Pipeline

Raw data (Input):

The raw data is coming from the reflectance data collected from the 'Hyperspectroscopy cameras'. Raw data is in the form of a ".csv" file with incident light wavelengths and reflectance values of 16 different sample materials. Comprising of 4 fabric materials and each material with 4 different colors (Figure 1 and Figure2). Figure 3 is a plot of raw data shown in Figure 2.

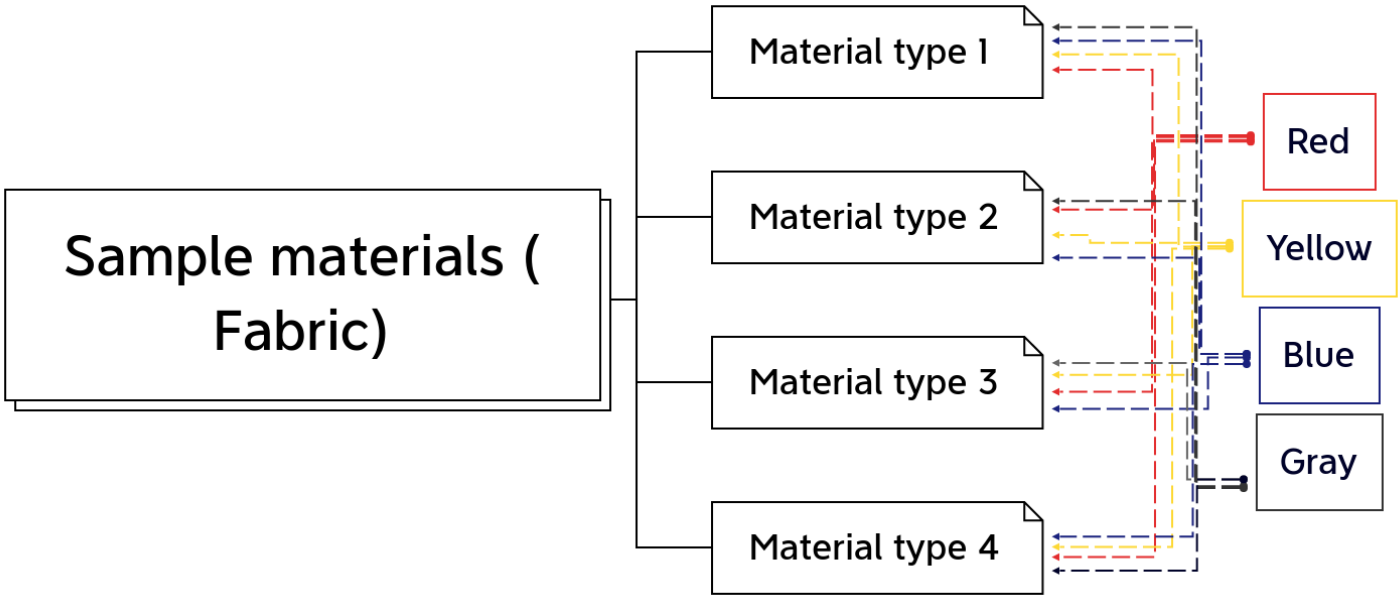


Figure 1 - Sample materials

Wavelength of the Indcident light, in 'nano meters'

Reflectance of the 'sample materials'

		1	2		287	288
Label Number	Material name	951.544641612916	956.998116091828	...	2511.23834258162	2516.69181706053
1	1075gelb	0.59115140858586	0.59030327826851	...	0.394922754793327	0.397803223546645
2	1075hellgrau	0.571173722772351	0.570794562955049	...	0.369977012696307	0.373473334759174
3	1075rot	0.586722799696014	0.586081682433926	...	0.39330496344823	0.397199899159867
4	1075royalblau	0.593474683813998	0.592612829642278	...	0.392190649413753	0.395550036250503
5	Corduragelb	0.510537327522183	0.509869579728641	...	0.270302565416906	0.273381209452828
6	Cordurahochrot	0.617322760768579	0.615917606431278	...	0.3826009572882	0.385767109268578
7	Corduraroyalblau	0.59307525534908	0.592044673798604	...	0.320296662524671	0.323151750784618
8	Cordurasilbergrau	0.473228041944632	0.471457960945864	...	0.324771623379043	0.32757200692709
9	Kochwollebordeaux	0.672600042645796	0.671780428284893	...	0.171787651842325	0.173639025745546
10	Kochwollecurcuma	0.655491935936317	0.65451373381008	...	0.167113429111089	0.169149058879548
11	Kochwolleindigo	0.556477006716623	0.556540500448218	...	0.175376291338795	0.17675413382213
12	Kochwollemelange	0.684516590580568	0.683331145817503	...	0.17977780361176	0.181717500381907
13	Leinengelb	0.599350984938449	0.598989577331253	...	0.216791001240517	0.221606939650882
14	Leinenhellgrau	0.554626578394312	0.556582169784338	...	0.218544271825407	0.223590630337517
15	Leinenhochrot	0.627794508923027	0.627019916877631	...	0.235059891561462	0.240406899152088
16	Leinenroyalblau	0.615877337459886	0.615118340660878	...	0.237979528520956	0.243726569908284

Figure 2 - Raw reflectance data looks like:

