Creating Genome Taxonomies with Growing, Hierarchical, Self-Organizing Maps

John Bethencourt
We classify organisms into hierarchies based on similarities we see.

When we match them on noticed similarities, unnoticed similarities often match up as well.

This is because we are mimicking the hierarchy of evolution with our taxonomies.

The closer our taxonomy is to the “natural” hierarchy of evolution, the better it should predict attributes of animals.
Motivation

- However, very simple microorganisms can be hard to classify because they are hard to observe.
- Many biologists do not have a way of placing microorganisms in a useful hierarchy.
- But now we have the complete genomes of many microorganisms.
- What if we let an unsupervised learning algorithm try to make a taxonomy for them?
The Idea

- Apply hierarchical clustering algorithm to the problem.
- Specifically, we apply the growing, hierarchical, self-organizing map (GH-SOM).
- The GH-SOM builds a tree of SOM’s. The tree represents a classification hierarchy.
- The structure of the tree is determined by the data; we don’t have to predefine it in any way.
Feature Selection

- We can’t put strings of DNA directly into the SOM algorithm’s; we need to generate feature vectors from the genomes.
- How can we generate meaningful features?
- It turns out that substring frequencies work pretty well.
Yep, It Works

We did this. Here is an example classification hierarchy it made:
Yep, It Works

We also made a tool that generates web pages for exploring the hierarchy: