A novel approach on protein classification

Xiaogeng Wan[1], Xin Zhao[1] and Stephen S. -T Yau[1,*]

[1] Department of Mathematical Sciences, Tsinghua University, 100084, Beijing, China.

* **Corresponding author.**

E-mail: wxgbj88@sina.com, zhao-x15@mails.tsinghua.edu.cn, yau@uic.edu (S. S. -T. Yau).

Abstract

Protein universe is a complex system with critical problem of protein evolution to be analyzed. Early studies have used geometric distances and polygenetic-trees to solve this problem. However, the traditional methods are bivariate, whose taxonomy classification relies on bivariate branching. This is not sufficient to describe the complex nature of protein universe. Therefore, we propose a novel approach on multivariate protein classification. The new method bases on the theory of information and network, can be used to analyze multivariate relationships of proteins. The new method is alignment-free and have wide-applications to both sequences and 3D structures. We demonstrate the new method on six protein examples, results show that the new method is efficient and can potentially be used for future protein classifications.

Keywords: Protein classification, information, network, sequence, evolution.

Introduction

The protein universe is diverse and has long been a mysterious entity and essential underpinning in biology [9, 21]. In protein universe, the protein sequences can be branched into families of fine hierarchy. Many researchers have spared their efforts to develop methods for future classification of unknown-lineage proteins [5, 6, 9, 12, 15, 19-21]. Early studies have used geometric methods in combination of protein biological nature to explore the universe of proteins. Prevalent idea is to use amino acid sequence homology to calculate their biological distances and draw polygenetic-trees according to the distances. They believe that sequence homology is highly related to protein relationships [6]. The polygenetic-trees show bivariate branching of protein lineages. Typical methods of this kind are the natural vector [21, 24, 25], protein map [19,20], K-string dictionary [22] and Yau-Hausdorff distance [15].

These early methods represent protein relationships in a bivariate manner. However, these methods may not be sufficient to describe the comprehensive nature of protein universe, in that protein relationships may not be only bivariate. In other words, in a big family of species, a parent species may not necessarily be evolved into exactly two children species, and one species may not necessarily has only one sister or brother species. In fact, like all other natural systems [14, 17], one protein may have multivariate connections to more than one other proteins. To reveal a more natural picture of protein evolution, one needs to globally survey the multivariate relationships of proteins. In this paper, we use networks [11] to model the space of proteins, in which each protein is a node, we aim to use network tools to analyze the global relationships of the protein nodes [14, 16, 17]. The theory of information and network provides ready tools to analyze the model of protein universe, where we aim to use property of networks to draw new global picture of protein universe.

The paper is divided into five parts. This section is an introduction to the study. In the next section, we describe the materials and methods of the new method. The third section describes six examples to demonstrate the application and efficiency of the method, where we present pictures on protein taxonomy classifications. The fourth and fifth sections are the discussion and conclusion to this paper, where we discuss and conclude the efficiency and properties of the new method.

Materials and methods

We combine information and network theories to develop a new approach in identifying global protein relationships. Protein amino acid sequence can be viewed as discrete time series, where the amino acid order is time and the species of the 20 amino acid are states. To start with, we map the amino acid sequence to integer sequence with states from 1 to 20. Since information theoretic measures are independent of the label of states [22], using different labels will not change the result. The discrete time series of integers are taken as inputs to the new global connectivity method.

The maximum mutual information rates

Before using the information theoretic measure, we first map the amino acid sequences into discrete time series. Each amino acid $a \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$ is uniquely mapped to an integer $b \in \{1,2,3,4...,20\}$. All protein sequences are transformed to sequences of discrete integers. For each pair of the integer sequences $x_1x_2 \cdots x_M$ and $y_1y_2 \cdots y_N$ (M and N denote the length of protein), without loss of generality, assuming $M \leq N$, we pick the length N segment for the longer sequence M, and calculate the mutual information [2,4,16,26] between $x_1x_2 \cdots x_M$ and $y_iy_{i+1} \cdots y_{i+M-1}$, for $1 \leq i \leq N - M + 1$,

$$I(X;Y_i)=\sum_{x\in S_x,y\in S_y} p(x_n=x,y_{n+i-1}=y)\log\frac{p(x_n=x,y_{n+i-1}=y)}{p(x_n=x)p(y_{n+i-1}=y)}\frac{1}{2}, \tag{1}$$

where X is the shorter sequence and Y_i is the length M segment of the longer sequences, S_x and S_y denote the state sets of the sequence X and Y respectively, which are subsets of positive integers with elements from 1 to 20.

Mutual information rate describes the mutual relationship between two proteins. Shifting i from 1 to $N-M+1$, we obtain a sequence of mutual information rates denoted as $I_1,\cdots,I_{N-M+1}$. Here we extract the maximum mutual information rates between X and Y,

$$I_{\max,xy}=\sum_{1\le i\le N-M+1} I(X;Y_i) \tag{2}$$

We set the maximum mutual information rate $I_{max,XY}$ as the (X,Y) elements of the adjacency matrix. Note that the mutual information rates are symmetric such that the adjacency elements

$$a_{XY}=a_{YX}=I_{max,XY}=I_{max,YX}. \tag{3}$$

We use the symmetric maximum mutual information rates as elements of the adjacency matrix to construct simple undirected protein network [11].

Without loss of generality, assume there are K protein sequences in a set. The adjacency matrix is a $K\times K$ symmetric matrix. The maximum mutual information rate for sequences X and Y is reached for some $1\le k\le N-M+1$:

$$I_{max,XY}=I(X;Y_k). \tag{4}$$

Knowing that the mutual information rate is up bounded by entropy [2,4,26]:

$$I(X;Y)\le \min\{H(X),H(Y)\}, \tag{5}$$

where $H(X)=-\sum_{x\in S_x} p(x_n=x)\log p(x_n=x)$ and $H(Y)=-\sum_{y\in S_y} p(y_n=y)\log p(y_n=y)$, thus

$$I_{max,XY}=I(X;Y_k)\le \min\{H(X),H(Y_k)\}. \tag{6}$$

Additionally, the entropy is non-decreasing as state number increases, we have

$$H(Y_k)\le H(Y), \tag{7}$$

because the number of states in Y is no less than the number of states in Y_k. Then, we have

$$I_{max,XY}\le \min\{H(X),H(Y)\}. \tag{8}$$

To make fair threshold, we normalize the adjacency matrix in terms of the entropies. Denote the sequences as $X_1,X_2,\cdots,X_K$, elements of the adjacency matrix now become:

$$a_{ij} = a_{X_i,X_j} = \frac{I_{max,X_iX_j}}{\max\limits_{k} H(X_k)}, \tag{9}$$

all elements of the new adjacency matrix are bounded between 0 and 1.

Connect component

Connect component is a basic concept in network theory, which is a good method for clustering. To unveil the evolutionary relationship among proteins, we set up a threshold to filter the corrected adjacency matrix. Elements below the threshold are set to zero, with the rest elements are unchanged. The threshold is defined as constant multiple of the maximum adjacency element, more specific, denote

$$T_c = c \cdot \max_{i,j} a_{ij} \tag{10}$$

as the threshold at multiplicity c, where the multiplicity c takes uniform distributed values from 0.1 to 1, with interval of 0.01. For each multiplicity c, we filter the corrected adjacency matrix A, the connect components of the protein network are the sets of proteins whose adjacency elements are all non-vanishing, the inclusion of any other new proteins in the set will break the law (i.e. introduce vanishing adjacency elements). In other words, connect component is a subset of all vertices in a network preserving the connection criterion. The criterion requires that each member of the subset has at least one path connecting to any other member of the subset, where the path is the joint of links that are connected end to end. The connect components are maximum, because no other vertex in the network can be added to the subset while preserving this property [11]. Connect components of undirected networks are called weakly connect components, to distinguish from the strongly connect components of directed networks. Nodes in one connect component are highly related to each other.

