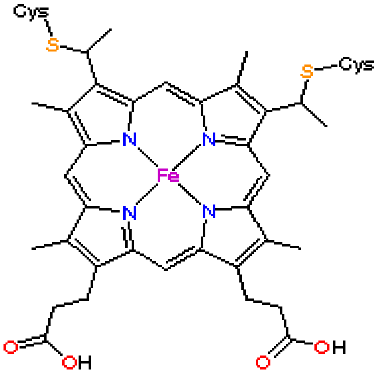
Martina Koeva
Biochemistry 218 Final Project
Winter 2001-2002

Strengths and weaknesses of parsimony and distance methods, used in the PHYLIP 3.6 software package

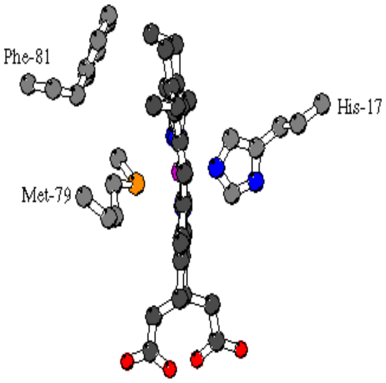
Introduction

Cytochrome c is an electron-transport protein that has been studied and described in a great detail, as well as has been used in a variety of studies for other purposes. It is a soluble mitochondrial matrix spherical (diameter = 34 A) protein that is associated to the inner mitochondrial membrane and has the function of transferring electrons between respiratory chains III and IV. Cytochrome c is present in almost all organisms that have mitochondrial respiratory chains, including plants, animals and microorganisms (Stryer 1988).



The sequence for cytochrome c has been determined for a large number of organisms (more than 70) and is usually 104 amino acid residues long. The consensus pattern for its family is Cys-X-X-Cys-His (or C-{CPWHF}-{CPWR}-C-H-{CFYW}). It consists of one polypeptide chain and has covalently attached heme group by means thioether bonds to two conserved cysteine residues. "The iron atom is bonded to the sulfur atom of the methionine residue and to the nitrogen atom of the histidine residue," (Stryer 1988) both of which have been conserved in all species. The sequence has 26 completely conserved residues, including the cysteine and histidine residues mentioned above. A large number of studies on the cytochrome family evolution have been done.

The purpose of this paper will be to discuss some strengths and weaknesses of the parsimony and distance methods, examined by means of phylogenetic analysis of the cytochrome c family, using the PHYLIP software package, provided by Joseph Felsenstein.



The following programs were chosen for the purposes of this study:
PROTRAPS – uses parsimony method from protein sequences to build a phylogenetic tree
PROTDIST – uses maximum likelihood estimate (based on PAM, or Kimura matrices) to calculate distance matrices for protein sequences;

FITCH – uses distance matrices to build a phylogenetic tree, assuming additivity of the branch lengths for the species
KITSCH – uses distance matrices to build a phylogenetic tree, assuming additivity of the branch lengths of the species, as well as a molecular evolutionary clock
CONSENSUS – uses the majority-rule consensus tree method to analyze and find the best tree (the strict consensus tree)

Methods

2. Background

- Searched the PROSITE list of documentation entries to find the cytochrome family (PDOC00169 or PS00190)
- Selected protein sequences for 36 species from 87 listed as true members of family
- Extracted sequences from the appropriate 36 Swiss-Prot entries
- Compiled 3 different datasets – 2 plant datasets (9 and 12 sequences each), and 1 animal dataset (15 sequences) – see Figure 1 in Appendix

2. File Formats

For PROTRAPS and PROTDIST all sequences were saved in 3 separate files in Notepad, with .txt extensions. The format for each entry in every file was as follows: the name of the species (at most 10 characters long), followed by the sequence itself on the same line. At the beginning of each file the number of species, as well as the length of the sequences was specified:

```
          9    104
MungBean  ASFDEAPPGNSKSGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQSGTTAGYSYSTANKNMA…
CastorBea  ASFBZAPPGBVKAGEKIFKTKCAQCHTVEKGAGHKQGPNLNGLFGRQSGTTAGYSYSAANKNMA…
Sesame      ASFBZAPPGBVKSGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQSGTTPGYSYSAANKNMA …
…
…
```

The matrices generated by PROTDIST were saved as .txt files for both the PAM and the Kimura matrices for subsequent use in the reconstruction of trees using FITCH and KITSCH. The number of sequences in each dataset was specified at the beginning of each file:

```
 9
MungBean    0.09240 11.77498  0.03939 … …
CastorBea    0.04044 11.41536  0.07352 … …
Sesame        0.06058 11.77662  0.06211 … …
… …
… …
```

The treefiles generated by PROTRAPS were saved as .txt files for subsequent use in the construction of the best (strict consensus) tree by Consensus:

```
(((CastorBea,(Maize,(Sesame,(LoveInMis,(MungBean,Pumpkin))))),(SeaIslCot,Leek          __ATF)),Indianma)[0.3333];
```

```
(((SeaIslCot,(CastorBea,(Maize,(Sesame,(LoveInMis,(MungBean,Pumpkin)))))),Leek        __ATF),Indianma)[0.3333];
((SeaIslCot,((CastorBea,(Maize,(Sesame,(LoveInMis,(MungBean,Pumpkin))))),Leek        __ATF)),Indianma)[0.3333];
```

3. **Experiment variation**

In PROTRAPS, 3 separate experiments were run on each dataset:

1.) no variation in the sequences' input order
2.) 5 randomized rearrangements of the sequences' input order
3.) 10 randomized rearrangements of the sequences' input order.

In PROTDIST, 2 separate experiments were run on each dataset:

1.) using a PAM matrix
2.) using a Kimura 1983 matrix

No additional experiments were done in FITCH, KITSCH, or CONSENSUS, except for the ones necessitated by the results from the other programs.

Results

PROTRAPS – For each dataset, 3 output files were generated. First, the sequences were aligned without variation in the input order, using the first sequence in each dataset as an outgroup for the phylogeny. Then, by means of the 'J' option in the PROTRAPS menu, the sequences' order was randomly rearranged and the new most parsimonious trees were searched for. The procedure was chosen to be performed 5 times, and then 10 times and the program selected the best trees over all phylogenies built. The results were as follows:

	# of trees for non-random order	# of trees for randomized input order – 5 times	# of trees for randomized input order – 10 times
Plant dataset1	7	3	3
Plant dataset2	4	4	4
Animal dataset	22	40	50

Figure1: Number of trees generated for each dataset with each procedure by PROTRAPS – non-randomized and randomized input sequence order

In search of the overall best tree, the trees generated in each procedure were inputted into CONSENSUS, which generated an overall strict consensus tree for each method used. The graphical and statistical results for each dataset have been shown in Section 2 of the Appendix.

