Towards a Statistical Species Concept and Speciation

A thesis submitted towards partial fulfilment of

BS MS Dual Degree programme

By

Suraj Chawla

(Roll No: 20091104)

Under the guidance of

Prof. Milind Gajanan Watve

(Professor, Department of Biology, IISER Pune)

Indian Institute of Science Education and Research

Pune

Certificate

This is to certify that this dissertation entitled "**Towards a Statistical Species Concept and Speciation"** towards the partial fulfilment of the BS-MS duel degree programme at the Indian Institute of Science Education and Research (IISER) Pune, represents original research carried out by Suraj Chawla at IISER Pune under the supervision of Prof. Milind Gajanan Watve, Associate Professor, Department of Biology, IISER Pune during the academic year 2013-2014.

Suraj Chawla (Roll no: 20091104)

Prof. Milind Gajanan Watve	
(Supervisor)	Dean (Biological sciences)
Date:	Date:
Place:	Place:

Declaration

I hereby declare that the matter embodied in the report entitled “Towards a Statistical Species Concept and Speciation” are the results of the investigations carried out by me at the Department of Biology, Indian Institute of Science Education and Research (IISER) Pune, under the supervision of Prof. Milind Gajanan Watve, Associate Professor, IISER Pune and the same has not been submitted elsewhere for any other degree.

Suraj Chawla (Roll no: 20091104)

Date:

Place:

Contents

Abstract

Defining species has been a central unsolved problem in biology. Modern taxonomy uses objective methods to reconstruct phylogenetic trees but there are no objective methods for delimitation of species as well as for higher taxonomic units. These algorithms do not make any null hypothesis and therefore do not show whether significant clusters exist or whether hierarchical classification exists. We propose a Statistical Species Concept (SSC) based on the Frequency Distribution plot (FDp) of the distances between individuals, which can reveal clusters by segregating the distribution of within and between cluster distances. This algorithm, based on SSC, is then tested with synthetic sequences and compared with the molecular phylogenetic analysis by Maximum Likelihood (ML) method. We find that out of the 100 cases we tested for, SSC predicted the correct hierarchy 94 times while ML gave satisfactory results only 31 times. We also tested the SSC algorithm on real genetic data to compare the predictions of SSC with existing taxonomic ranks of those individuals. We found that the species level of classification corresponded to first level of clustering in the FDp as was predicted based on SSC. We also found clustering that corresponded to the genera level of classification. Finally we develop a platform of computational models to understand speciation and use the objective criteria developed for SSC. We find interesting preliminary results; i) Neutral drift can cause speciation for asexual populations under specific circumstances ii) Patch-dynamics does not facilitate speciation and iii) Competition adds substantial stability to speciation.

List of Figures:

Acknowledgement

If I were to thank Watve Sir in a statistically proportionate way, then I would surely run out of pages. He has been like one of those Captains at the sea, the one who could understand his ship (me), was deeply aware of the beauties and the dangers of the sea (life) and never lost track of the direction in which the ship was headed for.

I would also like to thank my physics mentor Prof. Varun Sahni from IUCAA, if it would not have been for his motivation, I would not have taken this decision of doing my fifth year project in biology. I thank my teachers who have been like friends to me; Prof. L.S. Shashidhara for allowing me to make this big switch from physics to biology, K P Mohanan for the invaluable discussions we had, Prof. K Thangaraj and Dr. Neelesh Dhanukar for the data that they have so kindly provided.

I thank IISER Pune for providing me with all the facilities; I love every changing aspect of it; the labs, the classrooms, the library, the canteen and the people. I can not make a justified thanks to my friends who were like teachers to me, criticizing and supporting all the time; Adwiteey, Suryesh, Pramod, Shubhankar and Manwa.

Finally to mention those, who have always been there to support me in all of my decisions, I thank Mom and Dad.

Suraj Chawla

Introduction

Taxonomy in biology is over 300 years old and the concept of species perhaps much older. People have been identifying and naming species much before the emergence of biology as science. The Linnaean system consisted of assigning every organism to a particular taxon and a corresponding rank for that taxon in a hierarchical taxonomic structure. The concepts of species and a hierarchy of all the higher ranks originated much before the concept of evolution.Therefore the classical species concept was static, while evolution is a dynamic process. The concept of species is so deep rooted in the human mind that it influenced the thinking in evolution. In reality the principles of evolution do not need the concept of species but species is so deeply rooted in the minds of biologists that evolution was depicted as the 'origin of species'. Is species a fundamental unit of evolution? Is evolution possible without species? There is no apparent reason why it should not be. The fundamental processes that drive evolution namely inheritance, variability, mutations, selection and drift are not dependent on whether or not there are species. In that case why do species exist? Do they really exist or is it a construction of the human mind?

For a fundamental level understanding we need to make a fresh enquiry with minimum assumptions or preconceptions. We can start with two most fundamental question of taxonomy as to whether natural clusters exist and whether they are naturally arranged in a hierarchical manner. At the first level we need to have methods to answer the questions. While it is possible to make some classification out of a continuum such as demarcating states or districts in continuous land, giving different names to different districts is not the same as giving names to islands in an archipelago. While islands can be said to be naturally segregated, district boundaries in a continuous land is a man-made division. An intermediate and more complex case is that of peaks in a mountain range. Although peaks are not as clearly separated as islands by distinct gaps, they certainly have natural existence. In the case of peaks although demarcations of where one peak ends and another begins cannot be drawn, the main body of the peak can be distinctly made out with little subjectivity. Do species exist as distinct as islands, or have peak like identities or are man-made demarcations like districts? We need to have methods to answer this question first and that is the first objective of this study. We would than look at the second question, if species are natural clusters, why and how do they arise?