The members of the connect components are all mutually connected by at least one path in the network, no matter the length of the path. By nature of the undirected networks, in a network of n nodes, the lengths of such paths should be no longer than $n-1$. In matrix form, there exists a path from node j to node i of length m ($\leq n$) if and only if the (i,j)-th element in the power m adjacency matrix A^m is positive [11]. To identify the connect components, we need to find all such paths from length 1 to length $n-1$. In matrix notation, denote the sum of the 1 to $n-1$ power adjacency matrix as

$$A_{sum} = \sum_{i=1,\cdots,n-1} A^i$$

via reversible matrix transformations, the matrix A_{sum} can be written in block diagonal form as

$$A_{sum} = \begin{pmatrix} D_1 & 0 & \cdots \\ 0 & D_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

if the network has more than one component. This implies that the non-zero elements of the sum matrix are confined to square blocks along the diagonal of the matrix, with all other elements being zero.

Changing the multiplicity of the threshold, we can see the variations of the connect components. As the multiplicity c varies from 1 to 0.1, the connect components of higher threshold bond together to form larger components at lower threshold. The components of higher T_c value indicate stronger mutual relations among the members of the components, whereas the components of lower T_c value implies weaker mutual relations among the member of the components.

By varying the threshold, we can draw a graph of sets with inclusion and exclusion of the connect components (sets) at different thresholds. We take the advantages of these set relations to inspect the evolutionary relations among proteins. For each multiplicity threshold, the connect component is drawn as a set enclosing all its members (proteins) and labeled with the multiplicity c of the threshold. The components of lower thresholds may contain the components of higher thresholds, in that the thresholding condition is looser when the threshold is lower. We put the sets of higher thresholds into the sets of lower thresholds if the connect components of the latter contain the connect component of the former. The graph of sets representing the connect components is a fine approach to represent the sequence relations among proteins. The hierarchy of protein evolution can be well delineated by the connectivity of the network.

Results

We use six protein data sets to illustrate the method. In each protein data set, we draw a graph of sets representing the connect components of the protein network at different thresholds. For each connect component, all members in the component are mutually connected at the given threshold. We take advantages of the changes of the connect components on the varying thresholds to demonstrate evolutionary relationships of the proteins. Note that the protein networks are undirected, given the symmetric characterization of the maximum mutual information rates.

Mitochondrial proteins of 28 mammal species

In the first example, we analyze the data set of 28 mitochondrial proteins formerly used by [1, 7, 18, 20]. This dataset consists of 28 proteins encoded by the mitochondrial genome of 28 different mammal species. Each of the 28 protein sequences is concatenated from 10 proteins (COI, COIII, COII, Cyt-b, ND1, ATPase 6, ND4, ND5, ND6, ND2) encoded by the same strand of the mitochondrial genome [1, 7, 18, 20]. Among the 13 protein-coding mitochondrial genes, the 3 shortest genes (ATPase 8, ND3, and ND4L) are excluded, and the 10 proteins (COI, COIII, COII, Cyt-b, ND1, ATPase 6, ND4, ND5, ND6, ND2) are coded by the 10 genes left. The 28 mammal species and their Genbank accession number are namely, the hedgehog (GenBank accession number X88898), mouse (J01420), rat (X14848), cat (U20753), gray seal (X72004), harbor seal (X63726), horse (X79547), donkey (X97337), rhinoceros (X97336), cow (V00654), fin whale (X61145), blue whale (X72204), gibbon (X99256), Sumatran orangutan (X97707), Bornean orangutan (D38115), gorilla (X93347), pygmy chimpanzee (D38116), chimpanzee (D38113), and human (X93334), tiger (EF551003), dog (U96639), wolf (EU442884), black bear (DQ402478), brown bear (AF303110), polar bear (AF303111), opossum (Z29573), wallaroo (Y10524), and platypus (X83427).

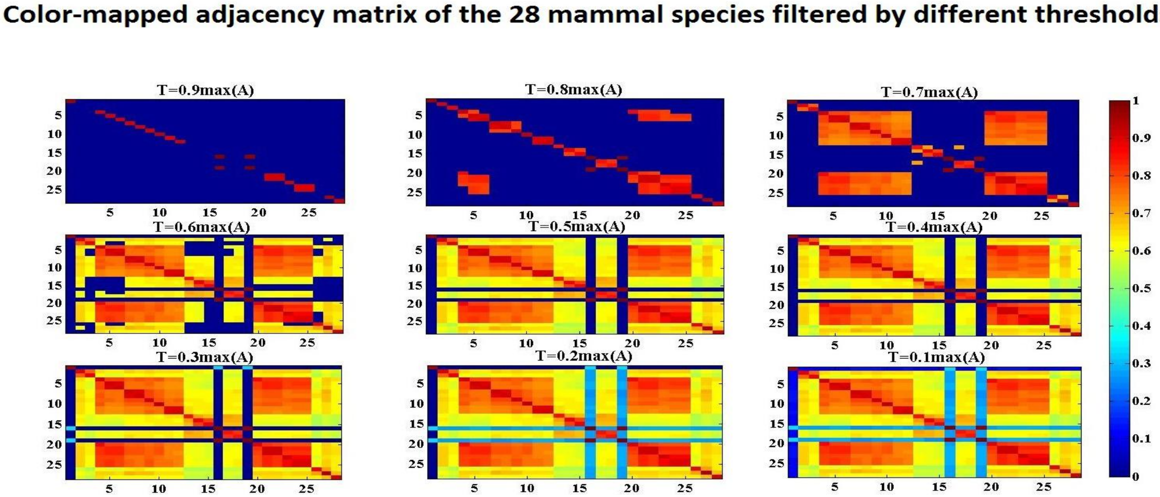

Fig 1. **Color-mapped adjacency matrix of the 28 mammal species filtered by different thresholds. This figure shows the color-map of the filtered adjacency matrices for the 28 mammal species. The multiplicity of the threshold ($T_c = c \cdot A_{max}$) is varied from $c = 0.9$ to $c = 0.1$. The elements of the adjacency matrix below the thresholds are filtered to zero, while the other elements remain unchanged. The adjacency elements are mapped to colors ranging from cold (dark blue, minimum) to warm (dark red, maximum) as shown in the color-bar.**

The color-map of the adjacency matrix filtered by the different thresholds are shown in Fig 1. In this figure, the whole adjacency matrix after filtering is mapped to colors as indicated by the color-bar. The color is ranged from cold (blue) to warm (red), indicating the causality values from minimum to maximum. Two nodes are connected if the corresponding adjacency element is positive, which is indicated by a bright color in the color-map. We can see that higher threshold filters out more connections (less bright areas in the color-maps), which leaves out fewer nodes to be connected. Decrease the threshold, more nodes become connected. The detailed impact of the threshold variation can be seen from the case study of 5 primate species (Gibbon, Sumatran Orangutan, Bornean Orangutan, Pygmy Chimpanzee and Chimpanzee) as shown in Fig 2. In this figure, the clustering of the primate species can be seen by the positiveness of the connectivity values. The Pygmy chimpanzee and the Chimpanzee are first clustered at multiplicity c = 0.84, then the Sumatran Orangutan and the Bornean Orangutan are clustered at the multiplicity of c = 0.82, after which the cluster of the two chimpanzees and the cluster of the two orangutans are grouped together along with the Gibbon at a lower threshold with multiplicity c = 0.72. The connect component is enlarged as the threshold multiplicity decreases.

The classification result is shown in Fig 3. In this figure, the proteins of the 28 mammal species are classified by the contour of connect components at different threshold multiplicity. The higher the multiplicity implies stronger mutual relations among the members of the components. Each member protein of the component is represented by their animal species, e.g. the mitochondrial protein of the hedgehog is represented by the name of hedgehog in the figure. This figure shows that the mammal species of Carnivora are first classified according to their families: Phocidae (Gray seal and Harbor seal, c = 0.9), Canidae (dog and wolf, c = 0.9), Ursidae (brown bear, polar bear, and black bear, c = 0.89), Felidae (cat and tiger, c = 0.88). The Carnivora families (Phocidae, Canidar, Ursidae) are grouped into a larger cluster at c = 0.85, which is later joined by another Carnivora family i.e. the family of Felidae, along with the two families (Equidae: horse and donkey, c = 0.9, and Rhinocerotidae: rhinoceros) of Perissodactyla order and one species (cow) of Artiodactyla order, at the multiplicity of c = 0.82. The Infra-class of Marsupialia (Opossum and Wallaroo, c = 0.73) and the order of Cetacea (Fin whale and Blue whale, c = 0.9), another species of Artiodactyla order (Platypus), and the order of Rodentia (Mouse and Rat, c = 0.79), are added into the original group level by level. As to the primates, the Ponginae subfamily (Sumatran orangutan and Bornean orangutan, c = 0.82), the Homininae subfamily (Pygmy chimpanzee and Chimpanzee, c = 0.84) of the Hominidae family are first grouped with the species (Gibbon) of the Hylobatidae family at c = 0.72, then this group of primates first joins the big group of mixed Carnivora, Perissodactyla, Artiodactyla, Marsupialia, Cetacea, and Rodents species. The network of the proteins of 28 mammal species is entirely formed, once the mixed large group is joined by another small mixed group of two primates: Hominoidea family (Gorilla and Human, *c* = 1) and one Eulipotyphla (Hedgehog) at *c* = 0.3.