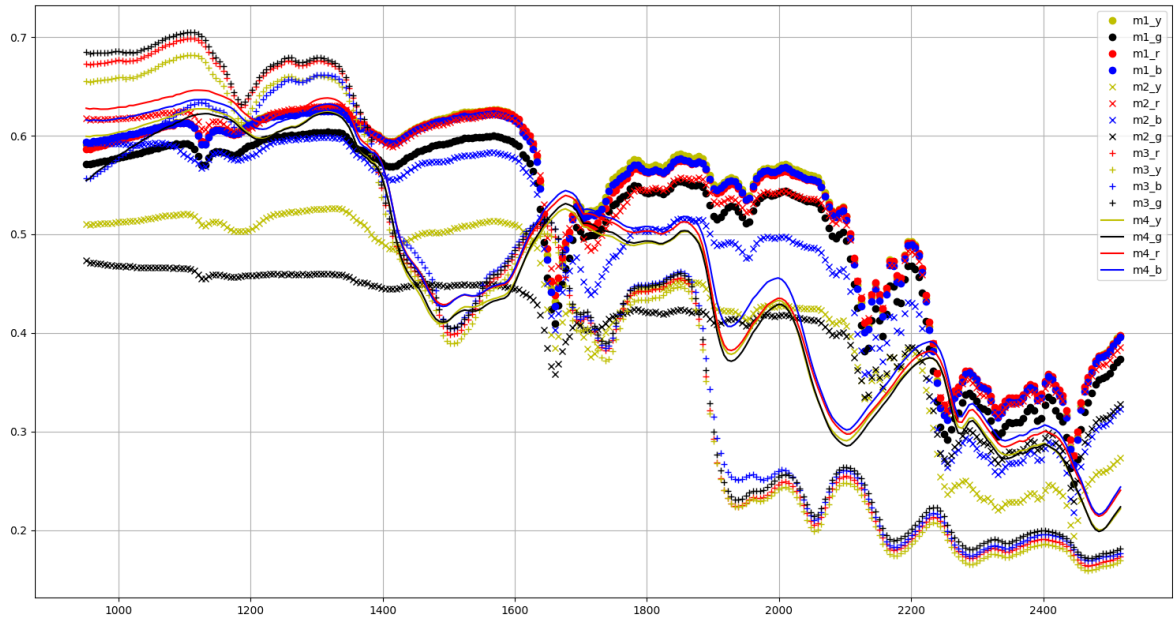


Figure 3 - X-axis: Wavelength of incident light (~900nm to ~2500nm). Y-axis: Reflectance values (0 to 1)

Normalisation 1 - 'Independent material' wise :

Each material's reflectance sequence is min-max scaled between 0 and 1. This step is performed independent of other material's reflectance sequences.

Global normalisation (normalisation with respect to 'maximum' and 'minimum' across the reflectance values of all materials) did not yield significantly different results when compared to an unnormalised sequence, because, the maximum amongst all the materials is very close to 1 and the minimum amongst all the materials is very close to 0. That is the reason why the images 2 and 4 appear very similar.

Figures 4, 5 and 6 are Comparison between unnormalised sequence, material wise independently normalised sequence and Globally normalised sequences, respectively.