For plant dataset1, the results – both the structure of the trees and the number of times particular groupings occurred in the order given – were identical for two procedures, in which the sequence input order was randomized. Similarly, for dataset2 (including the 3 yeast sequences) the trees and the corresponding numbers were an exact match for all three procedures.

For the animal dataset, the results indicated that although the groupings were formed between the same species in each procedure, the number of times that the particular taxon occurred among the trees changed for the group, formed by the domestic pigeon, domestic duck, king penguin, and snap turtle. The number of times that the species occurred in those particular places was decreased as the randomization of the input order was increased from 5 to 10 times.

PROTDIST – For each dataset, 2 output files were generated. First, the sequences were aligned and a PAM matrix was calculated for them, then the same sequences were used to generate a Kimura matrix. Both results were later used as an input to FITCH and KITSCH. The matrices for each dataset have been included in Section 3 of the Appendix.

For plant dataset1, the distance length for leek was significantly different from all other values in the PAM matrix generated by PROTDIST. The difference was observed in a value approximately 11 times higher than all other distances between all other species – distances for the leek sequence varied between 10.79 and 12.05. In the Kimura matrix, the same trend was observed, although the value for the distances between leek and all other species was set to –1, which was significantly different from all other values, which were positive between 0.00 and 0.12756.

Similarly, for plant dataset2 the distance lengths for Yeast, Yeast Orientalis, and Yeast Occidentalis were significantly different from all other values for all other species. In both PAM and Kimura matrices, the values for the distances between the sequences of the three species were set to –1.

No abnormal (or significantly different) values were observed in the results for either the plant dataset2, when the yeast sequences were removed, or for the animal dataset.

FITCH – For each dataset, 2 output files were generated – for the PAM and Kimura matrices respectively. The trees generated for each dataset have been included in Section 4 of the Appendix.

The trees built using the given matrices generated the following results in terms of the average standard deviation and the sum of squares criterion:

	%SD with PAM matrix input	%SD with Kimura matrix input
Plant dataset1	6.55867	89.43981
Plant dataset2	75.44679	75.44674
Plant dataset2-no yeast included	8.24863	7.89052
Animal dataset	8.77120	8.09748

Figure2a: Results for average % standard deviation for the trees in FITCH generated over all datasets, using both PAM and Kimura matrices as

	Sum of Square with PAM matrix input	Sumof Square with Kimura matrix input
Plant dataset1	0.30111	55.99635
Plant dataset2	73.99883	73.99874
Plant dataset2-no yeast included	0.47628	0.43582
Animal dataset	1.60023	1.36384

Figure2b: Results for sum of squares criterion for the trees in FITCH generated over all datasets, using both PAM and Kimura matrices as input

KITSCH – For each dataset, 2 output files were generated – for PAM and Kimura matrices respectively. The trees generated for each dataset and each type of matrix have been included in Section 5 of the Appendix.

The trees built using the given matrices generated the following results in terms average standard deviation and sum of squares criterion:

	%SD with PAM matrix input	%SD with Kimura matrix input
Plant dataset1	11.23223	-
Plant dataset2	-	-
Plant dataset2-no yeast included	11.91564	8.24863
Animal dataset	14.50394	14.66250

Figure 3a – Results for average % standard deviation in KITSCH for the trees generated over all datasets, using both PAM and Kimura matrices as input

	Sum of Square with PAM matrix input	Sumof Square with Kimura matrix input
Plant dataset1	0.883	-
Plant dataset2	-	-
Plant dataset2-no yeast included	0.994	0.47628
Animal dataset	4.376	4.472

Figure3b – Results for sum of squares criterion in KITSCH for the trees generated over all datasets, using both PAM and Kimura matrices as input

DISCUSSION

PARSIMONY METHODS (PROTRAPS): The program was based on a Felsenstein method, which could be described in the following way: each tree is built by successively adding edges one at a time. First, three sequences are randomly chosen and placed together in an unrooted tree. Then, another sequence is chosen and added to the edge that gives the best score for the scoring position, and the same procedure is continued until the tree is complete and all species have been included. Process continues and each time around local rearrangements are made to see whether any changes will improve on the likelihood of the tree (Felsenstein 1981). Therefore, the method does not guarantee to find the best tree, and adding the sequences in a different order could possibly yield different final trees. This approach depends on the input order of the sequences, so a good test for the validity of the result is its consistency as the input order is changed. If the data were consistent, the same tree would result in each case as the one with the highest likelihood.

For plant dataset1, only the strict consensus trees built from the trees found by PROTRAPS were compared. The tree generated from the 7 best results (with the non-random input order of the sequences) was most likely not the best tree (i.e. the tree with the highest likelihood). The trees generated from the random rearrangements of the data 5 and 10 times were identical to each other, and different from the one generated with non-random input order. Based on the assumption that randomizing the input order of the data is a good way to test the consistency of the results, it is possible that there was another tree with higher likelihood generated from a different input order. It would be established that the tree with the highest likelihood was the one generated from the randomization of the input order, because it remained consistent, independently of the number of rearrangements.

For plant dataset2, the strict consensus trees for all procedures were completely identical and based on Felsenstein's prediction, it could be concluded that the generated tree could reliably be established as the tree with the highest likelihood.

For the animal dataset, the consensus tree for each trial was different. The trees generated after the randomization of the input order of the data, however, were similar in structure, and grouping, although the number of times a particular taxon had occurred in the corresponding places of the tree had decreased as the number of randomizations was increased. It should also be noted that the trend could have been observed due to the increase in the number of different trees used for the generation of the strict consensus trees. The consensus tree for the 5-randomizations of the input order was generated from the 40 trees, while the consensus tree for the 10-randomizations was generated from 50 presumably different trees. This could possibly account for the decrease in the numerical values at each fork. Another reason to discard the result from the non-random input order is that other studies have shown that Aptenodytes patagonicus (King Penguin) and Anas platyrhynchos (Domestic duck) have the shortest distance to each other, than to any other species (Fitch 1976). However, the tree resulting from the non-random input order failed to group them together before grouping them with other species in the dataset.

DISTANCE METHODS (FITCH and KITSCH): Distance methods are used to build phylogenies based on distance measures, calculated on the basis of a number of factors (including transitions and transversions). The reconstruction of the evolutionary trees relies on the additive nature of the data, as well as for some methods on the existence of a molecular evolutionary clock. As it could be summarized (Kitching 1992) an additive tree requires that the following four conditions be satisfied:

1.) no measurable distance between a taxon and itself – d(A,A) = 0
2.) symmetrical distance between taxa – d(A,B) = d(B,A)
3.) negative distances are not allowed – d(A,B)>=0
4.) the triangle inequality between any three taxa in the tree – d(A,B) <= d(A,C) + d(B,C)

If negative distances were to arise, the results would indicate non-additive distances. It should be noted that this fact does not mean that a tree could not be generated. It means that a distance method will fail to construct a correct phylogenetic tree, even if the program itself allows negative distances, because the data will fail to satisfy either the third, or the fourth condition set for additive trees. On another note, the existence of a disproportionately longer branch for some taxon, in comparison to the other species could also not generate a correct phylogenetic tree, due to a violation of the triangle inequality.

