new method for studying the evolutionary origin of the SAR11 clade marine bacteria. Mol. Phylogenet. Evol. 98, 271–279.

Zhao, X., Tian, K., He, R., Yau, S., 2017. Establishing the phylogeny of Prochlorococcus with a new alignment-free method. Ecol. Evol. 7, 11057–11065.

Zheng, H., Yin, C., Hoang, T., He, R., Yang, J., Yau, S., 2015. Ebolavirus classification based on natural vectors. DNA Cell Biol. 34, 418–428.