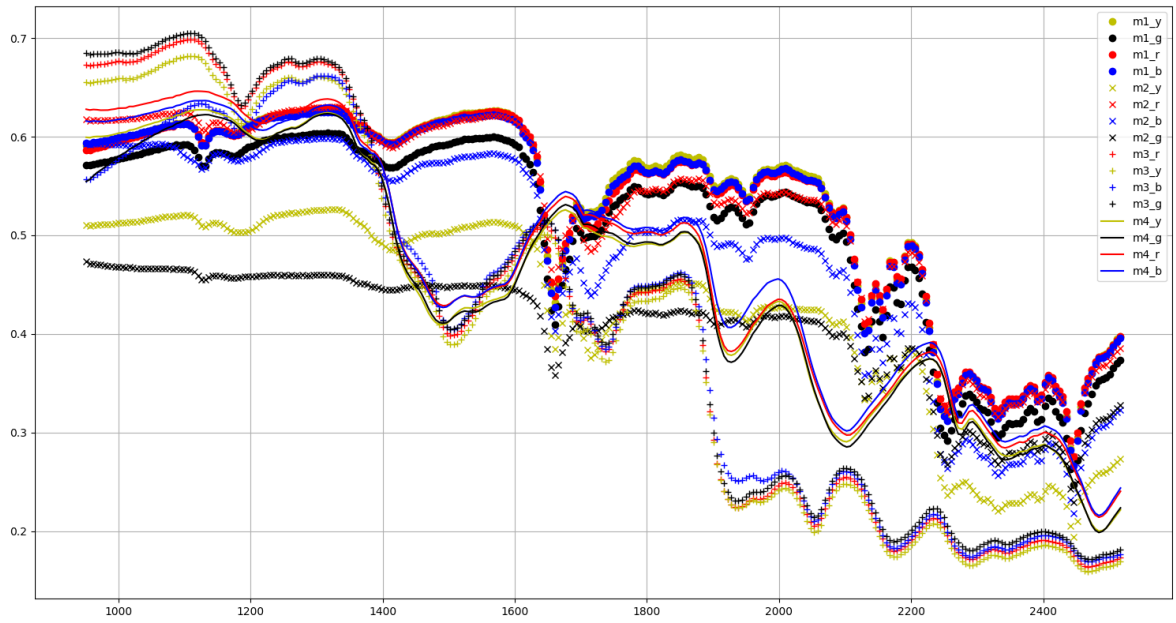


Figure 4 - unnormalised data

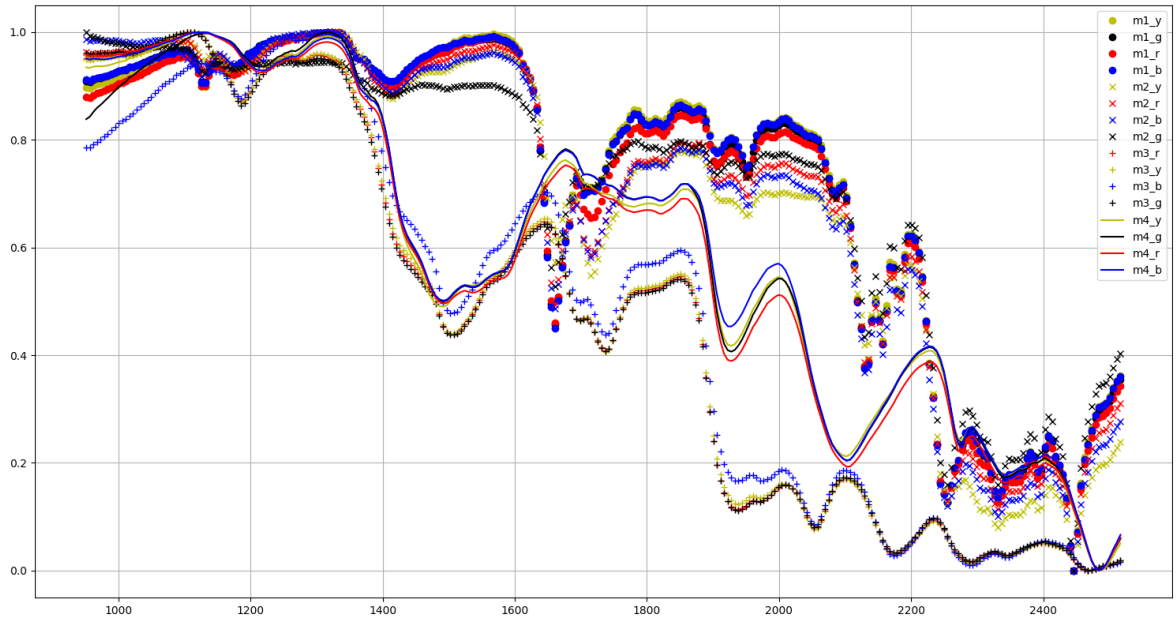


Figure 5 - Material wise independently normalised data

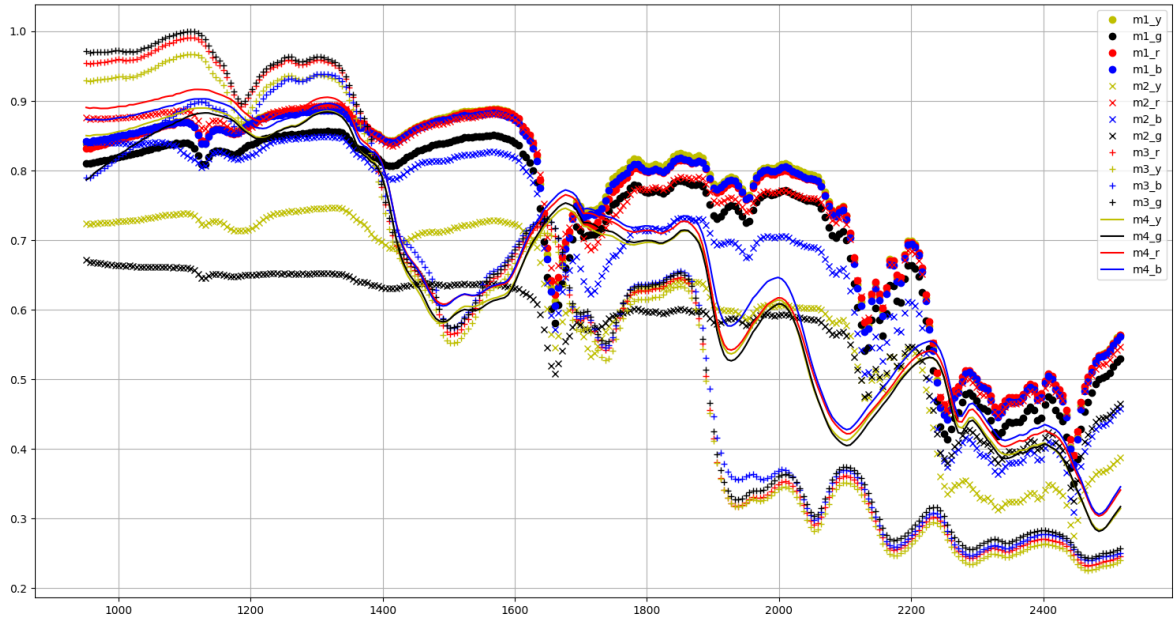


Figure 6 - Global normalised data

Numerical sequence to character sequences:

Phylogenetic algorithm expects the data to be in the format of characters. So, the reflectance values that range from 0 to 1 are mapped to ASCII characters that range from "A" to "~" as shown here:

> ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789!"#$%&'()*+,-/:;⇔?@[]^_`{|}~

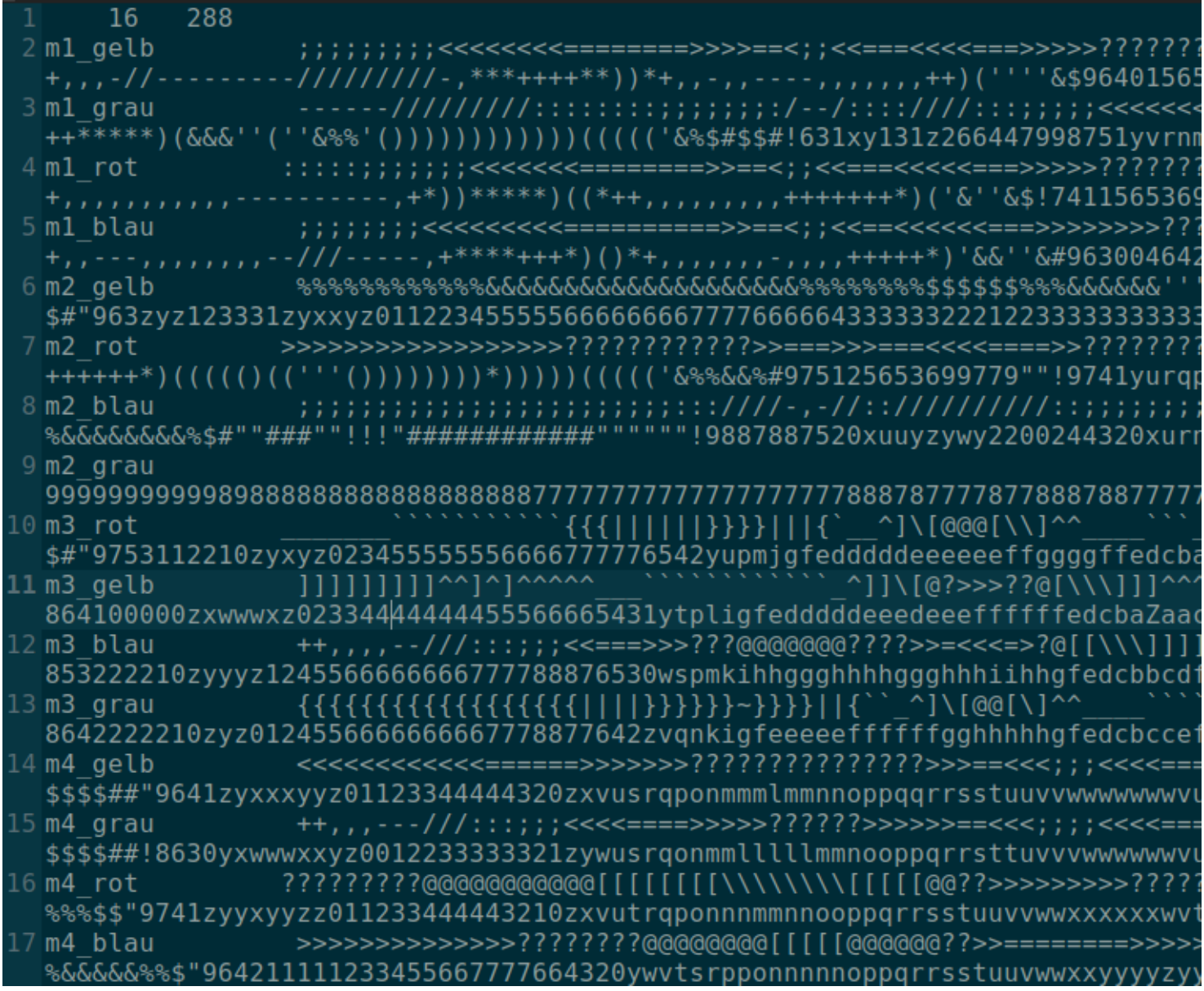


Figure 7 - Reflectance data mapped to 93 ascii characters

Variable binning:

- The idea is to control the number of ASCII characters used to map the reflectance values based on the similarity between the material sequences.
- The input original reflectance sequence of 16 materials are divided into, say, 'k' sections. in each section pairwise DTW distance of each material wrt every other material is calculated and summed. This sum corresponds to the "Overall Distance" of that 'section'.
- In a given section:
    - If the reflectance sequences are dramatically different to each other:
        - that means the "Overall Distance" is a big and the number of characters that encode the numerical sequence should be small in number.
            - Why? Because, since the sequences are pretty dissimilar with respect to each other, not many characters are needed to differentiate.
    - else:
        - "Overall Distance" is a small number if reflectance sequences are indistinguishable each other. Implies, the number of characters that encode the numerical sequence should be a large number.
            - Why? because, since the sequences are pretty similar with respect to each other, encoding with more number of characters means capturing more detail.