FITCH

For plant dataset1, both procedures generated trees that do not represent the correct phylogeny. In the case of the Kimura matrix, the conclusion that the tree is not correct is based on the result of the statistical analysis, which shows that the %SD for the given tree is 89.43981. However, in testing phylogenetic alternatives, one is seeking to minimize the percent standard deviation (Fitch 1967). In the case of the PAM matrix, the conclusion is based on the violation of the triangle inequality for some of the taxa. For example, if the data satisfied all conditions:

d(Leek, Pumpkin) < d(Leek, LoveInMis) + d(Pumpkin, LoveInMis)
12.35564 < 10.79360+0.11726
12.35564 < 10.91086

which is a contradiction. Thus, the triangle inequality was violated and the data could not have generated the correct phylogenetic tree.

Similarly, for plant dataset2, the trees generated using either procedure could not be taken as statistically reliable, because both matrices contain negative values (-1 for all Yeast sequences) and thus violate the triangle inequality, as well as the statistical analysis of both trees indicates %SD that is high enough for both trees to be dismissed as valid phylogenies: the PAM-based tree has %SD of 75.44679, and the Kimura-based tree – 75.44674.

However, when the yeast sequences were removed from the dataset, the newly generated trees by both the PAM and the Kimura matrices were statistically significant - %SD of 8.24863 and 7.89052 respectively, as well as they were almost identical in structure. The difference between the two trees consisted in the placement of the Thermomyces lanuginosus (Humicola •eonine•us) and Fagopyrum esculentum (Common buckwheat) group, which was further distance away from the "root" of the tree (**note**: these trees are unrooted, and the term refers to the chosen outgroup). The PAM-based tree was concluded to be correct, since it was not to be not only statistically significant, but also identical to the one generated by the parsimony method after 10 randomizations of the input order of the data (see Figure 2g in Section 1-Appendix).

For the animal dataset, both of the trees generated from the PAM and the Kimura matrices calculated different distances, where the structure was preserved in both trees and the distances were proportionately different for the species included.

KITSCH (assuming molecular clock)

For plant dataset1, the tree generated by the PAM matrix was not the correct phylogenetic tree, since the triangle inequality was not satisfied (similarly to the FITCH example). KITSCH did not generate a tree for the Kimura matrix at all due to the presence of negative values in the matrix that precluded the generation of a tree.

Similarly, for plant dataset2, due to the presence of negative values in both the PAM and the Kimura matrices, no trees were built. Even if such trees had been built, they would have not been correct and statistically significant, because the triangle inequality would have been violated.

However, the removal of the yeast sequences from the dataset again yielded statistically significant trees in KITSCH (similarly to FITCH). The structure and the groupings were similar to the ones generated by FITCH for both the PAM and Kimura matrices. One way of testing whether the data was truly ultrametric or not was to prove that for any three sequences x,y,z, the distances d(x,y), d(y,z), and d(x,z) were either all equal, or two of them were equal and the third one was smaller. However, based on the above described criterion:

d(Wheat, Buckwheat) = 0.14159
d(Wheat, Cauliflower) = 0.07209
d(Buckwheat, Cauliflower) = 0.10716,

which violates both conditions set by the ultrametric condition. Thus, the data is not ultrametric and the trees were not genuinely correct.

Similarly, for the animal dataset, the trees were shown to be statistically reliable and yet a study has shown that although "although the species, descendant from a common ancestor are equidistant with respect to time, they are not equidistant genetically. For example, there are 7.5 mutations in the descent of primates, and 5.8 mutations in the descent of other mammals, thus indicating that the change in the cytochrome c gene has been much more rapid in the descent of the primates, than in that of other mammals." (Fitch 1967) The data is not ultrametric also because it fails the criterion, as in the above shown example:

d(Domestic duck, Domestic Pigeon) = 0.02944
d(Domestic Duck, Snapping Turtle) = 0.07065
d(Domestic Pigeon, Snapping Turtle) = 0.08133,

which violates both conditions – none of the distances are equal. Therefore, the ultrametric condition is not satisfied and the data is not ultrametric and is not regulated by a molecular evolutionary clock.

Conclusions

Although the study described above could only give a glimpse of the variety of methods and their strengths and weaknesses due to the small number of datasets, and the limited number of procedures performed on the data, there are a number of observations that could be made.

Given the initial data and datasets, the parsimony method in this particular case proved more useful, however, only because of the small number of sequences in each dataset (9, 12, 15 sequences respectively). Felsenstein's maximum likelihood method would not work as well on a larger dataset due to the large number of iterations necessary if the optimality of the current best tree needs to be reassessed with the addition of each new taxon to the tree. Another problem with Felsenstein's method is that it does not guarantee to find the best tree overall, because it is dependent on the sequence input order. However, this problem as mentioned earlier in the paper could be resolved by randomizing the input order of the sequences.

The results of this study also showed some of the weaknesses of the distance methods, used in PHYLIP 3.6. One of the problems with the two distance methods used in the study, FITCH and KITSCH, was that they require that the nature of the data be additive, or additive & ultrametric. Unfortunately, the problem with such an approach is that a large percentage of data is non-additive, and an even larger percent of the data is non-ultrametric. In the PHYLIP software package, that problem has been partially accounted for by restricting the possible input to FITCH and KITSCH. "Although Felsenstein modified FITCH program in the PHYLIP software package to avoid negative branches, Farris' point remains: negative branch lengths identify non-additive data and removing this aspect avoids empirically testing this objective."

(Kitching 1992) Even, if the software were to allow non-additive data, the methods are such that the accuracy of the generated results would be largely decreased, and other methods should be used.

References

2.) Blanken, R. L. (1982). *Comouter Comparison of New and Existing Criteria for Constructing Evolutionary Trees from Sequence Data.*, J Mol Evol. 19: 9-19

3.) Dunn, B. S., (1982). An introduction to mathematical taxonomy, Cambridge University Press

4.) Durbin, R. *et al.*, (1998) Biological sequence analysis, Cambridge University Press

5.) Felsenstein, J. (1981). *Evolutionary trees from DNA sequences: a maximum likelihood approach.*, J Mol Evol.;17(6):368-76.