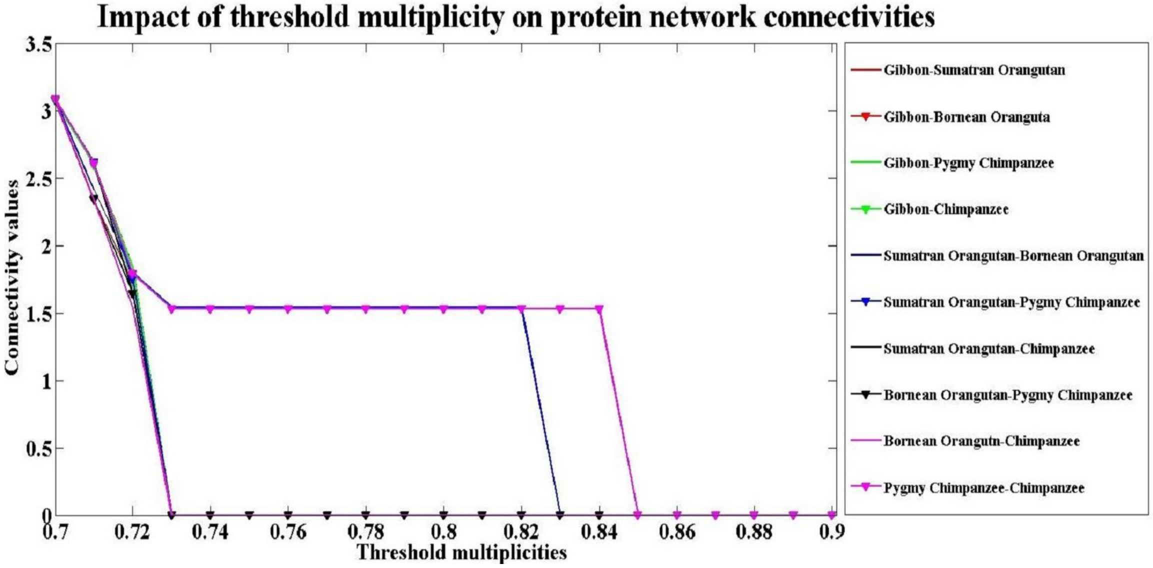

Fig 2. Impact of the threshold multiplicity on protein network connectivity. This figure shows the connectivity values varied against the threshold multiplicities (*c* = 0.7,0.71,··· ,0.9). The connectivity values are the mean value of the roots of the powered adjacency matrix $\overline{a}(i,j)=\frac{\sum_{n=1,2,\ldots,27} a_{n,ij}^{1/n}}{27}$ **, where the nominator is 27 because the entire network is consisted of 28 nodes so the maximum length of a path is 27, to get the connectivity values, we need to account all paths from length 1 to length 27, and get their square root averages from the 27 powered adjacency matrix, $a_{n,ij}$ is the *ij*-th element of the power *n* adjacency matrix A^n. The positiveness of $\overline{a}(i,j)$ indicates the existence of a connection between node i and node j in the protein network. This graph shows the connectivity values among the mitochondrial proteins of Gibbon, Sumatran Orangutan, Bornean Orangutan, Pygmy Chimpanzee and Chimpanzee, over the threshold multiplicity between *c* = 0.7 and 0.9.**

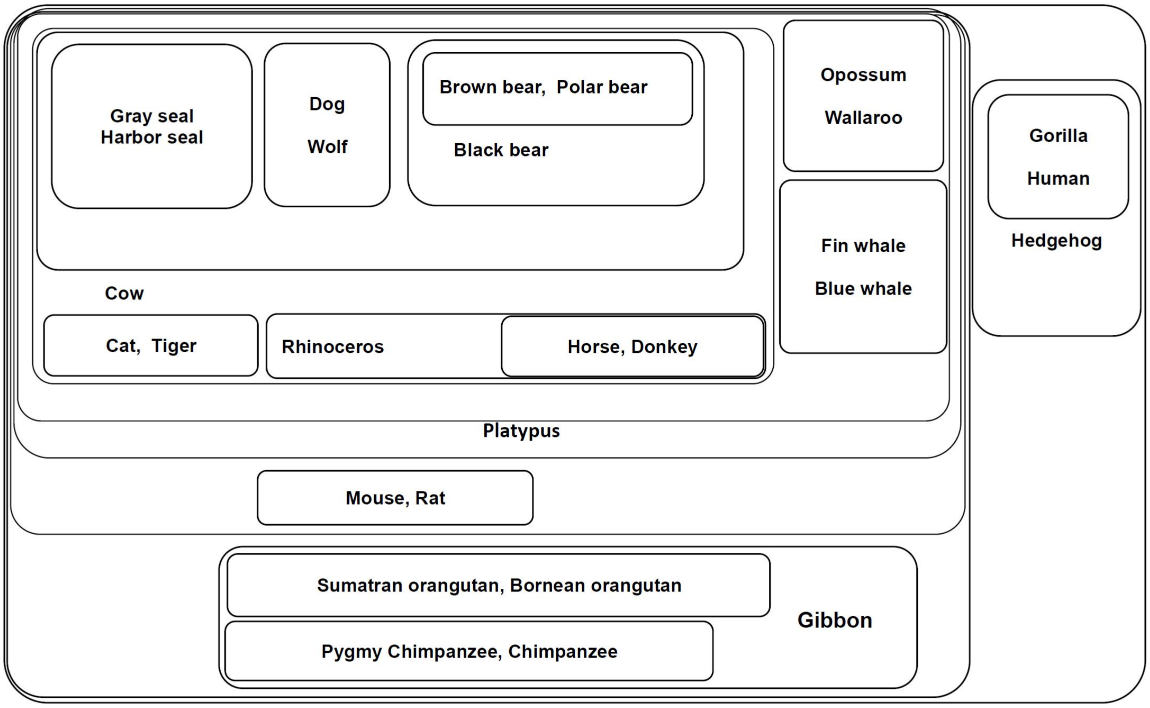

Fig 3. Component graph of mitochondrial proteins of 28 mammal species. This figure shows the graph of connect components of 28 mammal species at different thresholds. Each set represents a connect component whose members are mutually connected at a certain threshold (T_c = c • A_{max} , $c \in [0,1]$). Components of higher thresholds are included by components of lower threshold.

This result sheds light to the global relations among the 28 mammal species. In contrast to the bivariate branching of the polygenetic-trees [1, 7, 12, 18] and [20], the classification is more universal, indicating the parallel mutual connections among the different Carnivora families. This result brings a more natural explanation to the evolution compared to the conventional bivariate branching, because the relations among the different species may not necessarily be pairwise, i.e. it is insufficient to say that one species is close to only one other species in the universe.

Mitochondrial proteins of 35 mammal species

The second data set consists of 35 proteins of NADH dehydrogenase encoded by the mitochondrial genes from 35 different mammal species [23]. GenBank accession numbers of the 35 mammal genes [23] are human (V00662), pygmy chimpanzee (D38116), common chimpanzee (D38113), gorilla (D38114), gibbon (X99256), baboon (Y18001), vervet monkey (AY863426), ape (NC 002764), Bornean orangutan (D38115), Sumatran orangutan (NC 002083), cat (U20753), dog (U96639), pig (AJ002189), sheep

(AF010406), goat (AF533441), cow (V00654), buffalo (AY488491), wolf (EU442884), tiger (EF551003), leopard (EF551002), Indian rhinoceros (X97336), white rhinoceros (Y07726), harbor seal (X63726), gray seal (X72004), African elephant (AJ224821), Asiatic elephant (DQ316068), black bear (DQ402478), brown bear (AF303110), polar bear (AF303111), giant panda (EF212882), rabbit (AJ001588), hedgehog (X88898), Norway rat (X14848), vole (AF348082), squirrel (AJ238588).