Section cuts:

The incidend light from the Hyperspectroscopy camera is in the range of ~900 to ~2500nm. Reflectance behaviour response of the 16 sample materials within this wavelength range can be subdivided in many ways. Figure 5, 8, 9, 10, 11 show some of the ways the total wavelength range can be sub divided into.

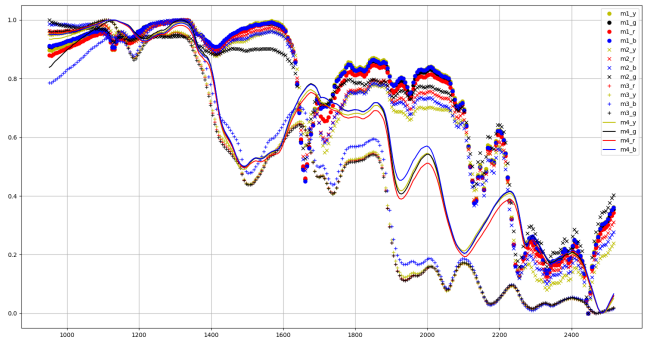


Figure 5 - Material wise independently normalised data

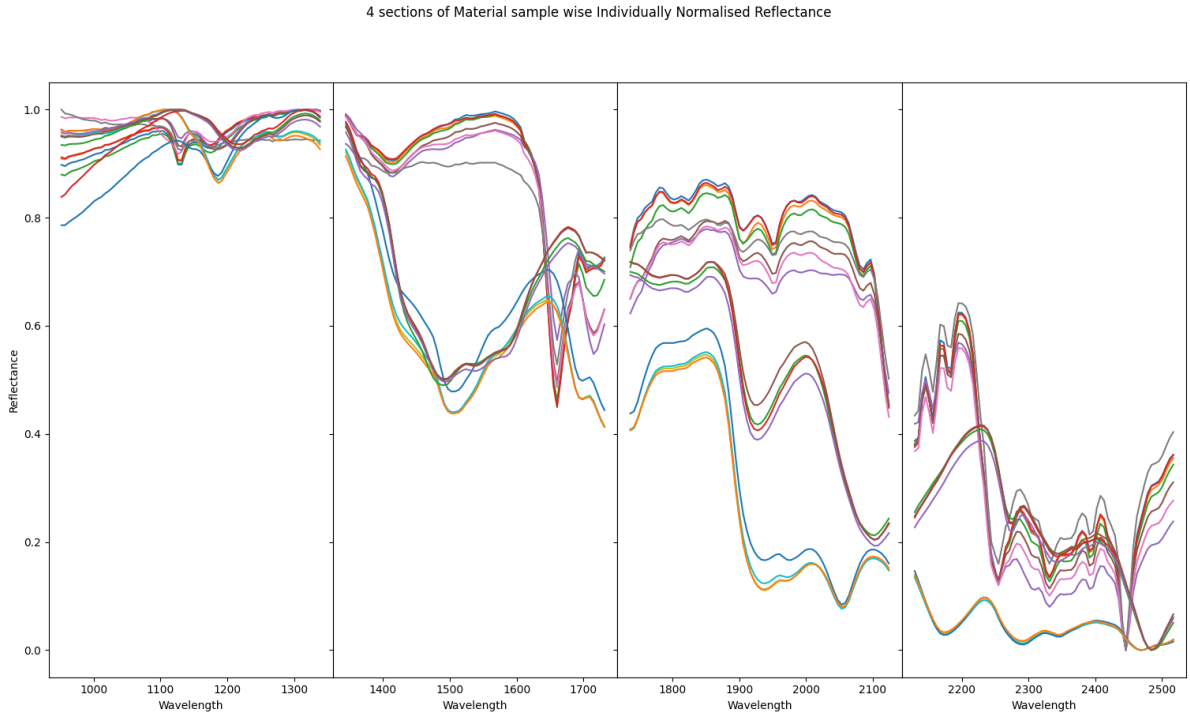


Figure 8 - 4 sections of Material sample wise Normalised Reflectance

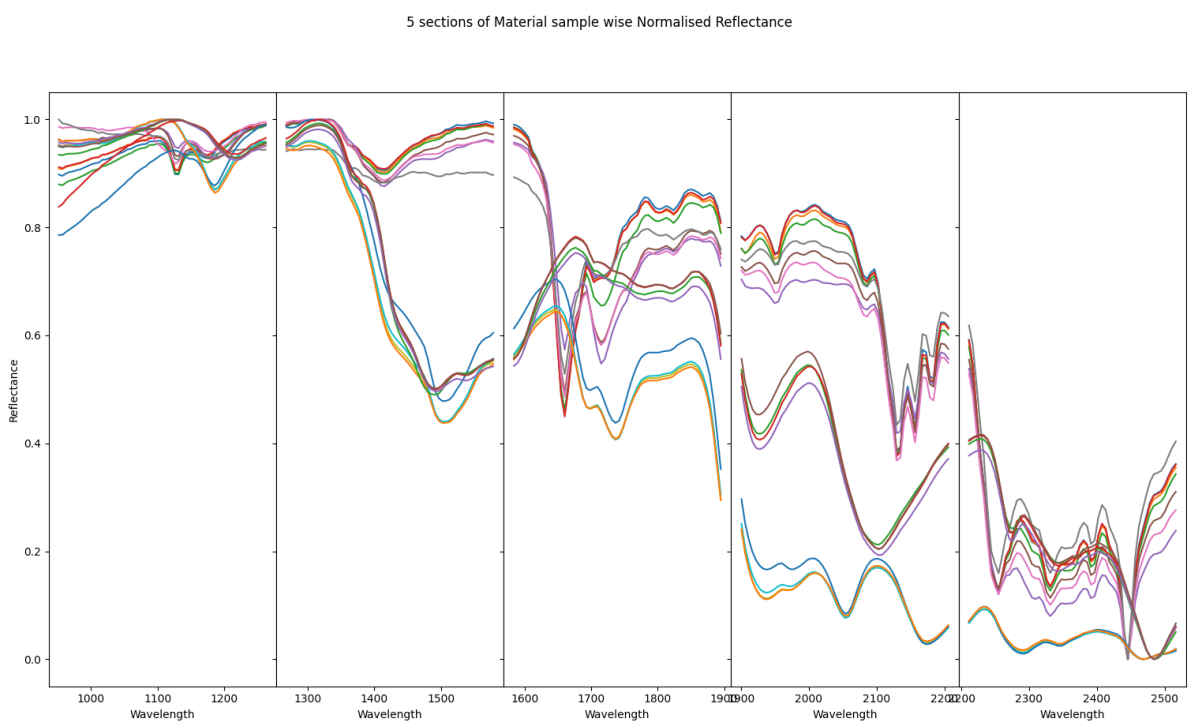