6.) Fitch, W. M., Margoliash, E. (1967). *Construction of Phylogenetic Trees*; Science 155: 279-284

7.) Fitch, W. M. (1976) *The Molecular Evolution of Cytochrome c in Eukaryotes.*, J Mol Evol., 8, 13-40

8.) Kitching, I. J. *et al.* (1992). Cladistics: A Practical Course in Systematics. Clarendon Press, Oxford

9.) Sourdis, J., Krimbas, C. (1987). *Accuracy of Phylogenetic Trees Estimated from DNA Sequence Data.*, Mol Biol Evol., 4(2): 159-166

10.) Stryer, L. (1988). Biochemistry, 3rd ed., Freeman and Company, New York

10.) Tateno, Y. *et al.*, *Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species.*, J Mol Evol, 18: 387-404

Appendix

Section 1 – Species and datasets

#	Name of species	Tax ID	Swiss-Prot Ascension Number	In dataset #:
1	Indian Mallow		P00059	1
2	Allium porrum (leek)	4681	P00064	1
3	Cucurdita maxima (pumpkin)	3661	P00051	1
4	Gossypium barbadense (Sea-Island Cotton)	3634	P00058	1
5	Zea mays (Maize)	4577	P00056	1
6	Nigella damascena (Love-in-a-mist)	3444	P00066	1
7	Phaseolus aureus (Mung bean)	3916	P00052	1
8	Ricinus communis (Castor bean)	3988	P00057	1
9	Sesamum indicum (Oriental sesame)	4182	P00054	1
10	Tropaeolum majus (Common nasturtium)	4020	P00067	2
11	Triticum aestivum (Wheat)	4565	P00068	2

12	Acer negundo (Box Elder)	4023	P00063	2
13	Brassica oleracea (Cauliflower)	3712	P00050	2
14	Fagopyrum esculentum (Common buckwheat)	3617	P00072	2
15	Guizotia abyssinica (Niger) (Ramtilla)	4230	P00069	2
16	Oryza sativa (Rice)	4530	P00055	2
17	Solanum tuberosum (Potato)	4113	P00061	2
18	Thermomyces lanuginosus (Humicola •eonine•us)	5541	P00047	2
19	Candida Albicans(yeast)	5476	P53698	2
20	Debaryomyces occidentalis	27300	P19681	2
21	Issatchenkia orientalis (Yeast – Candida krusei)	4909	P00041	2
22	Canis Familiaris (Dog)	9615	P00011	3
23	Chelydra serpentina (Snapping turtle)	8475	P00022	3
24	Columbia livia (Domestic pigeon)	8932	P00021	3
25	Anas platyrhynchos (Domestic duck)	8839	P00020	3
26	Equus caballus (Horse)	9796	P00004	3
27	Katsuwonus pelamis (Skipjack tuna) (Bonito)	8226	P00025	3
28	Homo sapiens (Human)	9606	P00001	3
29	Macaca mulatta (Rhesus macaque)	9544	P00002	3
30	Lampetra •eonine•us (Pacific lamprey)	7751	P00028	3
31	Macropus giganteus (Eastern gray kangaroo)	9317	P00014	3
32	Hippopotamus •eonine•us (Hippopotamus)	9833	P00007	3
33	Mirounga •eonine (Southern elephant seal)	9715	P00012	3
34	Oryctolagus cuniculus (Rabbit)	9986	P00008	3
35	Aptenodytes patagonicus (King Penguin)	9234	P00017	3
36	Eschrichtius gibbosus (California Grey Whale)	9764	P00010	3

Figure 1: The 36 species, their corresponding ID numers in the NCBI Taxonomy database, Swiss-Prot ascension numbers and indication of the dataset, to which the sequence belonged for the purpose of the study

Section 2 –PROTRAPS results

1. **Plant Dataset1**

```
           +----LoveInMis
        +--1.0
     +--1.0  +----Maize
     !  !
  +--1.0  +---------MungBean
  !  !
  !  +--------------Pumpkin
  !
+--1.0          +----SeaIslCot
!  !        +--0.7
!  !  +--0.6  +----Leek    ATF
!  !  !  !
!  +--0.6  +---------Indianma
!     !
!     +--------------CastorBea
!
+------------------------Sesame
```

Figure 2a: CONSENSUS tree, generated for plant dataset1 (assuming non-random input order)

```
               +----SeaIslCot
            +--0.3
   +------------1.0   +----Indianma
   !        !
   !        +---------Leek          ATF
   !
   !           +----Pumpkin
+--1.0        +--1.0
!  !     +--1.0   +----MungBean
!  !     !  !
!  !  +--1.0   +---------LoveInMis
!  !  !  !
!  +--1.0   +--------------Sesame
!     !
!     +-------------------Maize
!
+-----------------------------CastorBea
```

Figure 2b: CONSENSUS tree, generated for plant dataset1 (assuming 5 random rearrangements)

```
               +----SeaIslCot
            +--0.3
   +------------1.0   +----Indianma
   !        !
   !        +---------Leek          ATF
   !
   !           +----Pumpkin
+--1.0        +--1.0
!  !     +--1.0   +----MungBean
!  !     !  !
!  !  +--1.0   +---------LoveInMis
!  !  !  !
!  +--1.0   +--------------Sesame
!     !
!     +-------------------Maize
!
+-----------------------------CastorBean
```

Figure 2c: CONSENSUS tree, generated for plant dataset1 (assuming 10 random rearrangements)

2. Plant Dataset2

```
                           +----Potato
                        +--1.0
                     +--1.0   +----Rice
                     !  !
                  +--1.0   +---------BoxElder
                  !  !
                  !  !     +----Wheat
               +--0.5   +-------1.0
               !  !        +----Cauliflow
            +--0.5   !
            !  !  +-------------------Nasturtiu
         +--1.0   !
         !  !  +------------------------Buckwheat
      +--1.0   !
      !  !  +-----------------------------Ramtilla
   +--1.0   !
   !  !  +----------------------------------Humicola
+--1.0   !
!  !  +---------------------------------------Yeast             PA
!  !
!  +--------------------------------------------YeastOcci
!
```

Figure 2d: CONSENSUS tree, generated for plant dataset2 (assuming non-random input order)

```
+----------------------------------------------YeastOrie
```

```
                           +----Potato
                        +--1.0
                     +--1.0  +----Rice
                     !  !
                  +--1.0  +---------BoxElder
                  !  !
                  !  !     +----Wheat
               +--0.5  +-------1.0
               !  !        +----Cauliflow
            +--0.5  !
            !  !  +------------------Nasturtiu
         +--1.0  !
         !  !  +-----------------------Buckwheat
      +--1.0  !
      !  !  +---------------------------Ramtilla
   +--1.0  !
   !  !  +--------------------------------Humicola
+--1.0  !
!  !  +-------------------------------------Yeast              PA
!  !
!  +-----------------------------------------YeastOcci
!
+----------------------------------------------YeastOrie
```

Figure 2e: CONSENSUS tree, generated for plant dataset2 (assuming 5 random rearrangements)

```
                           +----Potato
                        +--1.0
                     +--1.0  +----Rice
                     !  !
                  +--1.0  +---------BoxElder
                  !  !
                  !  !     +----Wheat
               +--0.5  +-------1.0
               !  !        +----Cauliflow
            +--0.5  !
            !  !  +------------------Nasturtiu
         +--1.0  !
         !  !  +-----------------------Buckwheat
      +--1.0  !
      !  !  +---------------------------Ramtilla
   +--1.0  !
   !  !  +--------------------------------Humicola
+--1.0  !
!  !  +-------------------------------------Yeast              PA
!  !
!  +-----------------------------------------YeastOcci
```

Figure 2f: CONSENSUS tree, generated for plant dataset2 (assuming 10 random rearrangements)

3. Plant Dataset2 – no yeast sequence included

```
               +----Humicola
      +------------1.0
      !        +----Buckwheat
      !
   +--0.8           +----Cauliflow
   !  !  +-------1.0
   !  !  !     +----Wheat
   !  +--0.5
   !     !  +---------BoxElder
+--1.0      +--0.8
!  !        !  +----Rice
!  !        +--1.0
```

Figure 2g: CONSENSUS tree, generated for plant dataset2 without yeast sequences (assuming 10 random rearrangements)

```
!  !             +----Potato
!  !
!  +------------------------Nasturtiu
!
+-----------------------------Ramtilla
```

4. Animal Dataset

```
                               +----DomDuck
                        +--0.6
                     +--1.0  +----KingPengu
                     !  !
                  +--1.0  +---------SnapTurtl
                  !  !
               +--0.4  +--------------DomPigeon
               !  !
            +--1.0  +------------------Rabbit
            !  !
            !  +------------------------CGrayWhal
         +--1.0
         !  !                +----Human
         !  !             +--1.0
      +--0.5  +-----------------0.5  +----Monkey
      !  !                !
      !  !                +---------EastGreyK
      !  !
   +--1.0  +----------------------------------Hippopota
   !  !
   !  !                      +----PacifLamp
   !  !                   +--1.0
+--1.0  +---------------------------1.0  +----Tuna
!  !                      !
!  !                      +---------Horse
!  !
!  +--------------------------------------------Dog
!
+------------------------------------------------ElephSeal
```

Figure 2h: CONSENSUS tree, generated for animal dataset (assuming non-random input order)

```
                     +---------SnapTurtl
                  +--0.7
                  !  !  +----DomDuck
               +--1.0  +--0.7
               !  !     +----KingPengu
               !  !
            +--0.7  +--------------DomPigeon
            !  !
            !  !        +----Rabbit
            !  +------------0.8
         +--0.7             +----CGrayWhal
         !  !
         !  !           +----Monkey
         !  !        +--1.0
      +--0.3  +------------0.3  +----Human
      !  !           !
      !  !           +---------EastGreyK
      !  !
   +--1.0  +-----------------------------Hippopota
   !  !
   !  !                 +----PacifLamp
   !  !              +--1.0
+--1.0  +----------------------1.0  +----Tuna
!  !                 !
!  !                 +---------Horse
```

Figure 2i: CONSENSUS tree, generated for animal dataset (assuming 5 random rearrangements)

```
!  !
!  +--------------------------------Dog
!
+------------------------------------ElephSeal

                  +---------SnapTurtl
               +--0.6
               !  !  +----DomDuck
            +--1.0  +--0.5
            !  !     +----KingPengu
            !  !
         +--0.7  +--------------DomPigeon
         !  !
         !  !        +----CGrayWhal
         !  +------------0.8
      +--0.6           +----Rabbit
      !  !
      !  !           +----Monkey
      !  !        +--1.0
      !  +------------0.4  +----Human
      !           !
   +--1.0           +---------EastGreyK
   !  !
   !  !              +----Tuna
   !  !           +--1.0
   !  !        +--1.0  +----PacifLamp
+--1.0   !        !  !
!  !  +------------0.3  +---------Horse
!  !           !
!  !           +--------------Hippopota
!  !
!  +----------------------------Dog
!
+--------------------------------ElephSeal
```

Figure 2j: CONSENSUS tree, generated for animal dataset (assuming 10 random rearrangements)

Section 3 – PROTDIST results

1. Plant Dataset1

a) PAM matrix

```
Indianma     0.00000 11.15018  0.07205  0.02972  0.07371  0.16373  0.09240  0.04044  0.06058
Leek  ATF   11.15018  0.00000 12.35564 11.13064 12.05775 10.79360 11.77498 11.41536 11.77662
Pumpkin      0.07205 12.35564  0.00000  0.07278  0.08152  0.11726  0.03939  0.07352  0.06211
SeaIslCot    0.02972 11.13064  0.07278  0.00000  0.07260  0.15459  0.07198  0.02978  0.06118
Maize        0.07371 12.05775  0.08152  0.07260  0.00000  0.13272  0.09173  0.06188  0.08233
LoveInMis    0.16373 10.79360  0.11726  0.15459  0.13272  0.00000  0.10552  0.15577  0.13246
MungBean     0.09240 11.77498  0.03939  0.07198  0.09173  0.10552  0.00000  0.07271  0.06134
CastorBea    0.04044 11.41536  0.07352  0.02978  0.06188  0.15577  0.07271  0.00000  0.05020
Sesame       0.06058 11.77662  0.06211  0.06118  0.08233  0.13246  0.06134  0.05020  0.00000
```

b) Kimura matrix

```
Indianma     0.00000 -1.00000  0.07211  0.03003  0.07363  0.12756  0.09408  0.03033  0.05129
Leek  ATF   -1.00000  0.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000
Pumpkin      0.07211 -1.00000  0.00000  0.07211  0.08301  0.10267  0.03953  0.07286  0.06199
SeaIslCot    0.03003 -1.00000  0.07211  0.00000  0.07363  0.12756  0.07211  0.02008  0.05129
Maize        0.07363 -1.00000  0.08301  0.07363  0.00000  0.10507  0.09408  0.05236  0.07441
LoveInMis    0.12756 -1.00000  0.10267  0.12756  0.10507  0.00000  0.09053  0.12756  0.12756
MungBean     0.09408 -1.00000  0.03953  0.07211  0.09408  0.09053  0.00000  0.07286  0.06199
CastorBea    0.03033 -1.00000  0.07286  0.02008  0.05236  0.12756  0.07286  0.00000  0.05129
Sesame       0.05129 -1.00000  0.06199  0.05129  0.07441  0.12756  0.06199  0.05129  0.00000
```

