✦ Member-only story

Hierarchical Clustering

Ravasz and Girvan-Newman Algorithms

Luís Rita · Follow

Published in The Startup

5 min read · May 29, 2020

Listen Share

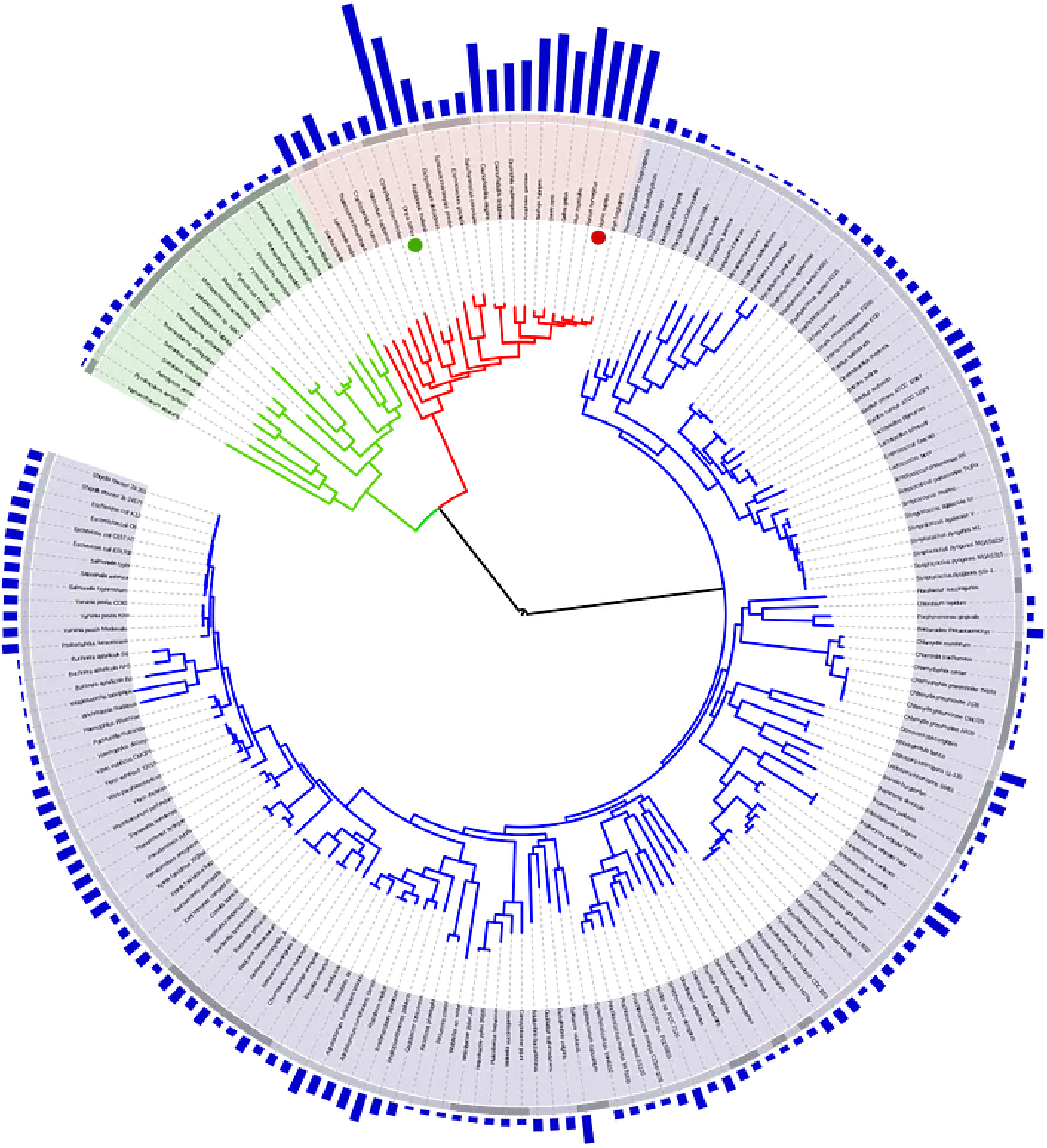

Figure 1 Tree of Life. Hierarchical clustering defining the 3 biological domains: Archaea (red), Bacteria (blue) e Eukarya (green). Source

Hereby are presented two categories of hierarchical clustering algorithms: agglomerative (Ravasz algorithm [1]) and divisive (Girvan-Newman algorithm [2, 3]).

Ravasz Algorithm

It is divided into 4 sequential steps.

