

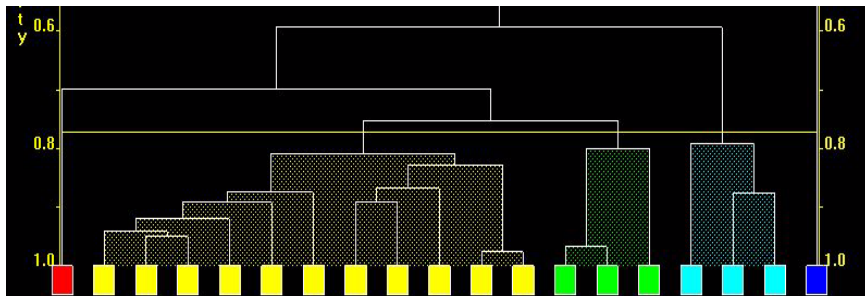
Reading a Dendrogram

Selecting *Dendrograms* from the tab bar running along the top of the screen opens the dendrogram display.

Basics:

Each box at the bottom of the tree diagram represents a pattern from the dataset provided. The label or filename of each sample is displayed below each box. If the dataset is very large it may not be possible to see each box individually and labels will not be displayed (hovering the mouse over a box shows a tooltip label instead). Each box has a vertical line coming from it, which is joined to the other vertical lines by horizontal lines called *tiebars*.

Figure 1.a - a dendrogram displaying good clustering

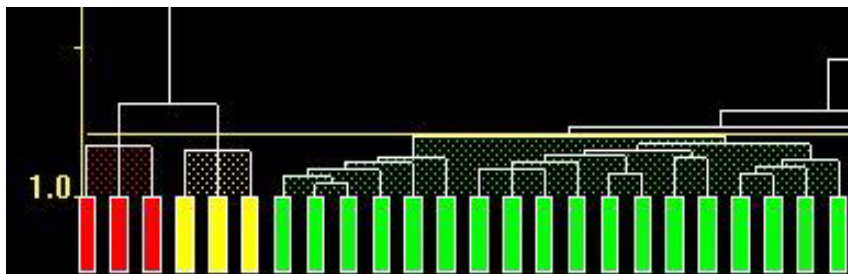


There is a *similarity scale* running vertically up the side of the display. This is a measure of how similar the any two samples are. **The lower a horizontal tie bar connects two samples, the more similar they are.** Two samples joined by tie-bars higher up the dendrogram will be less similar.

Example:

In figure 1.b the samples in the red cluster are similar and are joined by a low horizontal tie-bar at around 0.95. The samples in the yellow cluster are also similar to **each other** but not to those in the red cluster as they are joined to the red group by a higher tie-bar, so at this cut-level are in a different cluster. The samples in the green cluster are even less similar to those in the red cluster as they are joined by an even higher horizontal tie-bar (out of shot).

Figure 1.b - clusters in a dendrogram



Cut-level:

There is a horizontal yellow line intersecting the display. This is called the *cut-level*. It is a calculated cut-off point that determines when samples are similar enough to be classified together in a group or *cluster*. **Samples joined by tie-bars below the cut-level are considered to be in the same cluster and are coloured to show this.** These colours are then used in most of the other display screens.

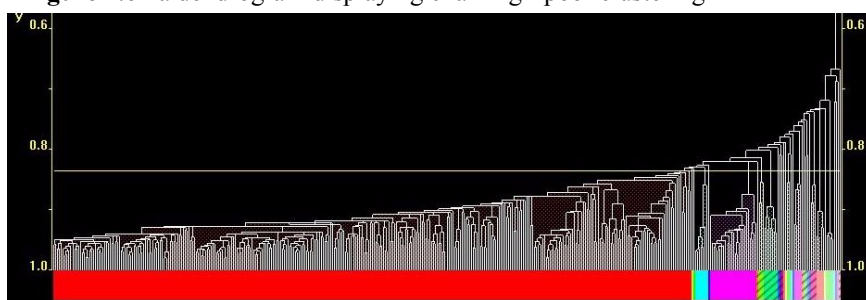
Good and bad clustering:

A good cluster in the dendrogram should have all the sample boxes corresponding to a single cluster joined by reasonably low tie bars (having high similarity), while different clusters are separated linked by high tie bars (low similarity). A good indicator of good clustering is when there is a large separation between the tie bars within a cluster and the tie bars linking clusters (figure 1.a - where the light blue cluster is well separated from the rest).

A dendrogram which shows 'chaining', *i.e.* where each sample is linked to the next by a slightly higher tie bar, gradually stepping up evenly from left to right, suggests that there are no clusters in the dataset (figure 1.c)

A dendrogram where two sample in the same cluster are joined by a relatively high tie bar may indicate that the analysis should be more discriminate in classifying clusters, and that the cut-level should be lowered to split the cluster into smaller groups. Sample effects - such as preferred orientation, or height

Figure 1.c - a dendrogram displaying chaining - poor clustering



variances, can affect how similar two patterns appear - these issues need to be kept in mind when determining the optimum cut-level.

Manipulating the dendrogram:

The cut-level is calculated for each individual dataset and can be adjusted and experimented with by the user. This is useful as *PolySNAP* only defines an initial cut-level according to the results of calculations, while the user can examine the data and decide what would be a sensible and appropriate cut-level based on the level of differences being investigated.

To do this either: Hold down the *Control* key and the left mouse button and drag the mouse up or down to raise or lower the cut-level, or select *Set Cut-level to...* from the *Tools* menu. Enter in a numerical value between 0 and 1.0 corresponding to the similarity scale and click *OK*.

The cut-level can be reset to the original calculated value at any time by selecting *Undo Saved Dendrogram Modifications...* from the *Tools* menu and selecting the *Original* setting.