2. Plant Dataset2

a) PAM matrix

```
Nasturtiu    0.00000  0.11966  0.11814  0.08284  0.14667  0.08249  0.13605  0.13944  0.69085 -1.00000 -1.00000 -1.00000
Wheat        0.11966  0.00000  0.09406  0.07209  0.14159  0.12606  0.10290  0.11773  0.62028 -1.00000 -1.00000 -1.00000
BoxElder     0.11814  0.09406  0.00000  0.08351  0.12669  0.11628  0.09237  0.07173  0.65318 -1.00000 -1.00000 -1.00000
Cauliflow    0.08284  0.07209  0.08351  0.00000  0.10716  0.09151  0.09114  0.09349  0.66222 -1.00000 -1.00000 -1.00000
Buckwheat    0.14667  0.14159  0.12669  0.10716  0.00000  0.10412  0.12488  0.10386  0.59086 -1.00000 -1.00000 -1.00000
Ramtilla     0.08249  0.12606  0.11628  0.09151  0.10412  0.00000  0.11363  0.11417  0.63875 -1.00000 -1.00000 -1.00000
Rice         0.13605  0.10290  0.09237  0.09114  0.12488  0.11363  0.00000  0.07019  0.65869 -1.00000 -1.00000 -1.00000
Potato       0.13944  0.11773  0.07173  0.09349  0.10386  0.11417  0.07019  0.00000  0.60432 -1.00000 -1.00000 -1.00000
Humicola     0.69085  0.62028  0.65318  0.66222  0.59086  0.63875  0.65869  0.60432  0.00000 -1.00000 -1.00000 -1.00000
Yeast  PA   -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000  0.00000 -1.00000 -1.00000
YeastOcci   -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000  0.00000  0.20359
YeastOrie   -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000  0.20359  0.00000
```

b) Kimura matrix

```
Nasturtiu    0.00000  0.11672  0.09408  0.08301  0.14007  0.05348  0.14007  0.12831  0.67660 -1.00000 -1.00000 -1.00000
Wheat        0.11672  0.00000  0.09216  0.07065  0.13711  0.11118  0.10314  0.11430  0.61411 -1.00000 -1.00000 -1.00000
BoxElder     0.09408   0.09216   0.00000   0.08133   0.12562   0.07602   0.09216   0.07065   0.67879  -1.00000  -1.00000  -1.00000
Cauliflow    0.08301  0.07065  0.08133  0.00000  0.10314  0.06467  0.09216  0.09216  0.65669 -1.00000 -1.00000 -1.00000
Buckwheat    0.14007  0.13711  0.12562  0.10314  0.00000  0.09927  0.12562  0.10314  0.59358 -1.00000 -1.00000 -1.00000
Ramtilla     0.05348  0.11118  0.07602  0.06467  0.09927  0.00000  0.11118  0.09927  0.63580 -1.00000 -1.00000 -1.00000
Rice         0.14007  0.10314  0.09216  0.09216  0.12562  0.11118  0.00000  0.07065  0.67879 -1.00000 -1.00000 -1.00000
Potato       0.12831  0.11430  0.07065  0.09216  0.10314  0.09927  0.07065  0.00000  0.63514 -1.00000 -1.00000 -1.00000
Humicola     0.67660  0.61411  0.67879  0.65669  0.59358  0.63580  0.67879  0.63514  0.00000 -1.00000 -1.00000 -1.00000
Yeast     PA -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000  0.00000 -1.00000 -1.00000
YeastOcci   -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000  0.00000  0.19732
YeastOrie   -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000 -1.00000  0.19732  0.00000
```

3. Plant Dataset2 – yeast sequences not included

a) PAM matrix

```
Nasturtiu    0.00000  0.11966  0.11814  0.08284  0.14667  0.08249  0.13605  0.13944  0.69085
Wheat        0.11966  0.00000  0.09406  0.07209  0.14159  0.12606  0.10290  0.11773  0.62028
BoxElder     0.11814  0.09406  0.00000  0.08351  0.12669  0.11628  0.09237  0.07173  0.65318
Cauliflow    0.08284  0.07209  0.08351  0.00000  0.10716  0.09151  0.09114  0.09349  0.66222
Buckwheat    0.14667  0.14159  0.12669  0.10716  0.00000  0.10412  0.12488  0.10386  0.59086
Ramtilla     0.08249  0.12606  0.11628  0.09151  0.10412  0.00000  0.11363  0.11417  0.63875
Rice         0.13605  0.10290  0.09237  0.09114  0.12488  0.11363  0.00000  0.07019  0.65869
Potato       0.13944  0.11773  0.07173  0.09349  0.10386  0.11417  0.07019  0.00000  0.60432
Humicola     0.69085  0.62028  0.65318  0.66222  0.59086  0.63875  0.65869  0.60432  0.00000
```

b) Kimura matrix

```
Nasturtiu   0.00000  0.11672  0.09408  0.08301  0.14007  0.05348  0.14007  0.12831  0.67660
Wheat       0.11672  0.00000  0.09216  0.07065  0.13711  0.11118  0.10314  0.11430  0.61411
BoxElder    0.09408  0.09216  0.00000  0.08133  0.12562  0.07602  0.09216  0.07065  0.67879
Cauliflow   0.08301  0.07065  0.08133  0.00000  0.10314  0.06467  0.09216  0.09216  0.65669
Buckwheat   0.14007  0.13711  0.12562  0.10314  0.00000  0.09927  0.12562  0.10314  0.59358
Ramtilla    0.05348  0.11118  0.07602  0.06467  0.09927  0.00000  0.11118  0.09927  0.63580
Rice        0.14007  0.10314  0.09216  0.09216  0.12562  0.11118  0.00000  0.07065  0.67879
Potato      0.12831  0.11430  0.07065  0.09216  0.10314  0.09927  0.07065  0.00000  0.63514
Humicola    0.67660  0.61411  0.67879  0.65669  0.59358  0.63580  0.67879  0.63514  0.00000
```

