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.