The evolutionary relationship of the species are shown in Fig 4. In this figure, the species of different families are clearly classified into separate groups according to their mammal orders (Carnivora, Artiodactyla, Perissodactyla, Lagomorpha, Rodentia, Proboscidea, Primate, and Eulipotyphla). It is identified in the figure that the Carnivora is the core of the networks, which are closely surrounded by the species of the Perissodactyla order and the Artiodactyla order, the components of the three orders are then closely connected to the species of Lagomorpha order and the Proboscidea order. The Primate species are in the next order level that are fully connected to the core, which is followed by the Rodentia order and the Eulipotyphla order. This hierarchical relations are similar to those found by moment vectors on mitochondrial genes [23], except for the difference that the cousins or non-brother peers in moment vector analysis now become mutually related at certain levels by the multivariate nature of the new method.

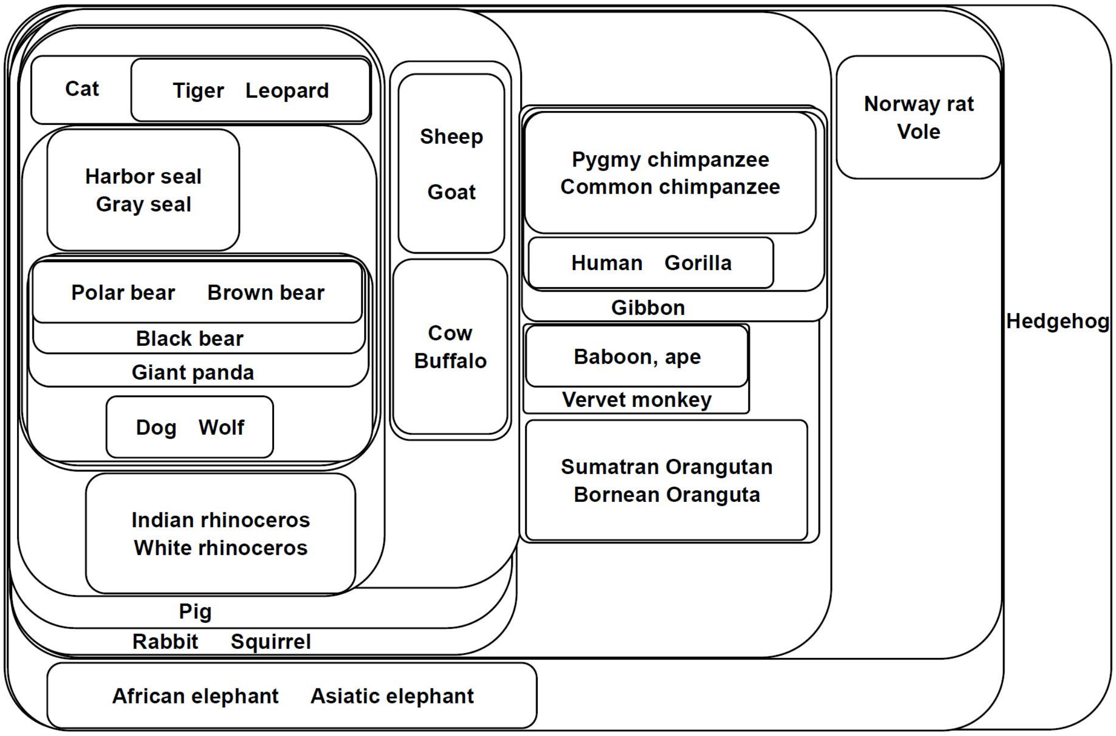

Fig 4. Component graph of mitochondrial proteins of 35 mammal species. This figure shows the graph of connect components of 35 mammal species at different thresholds. Each set represents a connect component whose members are mutually connected at a certain threshold ($T_c = c \bullet A_{max}$, $c \in [0,1]$). Components of higher thresholds are included by components of lower threshold.

Inside each different order, the mammal species are also well-classified. The Carnivora is classified into families of Phocidae (Gray seal and Harbor seal, c = 0.98), Ursidae (brown bear, polar bear, black bear and Giant panda c = 0.94), Canidae (dog and wolf, c = 1), and Felidae (cat and tiger, leopard, c = 0.94). The only one family Rhinocerotidae (Indian rhinoceros and White rhinoceros, c = 0.93) of Perissodactyla order is inter-connected to the families of Carnivora. The Artiodactyla is divided into two groups of the same family (Bovidae) but different subfamilies: Bovinae (Cow and Buffalo, c = 0.97) and Caprinae (Sheep and Goat, c = 0.96), along with one species of the Suidae family (Pig). Aside from the other two Rodentia families: Muridae (Norway rat) and Cricetidae (Vole), the one species of Sciuridae family (squirrel) affiliated to the Rodentia order is close to the Carnivora, Perissodactyla, Artiodactyla, Lagomorpha (Leporidae family: rabbit), and Proboscidea (Elephantidae family: African elephant and Asiatic elephant, c = 0.98). The Primate is also classified into different families: Hominidae (Ponginae: Sumatran orangutan and Bornean orangutan, c = 0.8), and Homininae (Pygmy chimpanzee, Common chimpanzee, Gorilla and Human, c = 0.85), Hylobatidae (Gibbon), the Cercopithecidae (baboon, Vervet monkey) and Hominidae (ape). The species of the Cercopithecidae and Hominidae families are interconnected. The hedgehog of the Erinaceidae family of the Eulipotyphla order is the furthest species to the others in the protein network.

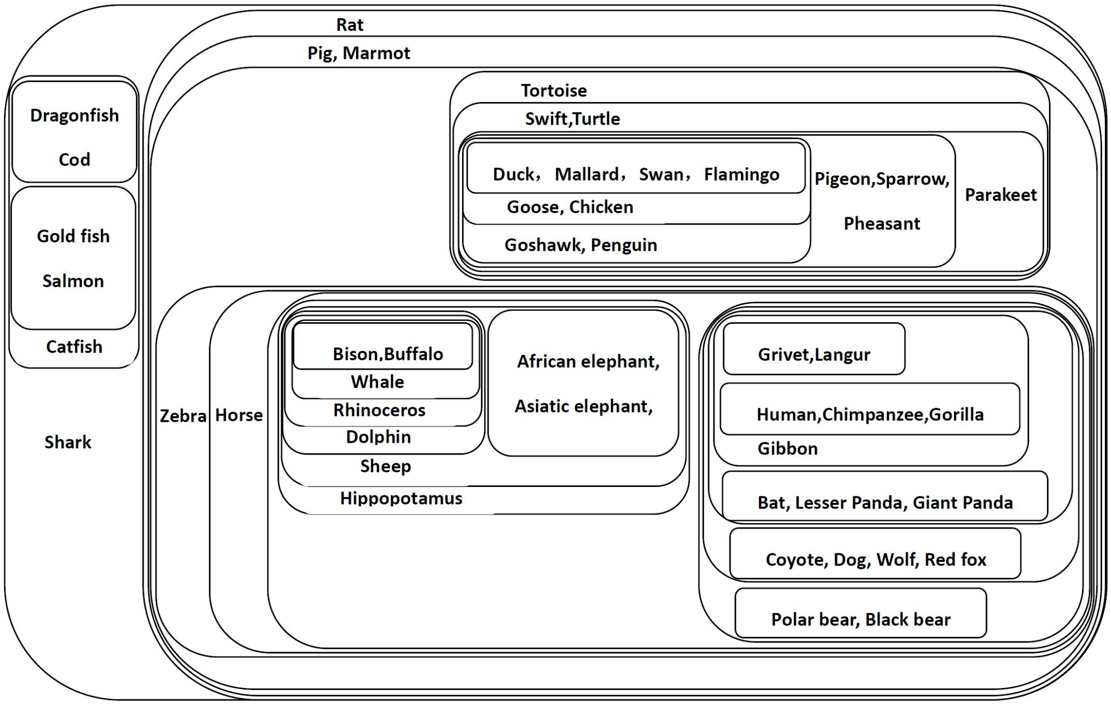

Fig 5. Component graph of beta globins of 50 animal species. This figure shows the connect components of 50 animal species at different thresholds. Each set represents a connect component whose members are mutually connected at a certain threshold ($T_c = c \bullet A_{max}$, $c \in [0,1]$). Components of higher thresholds are included by components of lower threshold.