4. Animal dataset

a) PAM matrix

```
Dog         0.00000  0.09445  0.09553  0.08553  0.06332  0.19305  0.01005  0.16062  0.03047  0.10561  0.04052  0.07135  0.05108  0.11599  0.10476
SnapTurtl   0.09445  0.00000  0.08131  0.06997  0.11627  0.18691  0.09471  0.21455  0.08285  0.07962  0.09363  0.11463  0.09370  0.16005  0.14839
DomPigeon   0.09553  0.08131  0.00000  0.02021  0.11765  0.19732  0.09634  0.21833  0.08382  0.01952  0.08333  0.11614  0.07339  0.12857  0.11708
DomDuck     0.08553  0.06997  0.02021  0.00000  0.10748  0.18589  0.08577  0.20821  0.07386  0.01831  0.08479  0.10628  0.06367  0.11962  0.10804
Horse       0.06332  0.11627  0.11765  0.10748  0.00000  0.20358  0.07430  0.18074  0.05267  0.12741  0.06301  0.07332  0.06242  0.12928  0.11775
Tuna        0.19305  0.18691  0.19732  0.18589  0.20358  0.00000  0.19520  0.20424  0.17918  0.19589  0.19171  0.19753  0.18787  0.23841  0.23966
ElephSeal   0.01005  0.09471  0.09634  0.08577  0.07430  0.19520  0.00000  0.17312  0.04087  0.10590  0.04086  0.08205  0.06167  0.12735  0.11596
PacifLamp   0.16062  0.21455  0.21833  0.20821  0.18074  0.20424  0.17312  0.00000  0.16982  0.21854  0.18061  0.18634  0.19349  0.22970  0.23091
CGrayWhal   0.03047  0.08285  0.08382  0.07386  0.05267  0.17918  0.04087  0.16982  0.00000  0.09380  0.02135  0.06101  0.00940  0.10517  0.09405
KingPengu   0.10561  0.07962  0.01952  0.01831  0.12741  0.19589  0.10590  0.21854  0.09380  0.00000  0.10468  0.10397  0.08383  0.13966  0.12807
Hippopota   0.04052  0.09363  0.08333  0.08479  0.06301  0.19171  0.04086  0.18061  0.02135  0.10468  0.00000  0.07101  0.05084  0.11541  0.10424
EastGreyK   0.07135  0.11463  0.11614  0.10628  0.07332  0.19753  0.08205  0.18634  0.06101  0.10397  0.07101  0.00000  0.06108  0.10374  0.11482
Rabbit      0.05108  0.09370  0.07339  0.06367  0.06242  0.18787  0.06167  0.19349  0.00940  0.08383  0.05084  0.06108  0.00000  0.09323  0.08234
Human       0.11599  0.16005  0.12857  0.11962  0.12928  0.23841  0.12735  0.22970  0.10517  0.13966  0.11541  0.10374  0.09323  0.00000  0.00991
Monkey      0.10476  0.14839  0.11708  0.10804  0.11775  0.23966  0.11596  0.23091  0.09405  0.12807  0.10424  0.11482  0.08234  0.00991  0.00000
```

b) Kimura matrix

```
Dog         0.00000  0.09216  0.09216  0.08133  0.06013  0.17455  0.00968  0.14878  0.02944  0.10314  0.03953  0.07065  0.04976  0.11430  0.10314
SnapTurtl   0.09216  0.00000  0.08133  0.07065  0.11430  0.17455  0.09216  0.20994  0.08133  0.08133  0.09216  0.11430  0.09216  0.16063  0.14878
DomPigeon   0.09216  0.08133  0.00000  0.02944  0.11430  0.18693  0.09216  0.20994  0.08133  0.03953  0.08133  0.11430  0.07065  0.12562  0.11430
DomDuck     0.08133  0.07065  0.02944  0.00000  0.10314  0.17455  0.08133  0.19732  0.07065  0.02944  0.08133  0.10314  0.06013  0.11430  0.10314
Horse       0.06013  0.11430  0.11430  0.10314  0.00000  0.18693  0.07065  0.17266  0.04976  0.12562  0.06013  0.07065  0.06013  0.12562  0.11430
Tuna        0.17455  0.17455  0.18693  0.17455  0.18693  0.00000  0.17455  0.19951  0.16237  0.18693  0.17455  0.18693  0.17455  0.22529  0.22529
ElephSeal   0.00968  0.09216  0.09216  0.08133  0.07065  0.17455  0.00000  0.16063  0.03953  0.10314  0.03953  0.08133  0.06013  0.12562  0.11430
PacifLamp   0.14878  0.20994  0.20994  0.19732  0.17266  0.19951  0.16063  0.00000  0.16063  0.20994  0.17266  0.18489  0.18489  0.22277  0.22277
CGrayWhal   0.02944  0.08133  0.08133  0.07065  0.04976  0.16237  0.03953  0.16063  0.00000  0.09216  0.02944  0.06013  0.01949  0.10314  0.09216
KingPengu   0.10314  0.08133  0.03953  0.02944  0.12562  0.18693  0.10314  0.20994  0.09216  0.00000  0.10314  0.10314  0.08133  0.13711  0.12562
Hippopota   0.03953  0.09216  0.08133  0.08133  0.06013  0.17455  0.03953  0.17266  0.02944  0.10314  0.00000  0.07065  0.04976  0.11430  0.10314
EastGreyK   0.07065  0.11430  0.11430  0.10314  0.07065  0.18693  0.08133  0.18489  0.06013  0.10314  0.07065  0.00000  0.06013  0.10314  0.11430
Rabbit      0.04976  0.09216  0.07065  0.06013  0.06013  0.17455  0.06013  0.18489  0.01949  0.08133  0.04976  0.06013  0.00000  0.09216  0.08133
Human       0.11430  0.16063  0.12562  0.11430  0.12562  0.22529  0.12562  0.22277  0.10314  0.13711  0.11430  0.10314  0.09216  0.00000  0.00968
Monkey      0.10314  0.14878  0.11430  0.10314  0.11430  0.22529  0.11430  0.22277  0.09216  0.12562  0.10314  0.11430  0.08133  0.00968  0.00000
```

Section 4 – FITCH results

1. Plant Dataset1

a. PAM matrix

```
   +SeaIslCot
   !
+--1  +CastorBea
!  !  !
!  +--6  +Maize
!     !  !
!     +--3  +Sesame
!        !  !
```

```
  !     +--7    +MungBean
  !       !  +--5
  !       +--4  +LoveInMis
  !          !
  !          +Pumpkin
  !
--2-----------------------------------------------------Leek         ATF
  !
  +Indianma
```

b. Kimura matrix

```
  +Sesame
  !
  ! +MungBean
  ! !
--7--5  +CastorBea
  ! ! !
  ! ! !     +SeaIslCot
  ! +--6    +--1
  !    ! +--4  +LoveInMis
  !    ! ! !
  !    +--2  +Leek              ATF
  !      !
  !      ! +Maize
  !      +--3
  !        +Pumpkin
  !
  +Indianma
```