Figure 9 - 5 sections of Material sample wise Normalised Reflectance

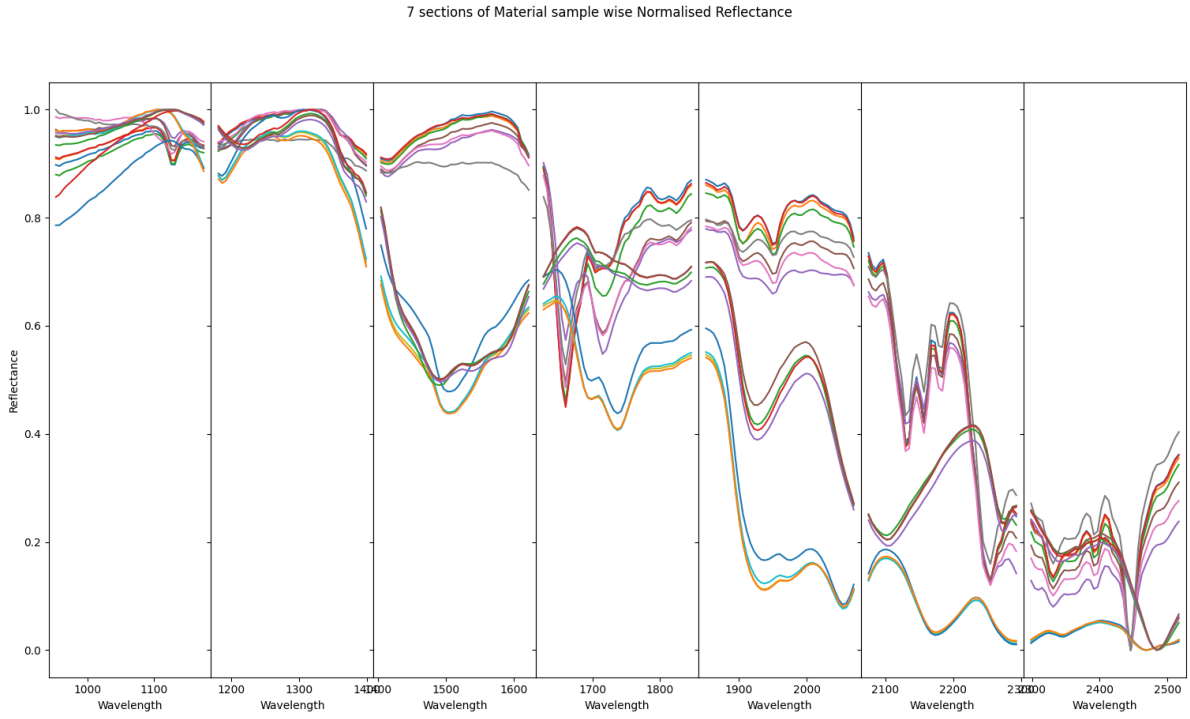


Figure 10 - 7 sections of Material sample wise Normalised Reflectance

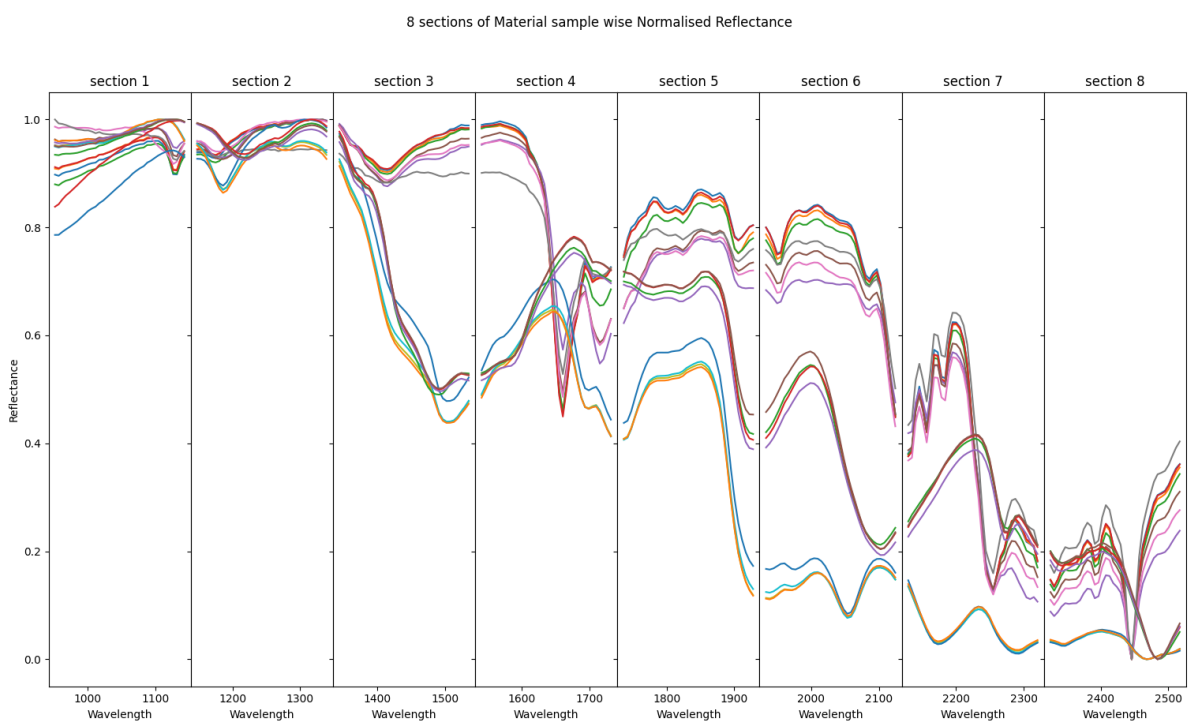


Figure 11 - 8 sections of Material sample wise Normalised Reflectance

Each section is used to generate a new phylogenetic tree. The numerical reflectance sequence is mapped to the 93 ASCII characters. Application of variable binning in each section to decide the number of the characters is optional. It is ideal to use as many characters as possible.

