Displaying Phylogenetic Tree with 3D Clustering

1. Introduction

By using DACIDR [1], each sequence is represented as a point in the target dimension space, i.e. the 3D space. Also, by using RAxML, all the sequences are represented as leaf nodes in the phylogenetic tree. Therefore each leaf node in the phylogenetic tree corresponds to a point in the 3D dimension reduction result. To display the clustering result with phylogenetic trees, one way to do that is to display the clusters of sequences on one side, then the phylogenetic tree on the other side. This can be generated using a cladogram, where each point from the tree node is projected into a corresponding point in the clustering result.

1. Cuboid Cladogram

As the clustering result is generated in a 3D space, one intuitive way to view the clustering result and the tree together is to display the clustering result on one side, while the tree is displayed on the other side. This method is solely used to verify the result between the clustering and phylogenetic analysis generated in a separate method using the same sequence set. The cladogram from a purely tree diagram can be project to the clustering result so that each point in the target dimension space correspond to a leaf node in the phylogenetic tree. The clustering result is also referred to as the MDS dimension reduction result (MDS Clustering). Instead of building tree from top down (like what most algorithm would do), the tree can be built from the MDS clustering result, i.e. from down to top. Once the phylogenetic tree is generated from a traditional method, the tree can be projected onto the clustering result.

To find the suitable plane for the phylogenetic tree to be projected to, one can either randomly select a plane or use the best possible plane. The random selection may have a problem that the tree is projected to a plane that points are not clearly separated. Figure 1 shows an example of a phylogenetic tree with 8 sequences and the same 8 sequences with clustering result. Assume that this tree is constructed using a traditional method, and if one needs to display the tree simultaneously with the clustering, one needs to choose which dimension it wants to be projected to. In current example, as the both the clustering and the trees are in 2D, one needs to project the tree into a one dimension plane (a line) within the 2D clustering where the points are lied on. A random choice could end up choosing a tree shown in Figure 2. It shows that the tree is not well projected, and the lines were overlapping with each other so that it hard to observe the connections between the leaf points.

Figure The left hand side of the graph representation is a cubic cladogram displayed with 8 sequences. The right hand side of the graph is the same 8 sequences visualized in 2D after dimension reduction.



Figure The example of choosing a random line to project all the sequences to and draw the given cubic cladogram accordingly.

And in practice, an obviously good choice of the line would be the same line as shown in Figure 3. This is because the points that projected on this line are spread out instead of crowded together as shown in Figure 2. One way to select a best possible plane is to use the principal component analysis (PCA) on that. The PCA is a popular method that used for dimensionality reduction based on the vectors in the original dimensionality, which is 2 in this example. It can transform the data to a new coordinate system such that the greatest variance comes to a certain coordinate (usually the first one). Suppose the target dimension for the clustering result is in L-dimension and there are number of N points, the MDS clustering coordinates can be represented in an N×L matrix, denoted as X. The general solution for PCA involves singular vector decomposition on the original matrix X. In in the example given as in Figure 1, x is an 8×2 matrix. So there are 2 dimensions generated by PCA that gives the largest variance and the smallest variance of the coordinates on that dimension. The Figure 4 actually shows the line which gives the largest variances. Compared to the random choice shown in Figure 2, the projected tree is much clearer, and the correlations between the clustering result and the phylogenetic tree can be intuitively observed.



Figure An example of a good choice of projection line as the dotted line within 8 sequences visualized in 2D space.



Figure The example of choosing a good projection line determined by PCA to project all the sequences to and draw the given cubic cladogram accordingly.

As shown in the 2D plot, the branches sometimes can overlap each other if a projection from clustering to a certain line is made. This could cause in-efficient display if some dimension similar to Figure 2 is chosen. Additionally, one cannot guarantee the clean coordinates system such as in Figure 4 exists in every clustering result and phylogenetic tree result. In this particular example, all the sequences have the same order as from clustering result to phylogenetic tree. If they have different order, the branches from the projected tree may be overlapping with each other, e.g. if sequence A and sequence E swap location. So to avoid that, one more dimension can be added to this method. If the clustering result is shown in 3D and the tree is projected onto a 2D plane, the overlapping of the branches won’t be an issue since the branches are in 1D. The example graph is shown in Figure 5. This graph is referred to as cuboid cladogram. This is because the phylogenetic tree displayed in the graph won’t show difference lengths between the leaf nodes and their parents. The phylogenetic tree here could be a rooted tree and an outgroup can be added as well. Once the plane with the largest variance is found for the 3D coordinates, the points are projected onto that plane. Then each pair of points, which correspond to each pair of leaf nodes shared the same parent are selected in the plane, their parent will be a new point which is the exact middle point of the connection between these two points. The parent will be draw one level higher than its children. The process is recursively done until all points, including the root and the outgroup points are drawn.