Definition of a Similarity Matrix

Matrix entries can represent evolutionary distances between nodes or the number of neighbors a pair of nodes has in common. In the second case, each entry is defined as:

$$x_{ij}^{o} = \frac{J(i,j)}{min\left(k_i, k_j\right) + 1 - \Theta(A_{ij})} \quad (1)$$

This implies when nodes and do not have neighbors in common:

$$J(i,j) = 0 \Rightarrow x_{ij}^{o} = 0 \quad (2)$$

The maximum value is obtained when both nodes are connected and have the same neighbors:

$$J(i,j) = min\left(k_i, k_j\right) \Rightarrow x_{ij}^{o} = 1 \quad (3)$$

Group Similarity Criteria

After joining the most similar nodes, clusters need to be compared with the remaining elements of the network (nodes/clusters). Three clustering approaches can be used:

- Single Linkage: similarity between two groups is equal to the similarity between the most similar elements;
- Complete Linkage: analogous to the previous measure but using as reference the most dissimilar nodes from each cluster;
- Average Linkage: considers the average distance of every possible pair combination in the 2 clusters.

Hierarchical Clustering Procedure

Having defined a similarity matrix and a similarity criterion to compare clusters, the following steps are executed:

- Assign a similarity value to every pair of nodes in the network;
- Identify the most similar community/node pair and join both. The similarity matrix is updated based on group similarity criteria;
- The second step is repeated until all nodes are in the same community.

Dendrogram Cut

At the end of the execution, a single tree joining all nodes is obtained — dendrogram. Although it is possible to identify the most similar nodes, it does not return the best partition of the network. In fact, the dendrogram can be cut in one out of several levels. To solve this, modularity is calculated for each partition and the one with the highest value is chosen.

Combining the four steps, the complexity of the algorithm is estimated:

- Step 1: similarity between every pair of nodes is calculated. Complexity should be , being the number of elements in the network;
- Step 2: each community is compared against the others. This requires calculations;
- Steps 3 and 4: using a convenient structure to represent data, in the worst-case scenario, the dendrogram can be built in steps.

$$O(N^2) + O(N^2) + O(NlogN) = O(N^2) \qquad (4)$$

Despite this algorithm is slower than some presented in the next sections, it is significantly faster than the brute force approach which requires operations.

Girvan-Newman Algorithm

Instead of connecting nodes based on similarity criteria, the algorithm developed by Michelle Girvan and Mark Newman removes edges based on centrality criteria. This is repeated until none remains.

Defining Centrality

GN algorithm identifies the pair of nodes that most likely belong to different communities and removes the link connecting them. Centrality matrix needs to be recalculated after each removal. Each entry is calculated using 2 alternative approaches:

- Link Betweenness: it is proportional to the number of shortest-paths connecting all pairs of nodes that cross the respective link. Complexity is or , in sparse networks.
- Random-Walk Betweenness: after picking random nodes and , a random path between those is traced. Doing it for every combination of nodes, the average number of times the link is crossed is recorded. is proportional to this value.

The first step of this calculation requires the inversion of a matrix, thus computational complexity is . The average flowing over all pairs of nodes requires steps. In the case of a sparse network, the overall complexity is .

Hierarchical Clustering Procedure

After choosing one of the two centrality criteria:

- Calculate the centrality for every pair of nodes;
- Link with the highest centrality is removed from the network. In case of a tie, one is randomly picked;
- Centrality matrix is updated;
- Two previous steps are repeated until any link is left in the network.

Dendrogram

Similarly to the Ravasz, Girvan-Newman algorithm does not predict the best partition. Again, modularity is used to determine the optimal cut in the dendrogram.

Regarding the complexity of the algorithm, the limiting step is the centrality calculation. If link betweenness is chosen, the complexity is . The overall complexity is obtained by multiplying the previous by the number of times the centrality matrix has to be calculated — (until all links are removed). This means or (sparse network) is the final complexity.

Respecting Ravasz and GN algorithms, it is important to ask whether hierarchical structure is really present in real networks or if the algorithms are imposing it. Are there nested modules inside bigger ones? Is it possible to assess, *a priori*, if a network has this structure?

One way to check whether hierarchical modularity is present is by analyzing the clustering coefficient:

This dependence with the node's degree lets us identify whether such pattern is present. In many real networks this phenomenon is present: scientific collaboration, metabolic and citation networks. As expected, under degreepreserved randomization, community structure disappears. Resembling Erdős-Rényi random networks, where these structures are not present.

References

[1] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabási, "Hierarchical Organization of Modularity in Metabolic Networks," *Science,* vol. 297, no. 5586, pp. 1551–1555, 2002.

[2] M. E.J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review. E, Statistical, nonlinear, and soft matter physics,* vol. 69, 2004.

[3] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, no. 12, pp. 7821–7826, 2002.

Ravasz Girvan Newman Algorithms Clustering Hierarchical

Follow

Written by Luís Rita

216 Followers · Writer for The Startup

Biomedical Engineer | PhD Student in Computational Medicine @ Imperial College London | CEO & Co-Founder @ CycleAI | Global Shaper @ London | Forbes 30 Under 30

More from Luís Rita and The Startup