Beta-globin of 50 animal species

The third data set is the set of 50 beta-globins from 50 different animal species. The data was originally used in [19, 22]. The animal species and protein accession number of the 50 beta-globins are Human (AAA16334.1), Pigeon (P11342.1), Goshawk (P08851.1), Black bear (P68012.1)), Lesser panda (P18982.1), Asiatic elephant (P02084.1), Giant panda (P18983.2), African elephant (P02085.1), Sheep (P02075.2), Tortoise (P83123.3), Duck (P02114.2), Grivet (P02028.1), Mallard (P02115.1), Gorilla (P02024.2), Goose (P02117.1), Shark (P02143.1), Rat (CAA33114.1), Hippopotamus (P19016.1), Penguin (P80216.1), Horse (P02062.1), Swift (P15165.1), Gibbon (P02025.1), Coyote (P60525.1), Whale (P18984.1), Catfish (O13163.2), Bat (P24660.1), Bison (P09422.1), Red fox (P21201.1), Swan (P68945.1), Marmot (P08853.1), Buffalo (P67820.1), Salmon (Q91473.3), Dog (P60524.1), Sparrow (P07406.1), Chimpanzee (P68873.2), Pheasant (P02113.1), Dolphin (P18990.1), Flamingo (P02121.1), Goldfish (P02140.1), Pig (P02067.3), Polar bear (P68011.1), Dragonfish (ADD73488.1), Rhinoceros (P09907.1), Parakeet (P21668.1), Chicken (P02112.2), Zebra (P67824.1), Wolf (P60526.1), Cod (O13077.2), Turtle (P13274.1), Langur (P02032.1).

The classification results using our new method are given in Fig 5. In this figure, we can see that the fish (Actinopterygii), avian, mammal and reptile are reasonably classified. The four big clusters are similar to those branches identified in the polygenetic tree by K-string dictionary method [22], with a few exceptions in the species sub-classes. The distinction between our results and the results of the other methods, is mainly due to the differences of the method orientations, in which our method pay more attention to the global connectivity rather than bivariate branching of the proteins.

In this result, the aves, reptiles, mammal, fish (Actinopterygii) are well-classified. All Aves species are clustered together, with Anatidae family (Duck, Mallard, Swan, c = 1) of the Anseriformes order as the core, all the rest aves species are enclosed around. The species of mammal class are categorized into clusters of different animal orders (Primate, Rodentia, Cetacea, Carnivora, Artiodactyla, Perissodactyla, Proboscidea, Chiroptera), where Primate species (Cercopithecidae family: Grivet, Langur, the Hominidae family: Human, Chimpanzee, Gorrila, and the Hylobatidae family: Gibbon) are the closest species to the Carnivora species (Ursidae family: Lesser panda, Giant panda, Canidae family: Coyote, Dog, Wolf, Red fox). The Bat of the Chiroptera Order is mixed with the Primates and the Carnivoras. In another part of the mammals, the Artiodactyla order (Bovidae family: Bison, Buffalo, Sheep, Hippopotamidae family: Hippopotamus), the Cetacea order (Whale, Dolphin), one species of the Perissodactyla order (Hinocerotidae family: Rhinoceros), and the Proboscidea order are closely connected. On a weaker threshold level (c = 0.58), the main mammal orders: Primates, Carnivoras, Artiodactylas, Cetaceas, and Proboscideas, as well as Perissodactyla order (Hinocerotidae family: Rhinoceros, Equidae family: Horse, Zebra) are all mutually connected. At a lower threshold (c = 0.5), the class of Aves, Mammals, Reptilia (Turtle, Tortoise), and a new joined species in the Artiodactyla order (Suidae family: Pig), as well as two other mammal species in the Rodentia order (Marmot, Rat), are all mutually connected. The fishes: Actinopterygii class (Dragonfish, Cod, Goldfish, Salmon, and the Catfish) and the Chondrichthyes class (Shark) are the last components joining the whole network, where the Chondrichthyes class (Shark) is the farthest class to all other animal species.

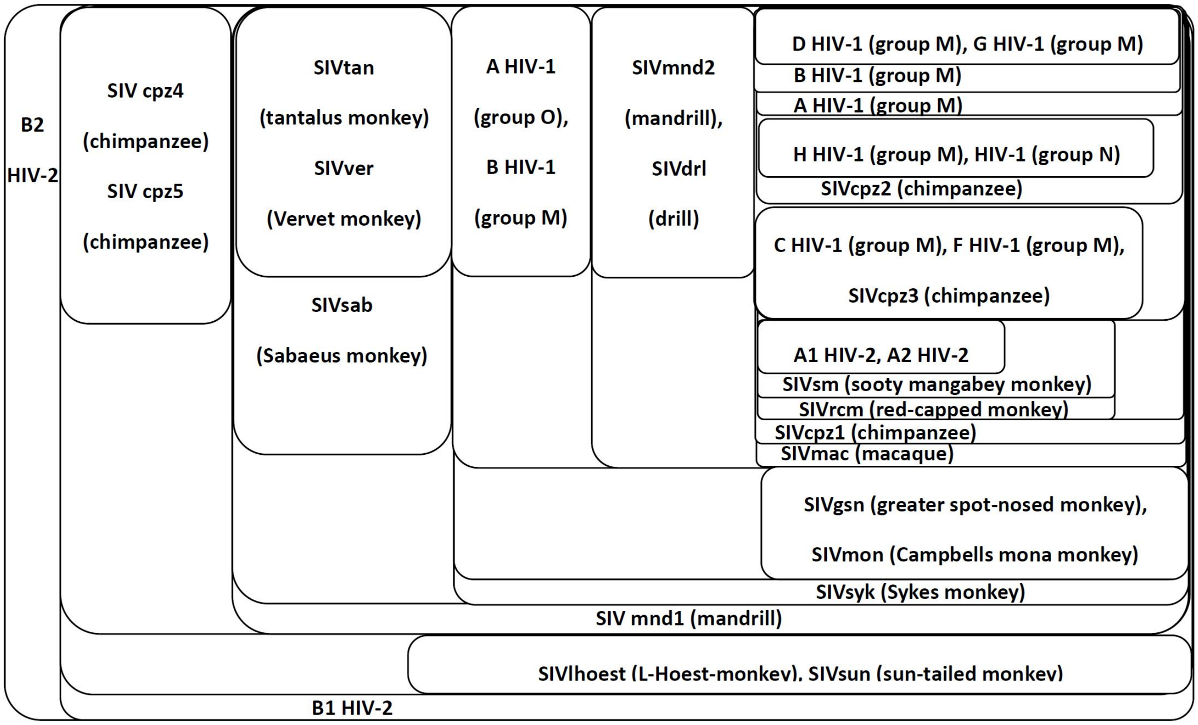

Fig 6. Component graph of proteins encoded by HIV virus. This figure shows the connect components of HIV virus proteins at different thresholds. Each set represents a connect component whose members are mutually connected at a certain threshold ($T_c = c \bullet A_{max}$, $c \in [0,1]$). Components of higher thresholds are included by components of lower threshold.