Normalisation 2 - Section wise:

Max-min scaling/normalisation can be applied again to the numerical reflectance sequence in each section. This ensures complete utilisation of all the characters during encoding of numerical sequence into character sequence. Figure 12 and Figure 14 illustrates

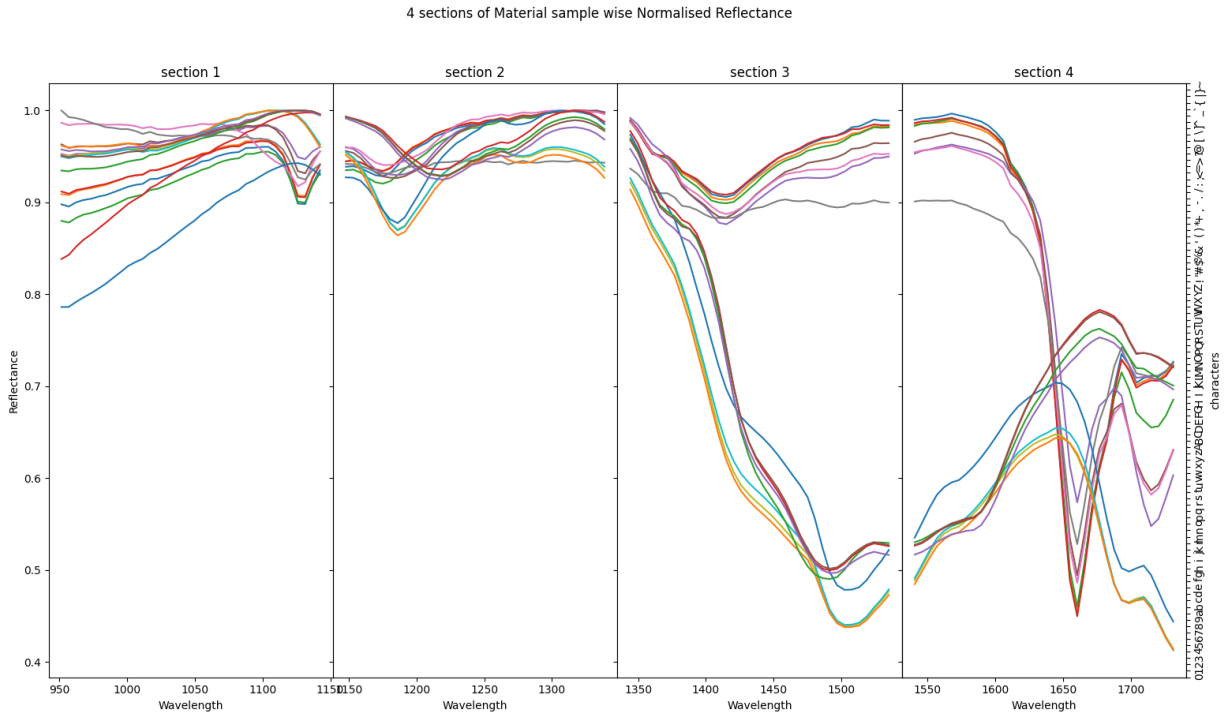


Figure 12 - 4 sections of Material sample wise Normalised Reflectance with 'Character'-axis

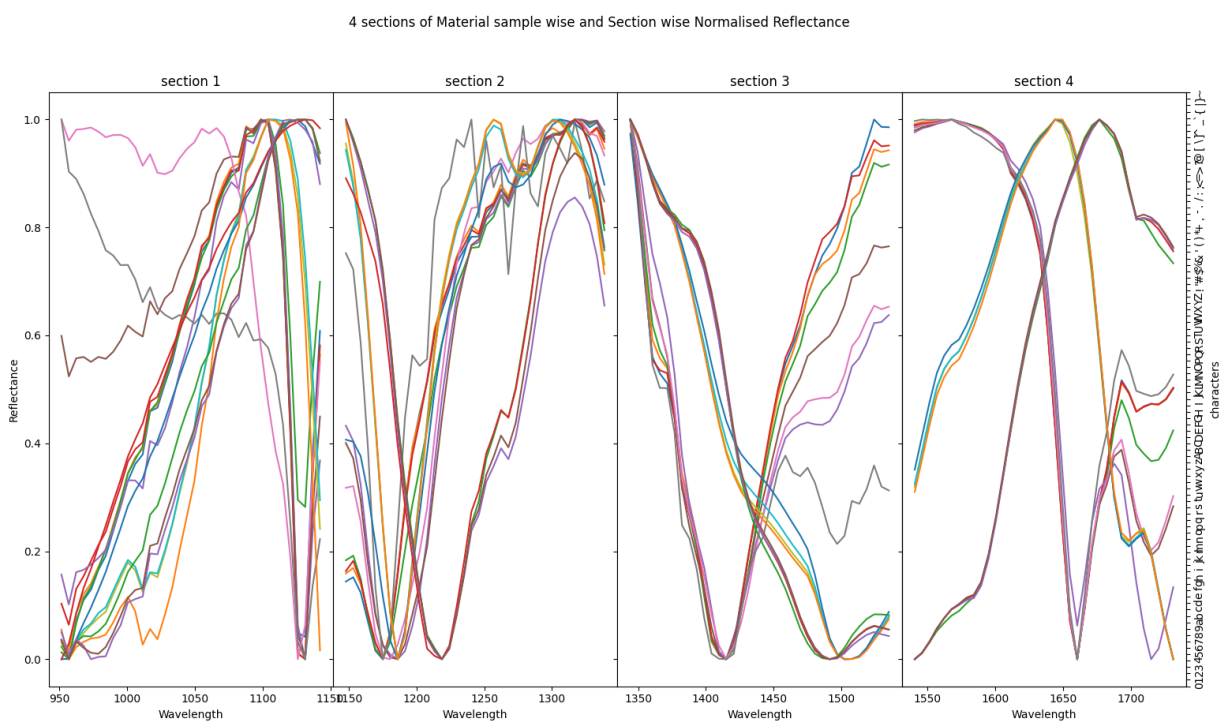


Figure 13 - 4 sections of Material sample wise and Segment wise Normalised Reflectance with 'Character'-axis

Character sequence to .phy file.

A ".phy" file is an acceptable input for the phylogenetic Maximum parsimony algorithm. One ".phy" file must be generated for every section. Figure 7 is one of such ".phy" files.