2. Plant Dataset2

a. PAM matrix

```
    +Wheat
    !
    !   +Buckwheat
    !   !
    !   !         +Cauliflow
  +--8  !         +--7
  ! !   !      +-10  +YeastOrie
  ! ! +--3     ! !
  ! ! ! !   +--6  +YeastOcci
  ! ! ! !   ! !
  ! ! ! ! +--9  +BoxElder
  ! +--5 ! ! !
  !   ! +--4  +Ramtilla
  !   !   !
  !   !   +Rice
  !   !
  !   +Humicola
  !
  ! +Yeast          PA
--2--1
```

b. Kimura matrix

```
    +Rice
    !
```

```
  +--3  +Buckwheat
  ! ! !
  ! ! !   +Wheat
  ! +--7   !
  !   ! +--2   +Ramtilla
  !   ! ! ! +--4
  !   ! ! +--8  +Yeast          PA
  !   +-10   !
  !     !   +Potato
  !     !
  !     +Humicola
  !
  !     +YeastOrie
  !   +--9
  ! +--1  +Cauliflow
  ! ! !
--6--5  +YeastOcci
  ! !
  ! +BoxElder
  !
  +Nasturtiu
```

3. Plant Dataset2 – no yeast sequences included

a. PAM matrix

```
  +-Ramtilla
  !
  !     +--Wheat
  !   +--3
  !   ! +Cauliflow
  ! +--1
  ! ! !   +-Potato
  ! ! ! +--6
  ! ! +--5  +-Rice
--4--2    !
  ! !   +-BoxElder
  ! !
  ! ! +--------------------------------Humicola
  ! +--7
  !   +-Buckwheat
  !
  +--Nasturtiu
```

b. Kimura matrix

```
  +Ramtilla
  !
  !       +-Potato
  !     +--6
  !   +--5  +-Rice
  !   ! !
  ! +--3  +-BoxElder
  ! ! !
  ! ! ! +---------------------------------Humicola
--4--1  +--7
  ! !   +-Buckwheat
  ! !
  ! ! +Cauliflow
  ! +--2
  !   +--Wheat
  !
  +-Nasturtiu
```

4. Animal Dataset

a. PAM matrix

```
  +ElephSeal
  !
  ! +Hippopota
  ! !
  ! !   +Rabbit
  ! ! +-11
--5--3  ! +CGrayWhal
  ! ! !
  ! ! !     +Monkey
  ! ! !   +--13
  ! ! !   ! +Human
  ! +--9    !
  !   ! +--7    +-----PacifLamp
  !   ! ! ! +--6
  !   ! ! ! ! +-----Tuna
  !   ! ! +--4
  !   ! !   ! +--SnapTurtl
  !   ! !   +--1
  !   +-10      ! +DomDuck
  !     !     +--2
  !     !       ! +KingPengu
  !     !       +--8
  !     !         +DomPigeon
  !     !
  !     ! +-EastGreyK
  !     +-12
  !       +-Horse
  !
  +Dog
```

b. Kimura matrix

```
  +ElephSeal
  !
  ! +Hippopota
  ! !
  ! ! +CGrayWhal
--5--3  !
  ! ! !   +-EastGreyK
  ! ! ! +-10
  ! ! ! ! +-Horse
  ! +--9  !
  !   ! !       +--SnapTurtl
  !   ! !       !
  !   ! !     +--1    +DomDuck
  !   ! !     ! ! +--2
  !   +-11      ! +--8  +-KingPengu
  !     !   +--4    !
  !     !   ! !   +DomPigeon
  !     !   ! !
  !     !   ! !  +Monkey
  !     ! +-12  +--13
  !     ! ! !    +Human
  !     ! ! !
  !     +--7  ! +-----PacifLamp
  !       ! +--6
  !       !   +-----Tuna
  !       !
  !       +Rabbit
  !
  +Dog
```

Section 5 – KITSCH results

1. Plant Dataset1

a. PAM matrix

```
                                                   +SeaIslCot
                                                 +--3
                                               +--7  +Indianma
                                               !  !
                                             +--8  +CastorBea
                                             !  !
                                           +--2  +Sesame
                                           !  !
                                           !  !  +MungBean
                                         +--4  +--6
                                         !  !    +Pumpkin
  +---------------------------------------------------5  !
  !                                      !  +Maize
--1                                       !
  !                                      +LoveInMis
  !
  +----------------------------------------------------Leek           ATF
```

b. Kimura matrix

No tree generated

2. Plant Dataset2

a. PAM matrix

No tree generated

b. Kimura matrix

No tree generated

3. Plant Dataset2 – yeast sequences not included

a. PAM matrix

```
    +-Ramtilla
  +--5
  ! +-Nasturtiu
  !
+--1      +-Potato
! !    +--7
! ! +--6 +-Rice
! ! ! !
```

```
              !  +--2  +-BoxElder
  +-------------4     !
  !           !    !  +-Cauliflow
  !           !    +--3
--8            !      +-Wheat
  !           !
  !           +---Buckwheat
  !
  +-----------------Humicola
```

b. Kimura matrix

```
                     +-Potato
                   +--1
                 +--7  +-Rice
                 !  !
                 !  +-BoxElder
                 !
               +--6       +-Ramtilla
               !  !    +--5
               !  !  +--2  +-Nasturtiu
               !  !  !  !
  +---------------4  +--3  +-Cauliflow
  !            !     !
  !            !     +--Wheat
--8             !
  !            +--Buckwheat
  !
  +-----------------Humicola
```

4. Animal Dataset

a. PAM matrix

```
       +Monkey
     +-14
     !  +Human
     !
     !           +Rabbit
     !         +-12
     !       +-10  +CGrayWhal
     !       !  !
     !     +--8  +Hippopota
   +-13      !  !
   !  !     !  !  +ElephSeal
   !  !   +--4  +--6
   !  !   !  !    +Dog
   !  !  +-11  !
   !  !  !  !  +-Horse
   !  !  !  !
   !  !  !  +-EastGreyK
  +--7  +--1
  !  !    !     +KingPengu
  !  !    !   +--9
  !  !    !  +--3  +DomDuck
  !  !    !  !  !
--5  !    +--2  +DomPigeon
  !  !       !
  !  !       +-SnapTurtl
  !  !
  !  +-----PacifLamp
  !
  +-----Tuna
```

b. Kimura matrix

```
       +Monkey
     +-14
     !  +Human
```

```
        !
        !                 +Rabbit
        !              +-12
        !           +-10  +CGrayWhal
        !           !  !
        !        +--8  +Hippopota
     +-13        !  !
     !  !        !  !  +ElephSeal
     !  !     +--4  +--6
     !  !     !  !     +Dog
     !  !  +-11  !
     !  !  !  !  +-Horse
     !  !  !  !
     !  !  !  +-EastGreyK
  +--5  +--1
  !  !     !        +DomDuck
  !  !     !     +--3
  !  !     !  +--9  +DomPigeon
  !  !     !  !  !
--7  !     +--2  +KingPengu
  !  !        !
  !  !        +-SnapTurtl
  !  !
  !  +----Tuna
  !
  +-----PacifLamp
```