HIV proteins

Human immunodeficiency virus (HIV) is a lenti virus that can lead to acquired immune deficiency syndrome (AIDS [13, 23]). To develop the anti-HIV drugs and vaccines, the research into the origins and evolution of this virus becomes very important. Rambaut et. al. [13] used maximum likelihood method to reconstruct the phylogenetic tree of the primate lenti viruses including HIV-1, HIV-2, and the simian immunodeficiency viruses (SIVs). It discovers that the two HIV viruses are related to different SIVs and therefore have different evolutionary origins. Here, we used the same dataset as they used to examine the global connections among proteins. The dataset consists of 33 protein sequences encoded by the DNA sequences of the 33 HIV and SIV viruses. The RNA genomes are transformed into DNA sequences (change U by T) before downloaded from the GenBank. The subtypes of HIV-1, HIV-2 and SIV viruses [13, 23] and their primate hosts and GenBank accession numbers are listed as follows: HIV-1, group M: A (AF004885); B (A04321); C (AF443079); D (K03454); F (AY173957); G (AY772535); H (AF190127); group N (DQ017382); group O: A (AY169802); B (AY169803); HIV-2: A1

(AF082339); A2 (M30502); B1 (L07625); B2 (X61240); SIV chimpanzee (Pan troglodytes troglodytes): SIVcpz1 (AY169968), SIVcpz2 (AJ271369), SIVcpz3 (DQ373063); SIV chimpanzee (Pan troglodytes schweinfurthii): SIVcpz4 (DQ374657), SIVcpz5 (DQ374658); SIVdrl, drill (AY159321); SIVgsn, greater spot-nosed monkey (AF468659); SIVlhoest, L' Hoest monkey (AF188114); SIVmac, macaque (D01065); SIVmnd1, mandrill (M27470), SIVmnd2, mandrill (AY159322); SIVmon, Campbells mona monkey (AY340701); SIVrcm, red-capped monkey (AF382829); SIVsab, Sabaeus monkey (U04005); SIVsm, sooty mangabey monkey (U72748); SIVsun, sun-tailed monkey (AF131870); SIVsyk, Sykes' monkey (L06042); SIVtan, tantalus monkey (U58991); SIVver, vervet monkey (M29975).

The evolutionary relationship interpreted by our method is shown in Fig 6. The classifications of our analysis are quite different from those found by moment vectors [23], but are similar to those found by the maximum likelihood method [13]. In our analysis, the different types of the HIV-1 and HIV-2 proteins have different lineages to the SIV of the primates. From the global connections, the group M proteins of HIV-1 virus are closest to the SIV proteins of chimpanzee (Pan troglodytes troglodytes). The HIV-2 A and HIV-2 B are separate. The HIV-2 A proteins are most closely related to the SIVsm (sooty mangabey monkey), SIVrcm (red-capped monkey), SIV chimpanzee (Pan troglodytes troglodytes) and the SIVmac (macaque), whereas HIV-2 B is closer to SIV chimpanzee (Pan troglodytes schweinfurthii), and SIVlhoest (L' Hoest monkey) and SIVsun (sun-tailed monkey). Separate clusters of HIV-2 A and of the group M of HIV-1, each along with some SIVs, are first enclosed into a larger connect component, then the other SIVs are connected to the joined group of the HIV-1 group M and HIV-2 A, when the threshold multiplicity decreases. The proteins of HIV-2 B is farthest to the proteins of HIV-1 and HIV-2 A, where HIV-2 B joins the HIV-1 and HIV-2 A at the lowest threshold.

Influenza A virus

Influenza A virus is a kind of negative-sense, single-stranded, segmented RNA viruses. Here, we use the dataset of 52 proteins encoded by the genes of 52 different influenza A virus [15]. These proteins are characterized by three factors: the virus subtypes, the geographical location of the occurrence and the host of the influenza A virus. The virus subtypes are labeled by the combination of an H number for the type of hemagglutinin and an N number for the type of neuraminidase. Our dataset is made up of six virus subtypes: H7N3, H11N9, H1N1, H7N9, H3N2, H5N1 [15].

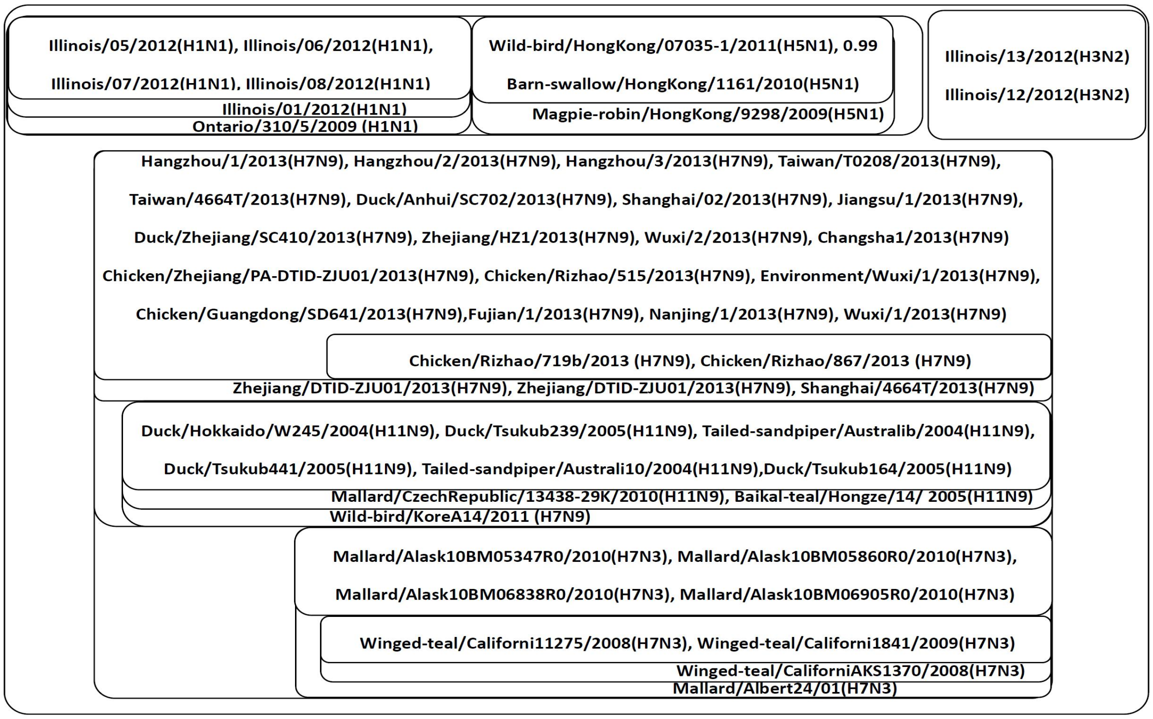
Illinois/05/2012(H1N1), Illinois/06/2012(H1N1),
Illinois/07/2012(H1N1), Illinois/08/2012(H1N1)
Illinois/01/2012(H1N1)
Ontario/310/5/2009 (H1N1)
Wild-bird/HongKong/07035-1/2011(H5N1), 0.99
Barn-swallow/HongKong/1161/2010(H5N1)
Magpie-robin/HongKong/9298/2009(H5N1)
Illinois/13/2012(H3N2)
Illinois/12/2012(H3N2)
Hangzhou/1/2013(H7N9), Hangzhou/2/2013(H7N9), Hangzhou/3/2013(H7N9), Taiwan/T0208/2013(H7N9),
Taiwan/4664T/2013(H7N9), Duck/Anhui/SC702/2013(H7N9), Shanghai/02/2013(H7N9), Jiangsu/1/2013(H7N9),
Duck/Zhejiang/SC410/2013(H7N9), Zhejiang/HZ1/2013(H7N9), Wuxi/2/2013(H7N9), Changsha1/2013(H7N9)
Chicken/Zhejiang/PA-DTID-ZJU01/2013(H7N9), Chicken/Rizhao/515/2013(H7N9), Environment/Wuxi/1/2013(H7N9),
Chicken/Guangdong/SD641/2013(H7N9),Fujian/1/2013(H7N9), Nanjing/1/2013(H7N9), Wuxi/1/2013(H7N9)
Chicken/Rizhao/719b/2013 (H7N9), Chicken/Rizhao/867/2013 (H7N9)
Zhejiang/DTID-ZJU01/2013(H7N9), Zhejiang/DTID-ZJU01/2013(H7N9), Shanghai/4664T/2013(H7N9)
Duck/Hokkaido/W245/2004(H11N9), Duck/Tsukub239/2005(H11N9), Tailed-sandpiper/Australib/2004(H11N9),
Duck/Tsukub441/2005(H11N9), Tailed-sandpiper/Australi10/2004(H11N9),Duck/Tsukub164/2005(H11N9)
Mallard/CzechRepublic/13438-29K/2010(H11N9), Baikal-teal/Hongze/14/ 2005(H11N9)
Wild-bird/KoreA14/2011 (H7N9)
Mallard/Alask10BM05347R0/2010(H7N3), Mallard/Alask10BM05860R0/2010(H7N3),
Mallard/Alask10BM06838R0/2010(H7N3), Mallard/Alask10BM06905R0/2010(H7N3)
Winged-teal/Californi11275/2008(H7N3), Winged-teal/Californi1841/2009(H7N3)
Winged-teal/CaliforniAKS1370/2008(H7N3)
Mallard/Albert24/01(H7N3)

Fig 7. Component graph of proteins encoded by Influenza A virus. This figure shows the connect components at different thresholds for proteins encoded by Influenza A virus genes. Each set represents a connect component whose members are mutually connected at a certain threshold (T_c = c • A_{max} , $c \in [0,1]$). Components of higher thresholds are included by components of lower threshold.