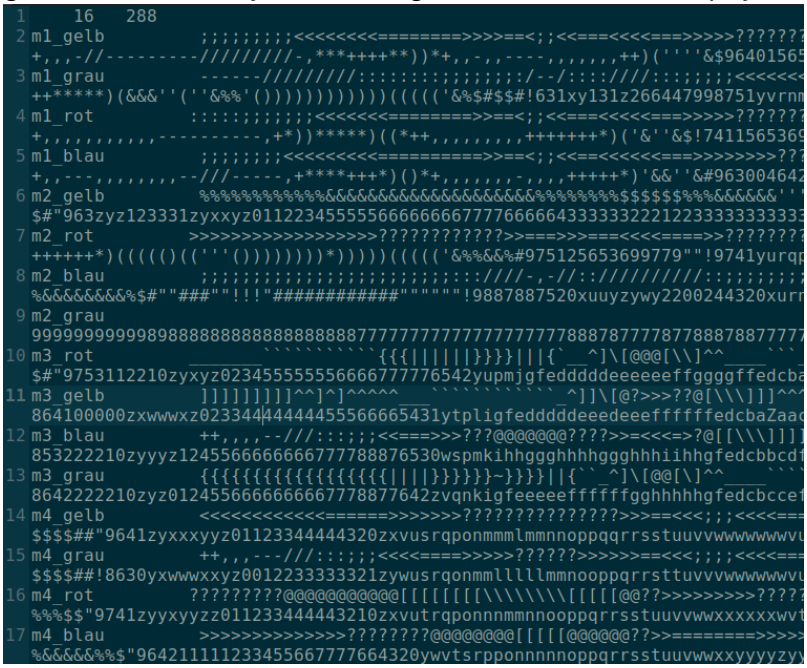


Figure 7 - Reflectance data mapped to 93 ascii characters

Character_sequence.phy to Unrooted phylogenetic tree:

Each .phy file is now the input to a tree building algorithm pipeline.

Tree building pipeline:

FIgure 14 illustrates the pipeline for maximum parsimony in a simplistic manner.

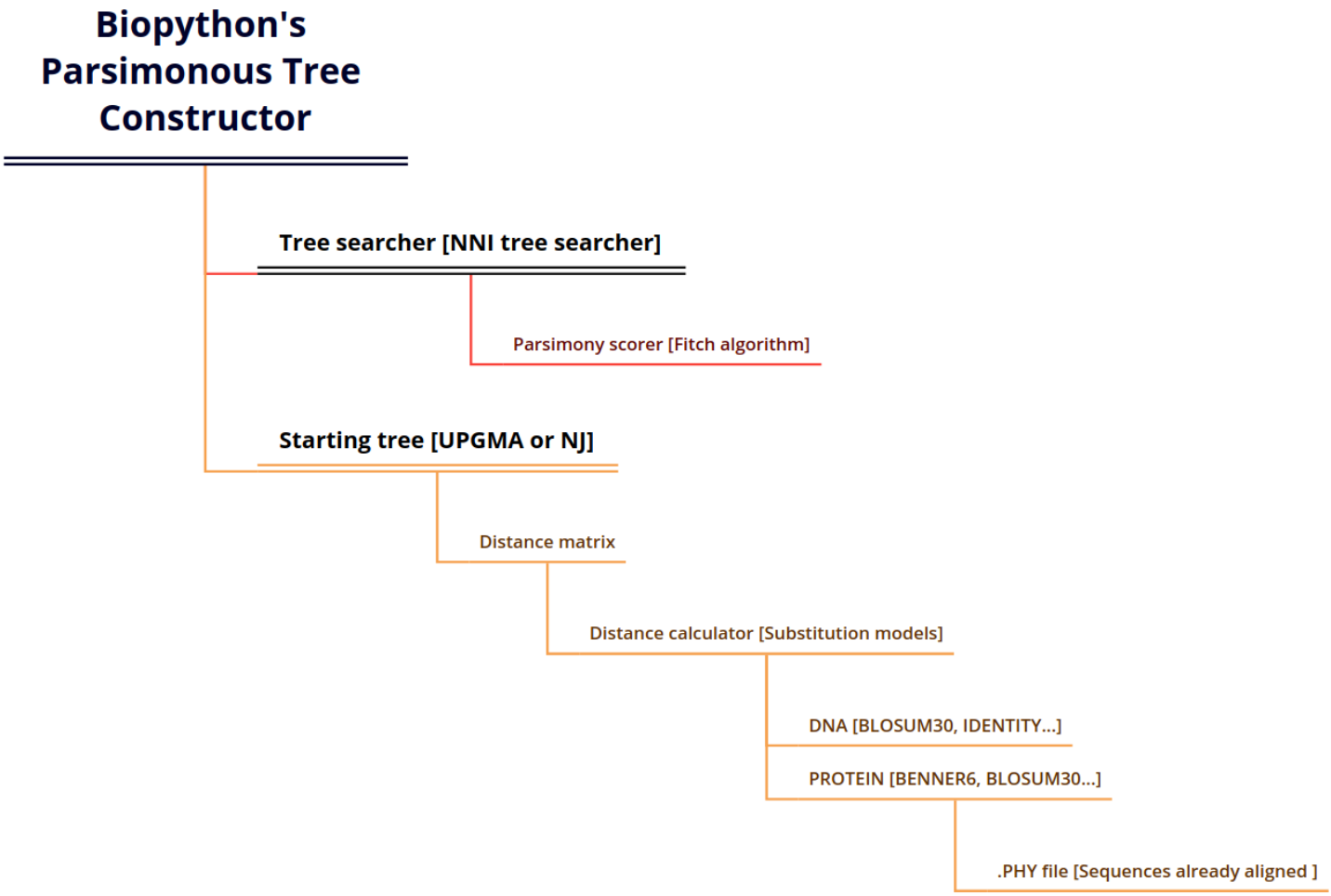


Figure 14 - Steps involved in building a 'Maximum Parsomony' tree.

- Input: .phy file with character sequences of 16 sample materials.
- Processing:
    - Substitution matrix
        - Benner93 - which is a substitution matrix that is obtained by modifying the original Benner6 substitution matrix. It contains 1's in the diagonal elements and -1's as every other element. i.e, any change is sequence is equally unappreciated. Figure 15 is the substitution matrix that is modified to contain only 1s and -1s.

```
#  S.A. Benner, M.A. Cohen, G.H. Gonnet:
#  "Amino acid substitution during functionally constrained divergent evolution
#  of protein sequences".
#  Protein Engineering 7(11): 1323-1332 (1994).
#  Figure 3A.
#  PMID 7700864
    A    B    C    D    E    F    G    H    I    J    K    L    M    N    O    P    Q    R    S    T    U    V    W    X    Y    Z
A  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
B -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
C -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
D -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
E -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
F -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
G -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
H -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
I -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
J -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
K -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
L -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
M -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
N -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
O -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
P -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
Q -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
R -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
S -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
T -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
U -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0
V -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0 -1.0
W -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0 -1.0
X -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0 -1.0
Y -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0 -1.0
Z -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0 -1.0  1.0 -1.0
```

Figure 15 - Modified Substitution matrix

- Distance matrix:
    - Biopython generates a distance matrix with the help of the defined substitution matrix and the input ".phy" file.