Figure 5, Figure 6, and Figure 7 illustrates the example of cuboid cladogram in 3d with the MDS clustering result. This result is generated from 446k fungal data, where 126 representative sequences from each cluster is selected, along with 74 sequence from GenBank shown in different color. The screen shot are from two different angel, one is from the side of the tree and the other is from the top of the tree. From the these figures, the correlations between the phylogenetic tree and the clustering can be easily observed. And in this graph, it shows that the phylogenetic tree and the MDS clustering have very high correlation.

To summarize, the cuboid cladogram took the following steps to generate: 1) generate the phylogenetic tree using traditional method; 2) generate the MDS clustering result; 3) Use PCA to find the plane with coordinates that has largest variance; 4) project the points from clustering onto the plane; 5) generate the cladogram from the points on the plane. Finally, by viewing the tree plot in 3D along with the clustering result, the clustering and phylogenetic analysis can be done simultaneously.



Figure The screen shot from the side of the cuboid cladogram by choosing a plane using PCA on 599nts data using MSA and WDA-SMACOF



Figure The screen shot from the bottom of the cuboid cladogram by choosing a plane using PCA on 599nts data using MSA and WDA-SMACOF



Figure The screen shot from the top of the cuboid cladogram by choosing a plane using PCA on 599nts data using MSA and WDA-SMACOF

1. Spherical Phylogram

The Spherical Phylogram has been proposed in [2]. The current distance calculation of the distances between the internal nodes and all other nodes follows the calculation method from Neighbor Joining. And WDA-SMACOF [3] is critical in this process since it can always find the global optima of the dimension reduction result. However, there may be additional option of generating the plot. One example would be use the tree distance directly. The tree distance is a distance generated during the process of a traditional tree generation method, such as RaXml. It will give a pairwise distance matrix between each pair of leaf nodes as well as the distances between the internal nodes and their direct parent or descendent. If one needs to calculate the distances as the pairwise distances among all the leaf nodes, one may use the distances as sum of branches as shown in figure 8.



Figure The example of distance calculation in a phylogenetic tree with 3 leaf nodes and 2 internal nodes.

The internal nodes cannot be directly observed because they represent hypothetical ancestor sequences, and therefore the distances from internal nodes to leaf nodes of the generated phylogenetic tree are unknown. By using RAxML, it is possible to calculate distance from an internal node to another node by using the summation over all the branches between them. For example, in Figure 8, the distance between point C and E can be calculated by summing over branch(C, B), branch(B, A) and branch(A, E). This distance calculation can generate a pairwise distance matrix for all the nodes based on all the branch lengths. However, the sum of branch lengths does not work to find the distance between pairs of leaf nodes since the pairwise distances between leaf nodes are already known from the MDS cluster visualization results. For example, the distance between leaf node C and D shown in Figure 5.5 is clearly not equal to branch(B, C) + branch(B, D). Therefore if the summation over the branches is used for defining distances during interpolation, the result will have a high bias because different distances were used for leaf nodes. But since leaf nodes already has coordinates in 3D, so this distance measurement maybe compared with the method in neighbor joining and do some future analysis.

[1] Ruan, Yang, Saliya Ekanayake, Mina Rho, Haixu Tang, Seung-Hee Bae, Judy Qiu, and Geoffrey Fox. "DACIDR: deterministic annealed clustering with interpolative dimension reduction using a large collection of 16S rRNA sequences." In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 329-336. ACM, 2012.

[2] Ruan, Yang, Geoffrey House, Saliya Ekanayake, Ursel Schütte, James D. Bever, Haixu Tang, and Geoffrey Fox. "Integration of Clustering and Multidimensional Scaling to Determine Phylogenetic Trees as Spherical Phylograms Visualized in 3 Dimensions." *Proceedings of C4Bio* (2014): 26-29.

[3] Ruan, Yang, and Geoffrey Fox. "A Robust and Scalable Solution for Interpolative Multidimensional Scaling with Weighting." In *eScience (eScience), 2013 IEEE 9th International Conference on*, pp. 61-69. IEEE, 2013.