The classification of influenza A virus is shown in Fig 7. In this figure, the influenza A viruse subtypes are well-classified, and within each subtype, the proteins are classified in terms of their host and geographic locations. For instance, the proteins of the H7N3 virus with the host of Mallard in Alaska are grouped together, and are separated to the group of H7N3 proteins with the host of winged-teal in California.

Within each subtype of the influenza A viruses, the proteins are grouped first according to the N number for the type of neuraminidase, and then the H number of the type of hemagglutinin. The proteins of H1N1 virus are closely grouped with the proteins of H5N1 virus, while the proteins of H7N9 virus are closely classified with the proteins of H11N9 virus. The evolutionary hierarchy of Influenza A virus is clearly shown in this classification. The connection core is formed by the proteins of H7N9 and H11N9, which is joined by the proteins of H7N3. The enlarged core is finally joined by the union of H3N2 proteins and the union group of H1N1 and H5N1.

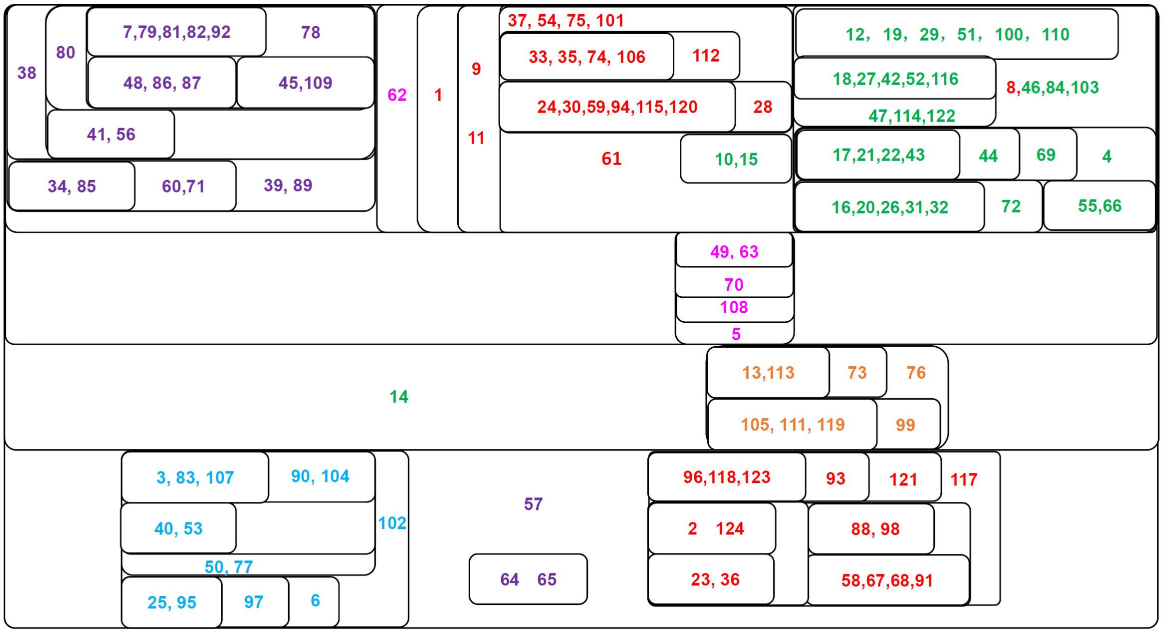

Fig 8. Component graph of protein Kinase C families. This figure shows the connect components of the PKC (protein kinase C) at different thresholds. Each set represents a connect component whose members are mutually connected at certain threshold ($T_c = c \bullet A_{max}$, $c \in [0,1]$). Components of higher thresholds are included by components of lower threshold. Each member of the 124 proteins is correspond to a unique index number between 1 and 124 as referenced in [21]. For presentation convenience, we only labeled the 124 unique index number to represent the 124 proteins. Proteins of different PKC subfamilies are labeled by different colors: aPKC (blue), cPKC (green), nPKC (red), PKC1 (purple), PRK (pink), PKCmu (orange). Description of the 124 PKCs can be found in supplementary materials of [21].

Protein kinase C

In the sixth example, we analyzed the protein kinase C families. Protein kinase C, in abbreviation the PKC, is a family of enzymes involved in controlling the function of other proteins through the phosphorylation of hydroxyl groups of serine and threonine amino acid residues on these proteins [21]. The entire PKC family can be divided into six subfamilies: cPKC, nPKC, aPKC, PKCμ, PKC1 and PRK. There are 124 protein sequences in total. The classification results of the PKC families are shown in Fig 8. In this figure, the six subfamilies of PKC are clearly clustered into separate groups: PKC1 (upper left block, purple), nPKC (Upper middle and the bottum right blocks, red), cPKC (upper right block, green), PRK (the center block, pink), PKCmu (the below center block, orange), aPKC (bottum left block, blue). All elements in each block are from the same PKC subfamilies.

Inside each block, the protein members are classified according to their NCBI descriptions. The nPKCs are divided into η (33, 35, 74, 106, 112), (24, 30, 59, 94, 115, 120, 28), δ (1, 2, 124, 23, 36, 88, 98, 58, 67, 68, 91), Serine\Threonine (9, 11), θ (96, 118, 123, 93, 121). The cPKCs are divided into subgroups of the γ (18, 27, 42, 52, 116, 47, 114, 122), α (16, 20, 26, 31, 32, 72), β (17, 21, 22, 43, 44, 69). The aPKCs are classified into ι (3, 83, 107, 90, 104, 40, 53) and ζ (25, 95, 97, 6). The PKCmu proteins are divided into nuPKCmu (13, 113, 73) and muPKCmu (105, 111, 119, 99). The classification results are similar to those found by natural vectors [21].

Clear hierarchy of the PKC families can be seen from this figure. The PKC1 (purple), cPKC (green) and the η, ϵ and Serine\Threonine sub-classes of the nPKC (red) are all parallel connected. The PRK (pink) is connected to the above at a lower threshold, which is followed by PKCmu (orange) and aPKC (blue). The δ and θ sub-classes of nPKC (red) is the farthest group to the core.

Discussion

Protein universe is a complex system can be modeled as an undirected network with evolutionary relations as interactions. In this paper, we described a global connectivity method to identify multivariate evolutionary relationship among proteins. This method bases on information and network theories is powerful. It takes advantages of the distribution of amino acids and use maximum mutual information rates to detect alignment-free mutual relationships among proteins. In analysis, protein universe is modeled as a protein network, where protein sequences as nodes and their relations as links, the evolutionary relationships of the proteins are identified by connect components of the network.

The key point and innovation of our method is that it considers the protein evolutionary relations as multivariate. Each taxon may have more than one sisters or brothers, i.e. their parents may have one, two or more than two children. Traditional protein classification methods inspect the protein relations pairwise, which limited the protein classification in a bivariate view. In contrast, our method examine the global relationships of proteins, the lineage of one species may be inherited by more than one sub-lineages. Our method explain reasonable multivariate evolutionary relationships among proteins of existing datasets, even though the true evolutionary hierarchy of some of the species are still controversial indeed. Compared to earlier polygenetic-tree representations, this method introduces brand-new ideas in protein evolutionary classification.