- "Starting tree" constructor:
    - A start tree is constructed using the distance matrix and "**Neighbour Joining**" algorithm. there are alternatives such as UPGMA(Unweighted Pair Group Method with Arithmetic Mean) algorithm.
- Tree searcher algorithm:
    - An algorithm called '**Nearest Neighbor Interchange**' tree search algorithm is used to search for maximum parisimonous tree in the tree search space. It starts the search with the "Initiation tree/ understandingStarting tree".
    - Tree searcher algorithms judge a given tree in the tree search space using parsimony scoring algorithms like "Fitch algorithm" and "Sankoff algorithm".

- Resultant tree: A maximum parsimonous rooted tree is obtained.

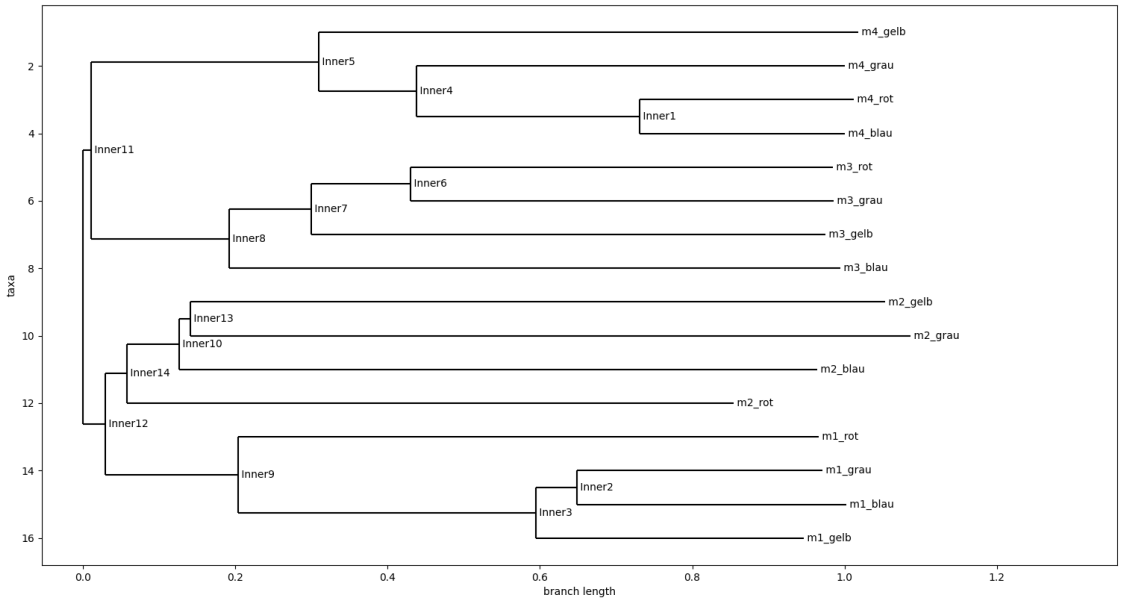
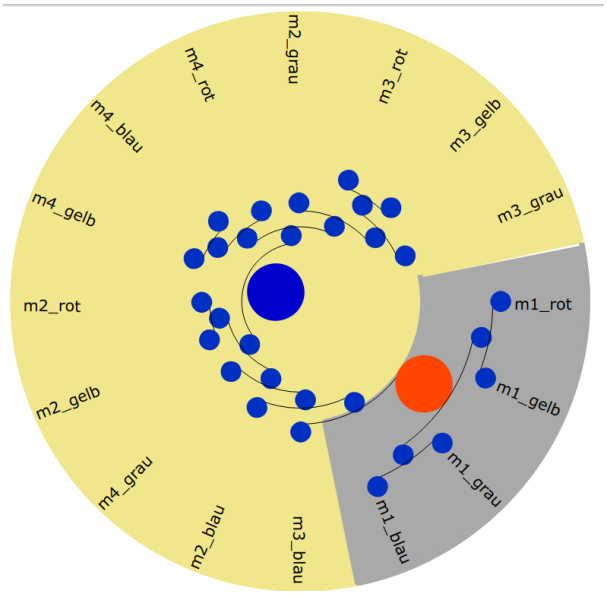


Figure 16 - Rooted tree generated from 4th section of 4 section cut seen in Figure 8.

- Representation:
    - The rooted trees are rearranged and represented in the following form for better visualization of the clustering.



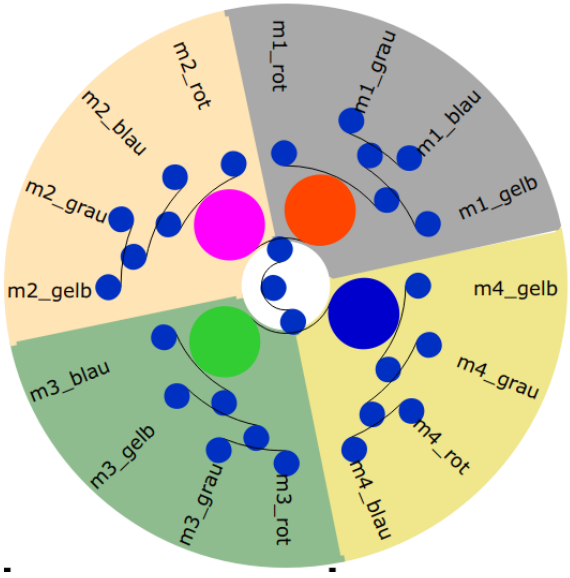


Figure 17 and 18 - 1st of 4 segments and 4th of 4 segments generated clustered trees as visualised here.

In the figures 17 and 18, the big nodes represent cluster heads, i.e, the cluster heads have 1 or more leaf nodes. In the next step called "Cluster check" the trees can be judged objectively.

Output: Cluster Check

A systematic check is performed on the resultant maximum parsimonous tree where the system we have built can confirm whether the tree has succefully clustered the tree or not.
In this test, If atleast one cluster head consists of leaf nodes that belong to more than 1 material type, The whole tree is labeled as "Bad Tree". If all the conditions are satisfied, It gives a "Good tree"