PRML WiSe 2016/17
Project: Leaf architecture

Supervisor: Jana Lasser Email: jana.lasser@ds.mpg.de
Phone: 0551 5176-218 Location: MPIDS, room 2.62

Description

Plants create the oxygen we breathe, therefore in a way, leaves are very crucial to our survival. Almost every plant has them and they come in very different shapes and sizes. If you look at a leaf, you will see an intricate venation pattern - a transport network. The plant uses this network to transport water from the roots up to the leaves and nutrients from the leaves down to the roots. The networks express interesting geometrical and topological features: most of the leaves have some kind of hierarchy of larger and smaller veins and express loopy structures. The idea of this project is, to identify leaves based on the venation networks they express.

Data

As this has been a research topic at the MPIDS, the data is already cleaned and readily available. The networks are stored in text files containing node-lists and edge-lists. You can download the files at `https://gitlab.gwdg.de/jana.lasser/PRML2016-projects/tree/master/leaf-architecture` (download both folders `BronxA` and `BronxB`). Node files contain number triplets (i, x, y) specifying the node index and its (x, y) coordinates in space. Edge files contain four entries (i, j, r, l): the two nodes (i, j) the edge connects and its radius r and length l. In addition to the networks there are two text files containing the network labels in pairs of plant genus and species. Note that most of the leaves in the dataset are not far away from each other in terms of their taxonomy, so don't expect too large differences. All of them are also evolutionary relatively young and stem from a tropical climate, look for example at a Ginkgo leaf to have a comparison to an evolutionary Methuselah among leaves.

Tasks

1) Make yourself familiar with the data

Investigate the database:

i) Read the graph information with your favorite network processing library (for Python we can recommend `networkX` or `GraphTool`).

ii) How large are the networks in terms of nodes and edges?

iii) Visualize a couple of leaves.

iv) Are there striking features you can distinguish by just looking at the leaves? (Look for example at the node degree).

v) Identify a set of features (geometrical as well as topological ones) that describe the networks. For example number of nodes, edge length, number of loops. Can you find a feature that captures the hierarchical organization of the leaves?

2) Data analysis

Employ methods you learned in the PRML class to analyze the dataset. Things to do:

i) Use principal component analysis to find out, which of the features contribute most to the variance in the dataset. Drop the most insignificant features. Be wary of the number of nodes: this feature might dominate all the others but be very insignificant in terms of information about the leaf venation structure it represents.

ii) Look at the labels and select the genera that have at least 4-5 members as groups. Use a clustering algorithm like KMeans to see, whether the networks belonging to the same genus cluster. Visualize the clusters, try to make sense of them in terms of physical appearance of the leaves. Be aware that the data might not cluster as nicely as the examples shown in class as the leaves within the same genus come from different species and also have quite some intra species variance.

iii) Select five pairs of networks that have the same (genus, species) label (so 10 networks in total). Cut each of them into 20 or more similarly sized pieces to have a larger sample size per label (how large do the pieces need to be to retain the network's characteristics?).

iv) See, whether the network pieces form clusters based on the leaf they belong to or rather based on their position within the leaf (for example being close to the tip or the main vein of the leaf).

v) Create a set of random artificial networks with the same node degree distribution, average edge length, average edge width and node number as the network pieces (they don't need to be planar if that is too difficult!). See whether your artificial networks form a separate cluster.

3) Optional

If you want to dig deeper into the topic, one way to go would be to try and add more and different leaf species to the dataset.

i) Have a look at `http://clearedleavesdb.org/all_collections` and select a suitable collection of leaves. Be wary of resolution/contrast differences in the images. The networks in our collection were extracted from images scanned with a resolution of 6400 dpi, networks from leaves scanned with significantly lower resolution might not be comparable. Why? Is there a way to overcome that problem?

ii) Extract networks from the images. We suggest to use this tool `https://github.com/JanaLasser/network_extraction`: it is native python and Jana created it, so she can help you. Or this one `http://nefi.mpi-inf.mpg.de/` which might be easier to get to work.

iii) Compute the features you identified as important for the new networks.

iv) Employ clustering again to compare the new networks to the old ones.