The classification process relies on the division of connect components and the variation of adjacency thresholds. The threshold of adjacency matrix acts as the cut-off to protein network connections. It cuts weak connections below the threshold, while keeping strong connections equal or above the threshold. By varying the threshold, one is able to classify proteins by examining the inclusion\exclusion of connect components.

Results of the mitochondrial proteins show that the animal species are classified first by their biological families, then theirs orders, and classes. The connection strength decreases as their biological similarity decreases. Animal species of the same family are strongly related, the relations or connections are weakened when their families in the same order differ, and are again weakened if their biological orders differ. Sometimes, the species in the same class, are cross related with different orders or families, which may be because their species lie on the same level of evolution. For instance, the species of Artiodactyla and Perissodactyla are not only intra-connected, but also inter-connected, and they are more closely related to Carnivora rather than Primates. Primates are comparatively far away than the other mammals, reflected by their weaker connections to the rest mammal orders. Rodentia is also a bit far away to other mammals, whose inter-connections to the other mammal species are weaker. Hedgehog is found close to the Hominoidea family (Gorilla and Human) of Primate in the 28 mammal analysis, which shows a consistency to [8] for the convergence of Hedgehog to Primates. Analysis of beta-globins shows that the mammals are firstly connected to Aves, then Reptilia, and finally the fishes (Actinopterygii class and the Chondrichthyes class), the fish is the farthest class to the mammals particularly the Primates.

We also found that the HIV-1 and HIV-2 proteins have close-connections to different SIVs. The HIV-1 group M is comparatively closer to HIV-2 A, and far away to HIV-2 B, where the HIV-1 group M is closest to the SIV chimpanzee (Pan troglodytes troglodytes), while HIV-2 A is also closer to the SIV chimpanzee (Pan troglodytes troglodytes) along with some other SIVs. The HIV-2 B is farthest to HIV-1 and HIV-2 A, but it is closer to the SIV chimpanzee (Pan troglodytes schweinfurthii) and some other SIVs. The results of Influenza A virus indicate that variations of the Influenza A virus are first gathered according to their neuraminidase types i.e. the N number, and then their hemagglutinin types, i.e. the H numbers. The classification results of Influenza A virus are much better by using our new method than by using Yau-Hausdorff distance [15]. For the same dataset, Yau-Hausdorff distance doesn't give a clear classification for the Influenza A virus. Our results of the PKC families are similar to those found by natural vectors [21], but we demonstrate more on the universal relationships of the six PKC subfamilies.

By taking advantages of connect components, our global connectivity method provides a universal view on the multivariate-connections among proteins. The new method is alignment-free, because mutual information rate only depends on the probability distribution of amino acids. The new method can also be used on protein 3D structure, which is done simply by replacing the discrete map of amino acid to real valued coordinates. However, we do not analyze protein 3D structures, because the datasets are only sequences. If we have 3D structure of proteins, the classification results may be improved. Our new method has advantage over other methods, because traditional methods such as K-string dictionary [22], protein map [20] and the natural vectors [21, 24, 25] can apply only to sequences or only to 3D structures (Yau-Hausdorff distance [15]), none of them can apply to both.

Conclusion

We have described a new method on protein classification. This new method innovate multivariate evolutionary relationships among proteins. In contrast to conventional methods, our new method is able to infer multivariate relationships among proteins, and is alignment-free that purely depend on probability distribution of amino acids. The new method can have wide-applications that it can be used to analyze both amino acid sequences and their 3D structures. This is an advantage of our method over traditional approaches, where old methods such as K-string dictionary, protein map, natural vector can only analyze on sequence rather than structure, and Yau-Hausdorff distance can only analyze 3D structures rather than sequences. The new method can help improve our understanding on complexity of protein universe from global connectivity prospective and is an efficient tool for future protein classification analysis.

Conflict of interest

The authors declare no conflict of interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Ethics Statement

N/A.

Acknowledgments

This work is supported by the National Natural Sciences Foundation of China (31271408), Tsinghua University startup fund and Tsinghua University independent research project grant. We further acknowledge Prof. Changchuan Yin for his valuable comments to this paper, Prof. Hing Sun Luk for the help on the languages, and Steve Yau for his kind assistance of computer service, and Department of Mathematics at Tsinghua University for providing the work space and library facilities.

References

1. Cao Y, et al. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. Journal of Molecular Evolution. 1998; 47, 3, 307-322.

2. Cover T, Thomas J. Elements of Information Theory. New York: John Wiley and Sons. 1991.

3. Gao F, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature. 1999; 397, 43641.

4. Gelfand I, Yaglom A. Calculation of amount of information about a random function contained in another such function. American Mathematical Society Translation Series. 1959; 2, 3-52.

5. Hashimoto T, Hasegawa M. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α\Tu and 2\G. Advances in Biophysics. 1996; 32, 96, 73-120.

6. Hlavackova-Schindler K, Palus M, Vejmelka M, Bhattacharya J. Causality detection based on information-theoretic approached in time series analysis. Physics Reports. 2007; 441, 1-46.

7. Janke A, Xu X, Arnason U. The complete mitochondrial genome of the wallaroo (Macropus robustus) and the phylogenetic relationship among Monotremata, Marsupialia and Eutheria. Proceedings of the National Academy of Sciencesm U.S.A. 1997; 94, 12761281.

8. Lawn RM, Schwartz K, Patthy L. Convergent evolution of apolipoprotein(a) in primatesandhedgehog. Proceedings of the National Academy of Sciences. 1997; 94, 22, 11992-7.

9. Levitt M. Nature of the protein universe. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106, 27, 11079-84.

10. Luo H. Evolutionary origin of a streamlined marine bacterioplankton lineage. The ISME Journal. 2014; 1-11.

11. Newman MEJ. Networks an introduction. Oxford University Press. 2010.

12. Raina SZ, Faith JJ, Disotell TR, Seligmann H, Stewart CB, Pollock DD. Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Research. 2005; 15, 66573.

13. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nature Review Genetics. 2004; 5, 5261.

14. Razak* FA, Wan* X, Jensen HJ. Information theoretic measures of causality: Music performance as a case study. Edward Elgar Handbook on Complexity Science Methods. In Press. 2016.

15. Tian K, Yang X, Kong Q, Yin C, He RL, Yau SS-T. Two Dimensional YauHausdorff Distance with Applications on Comparison of DNA and Protein Sequences. PLoS ONE. 2015; 10 (9): e0136577. doi:10.1371/journal.pone.0136577

16. Vejmelka M, Palus M. Inferring the directionality of coupling with conditional mutual information. Physical Review E. 2008; 77, 026214.

17. Wan X, Cruts B, Jensen HJ. The Causal Inference of Cortical Neural Networks during Music Improvisations. PLoS ONE.2014;9, 12, e112776. doi: 10.1371/ journal. pone. 0112776.

18. Xia X, Li W. What Amino Acid Properties Affect Protein Evolution? Journal of Molecular Evolution. 1998; 47: 557564.

19. Yau SS-T, Yu C, He RL. A protein map and its application. DNA and Cell Biology. 2008; 27, 241250.

20. Yu C, Cheng S, He RL, Yau SS-T. Protein map: An alignment-free sequence comparison method based on various properties of amino acids. Gene. 2011; 486, 110-118.

21. Yu C, Deng M, Cheng S, He RL, Yau SS-T. Protein space: A natural method for realizing the nature of protein universe. Journal of Theoretical Biology. 2013; 318,197-204.

22. Yu C, He RL, Yau SS-T. Protein sequence comparison based on K-string dictionary. Gene. 2013; 529, 250-256.

23. Yu C, Liang Q, Yin C, He RL, Yau SS-T. A Novel Construction of Genome Space with Biological Geometry. DNA Research. 2010; 17, 155168, doi:10.1093/dnares/dsq008.

24. Zhao B, He RL, Yau SS-T. A new distribution vector and its application in genome clustering. Molecular Phylogenetics and Evolution. 2011; 59, 438443.

25. Zhao X, Wan X, He RL, Yau SS-T. A new method for studying the evolutionary origin of the SAR11 clade marine bacteria. Molecular Phylogenetics and Evolution. 2016; 98, 271-279.

26. Zhou Y. The basics of information theory, 3rd Edition. Beijing University of Aeronautics and Astronautics Press. 2006.