The traditional approach in taxonomy is of making and labelling boxes and boxes within. Such as a phylum is a distinct box which contain several classes which contain orders and so on. This assumes and uses distinct clustering where a box needs to have clear demarcations. The more recent computational approaches, although popularly called 'clustering algorithms' are in reality some or the other form of 'joining' algorithms. These algorithms join individual units at different levels but do not make demarcated clusters themselves. The user may identify some clusters using arbitrary cut offs. Traditional approach is based on learned but subjective judgments of taxonomists, wherein no quantification or measurements are necessary. However, traditional approach is more ‘human friendly’ i.e. easy to name the boxes and thereby remember them and the ones within and develop mental perspectives. Modern approach makes use of objective/quantitative information, wherein dependence on human judgment is not eliminated but may be reduced. The key problem associated with the modern approach is that it creates complex trees which are more difficult to grasp, remember and comprehend i.e. less human friendly. It does not claim to identify boxes/levels of classification; in contrast to the traditional approach which makes an implicit assumption of underlying hierarchy.

It would be a dream to amalgamate the good points of both the approaches into an ideal method that is both objective as well as ‘human-friendly’. If we can have methods that are objective, quantifiable and computational, that give us a simple hierarchical structure comparable to the species, genus, family, order, class etc, we can have a clear cut sound and scientific taxonomy. In order to do so we need to start from the same question whether the boxes are *natural*? Or in other words can clustering exist at different levels so as to justify the purpose of use of 'boxes within boxes' ? Is there an objective way of identifying natural clusters and natural hierarchies? Is taxonomy naturally hierarchical? Can the levels be objectively identified? Both traditional and computational approaches have not answered (or even asked) the question whether there are natural clusters or whether clustering is imposed on the data. Also given natural clustering, is there clustering at different levels or in other words is a hierarchy of clustering is not a question systematically addressed so far. Both traditional and computational approaches assume ‘species’ and assume that there would be hierarchy and attempt to identify it.

The question of identification of natural clusters and significant hierarchy will be explored first in this study. Once the algorithm to detect natural clusters and hierarchy of clustering is ready, it will be used extensively in the second part of the project, primarily, for the objective quantification of the process of speciation.

The problem of choosing the characters for classification:
Inevitable for any classification scheme is some way to characterize the units to be classified. Examining the classical approach we get into a hen and egg problem. Mammals are defined by a handful of characters such as vivipary, mammary glands, hair on body etc. At the same time a number of other characters seem to be ignored or given minor importance in classifying such colour of hair. Were the groups conceived first or the characters? If groups came first, how were they perceived? If characters came first, why these specific characters were selected to define a group?
Is there an algorithmic/statistical way to do both starting from scratch from a set of raw data on all perceivable characters? That is can we assign differential importance to characters without compromising on objectivity. This is an important question since if the answer is yes, we can objectivise entire classification, if not it will remain subjective all the time and we will have to conclude that objective taxonomy is impossible in principle. We will show later that although it sounds impossible to begin with, there can be a possible solution.

Human perception and technology limit the set of characters available to us. This has changed in the history of taxonomy and will keep on changing in future too. For example, earlier classifications were based on visible external morphology which was later aided by anatomy and microanatomy, biochemistry, molecular biology and genomics step by step. Interestingly with greater access to data, although there have significant rearrangements in some groups, classification in several groups has remained surprisingly robust.

To be on the right path, we first need to have a set of principles and methods of classifying that can work with and have flexibility to incorporate any kind of characterization data. The methods should make robust groups and identifying at the same time the set of characters that are important for classification and the ones that are not.

CHAPTER 1: The Statistical Species Concept

1.1 Problems with the Existing Concepts:

Today the literature boasts of many species concepts, some of the most widely known are the Biological Species Concept (BSC), Phylogenetic Species Concept (PSC), Genotypic Cluster Definition (GCD) and Differential Species Concepts (DFSC). The interested reader may look into the following reviews (Mayden 1997; de Queiroz 2007; Coyne and Orr, 2004).Here we show some of the problems with Biological Species Concept (Mayr, 1942).

Biological Species Concept (BSC) states, “species are actually or potentially interbreeding individuals that are reproductively isolated from other such groups” .Some of the major problems with BSC are as follows;

- Is not applicable to asexual organisms that constitute majority of life's diversity.
- 10-30% plant and animal species are known to hybridise and exchange genes with others regularly (Abbott et al., 2013). Also, a very recent study on butterflies of genus heliconius shows gene flow across genomes continues to occur during speciation, which suggests that species can continue to diverge even when there is gene flow. (Martin et al., 2013)
- Another example would be the eutrophication of Lake Victoria that appears to break down sexual isolation of the cichlid colour morphs due to poorer light conditions. Exact enunciation of the conditions under which the populations are reproductively isolated, is difficult; is natural habitat sufficient? why not artificial habitats? It is impossible to answer these questions without putting subjectivity or human judgement.
- Can not be used for fossils.

BSC lacks objectivity, and this is the main purpose of a Statistical Species Concept (SSC); to remove as much as possible, subjectivity in the classification of organisms. Another thing that is more subtle and often misunderstood is the difference between a species concept and a species delimitation criterion (de Queiroz 2007). In the following sections we will be first developing a delimitation criteria, then develop it into statistically robust algorithms then re-trace back to the definition of a Statistical Species Concept.

Some more fundamental problems present in many other existing clustering algorithms as well are as follows;

- When tested against a null model of random scatter, many of them give apparent clustering when there is none. See the following illustration, dendrograms suggesting common origins and suggesting hierarchical clustering

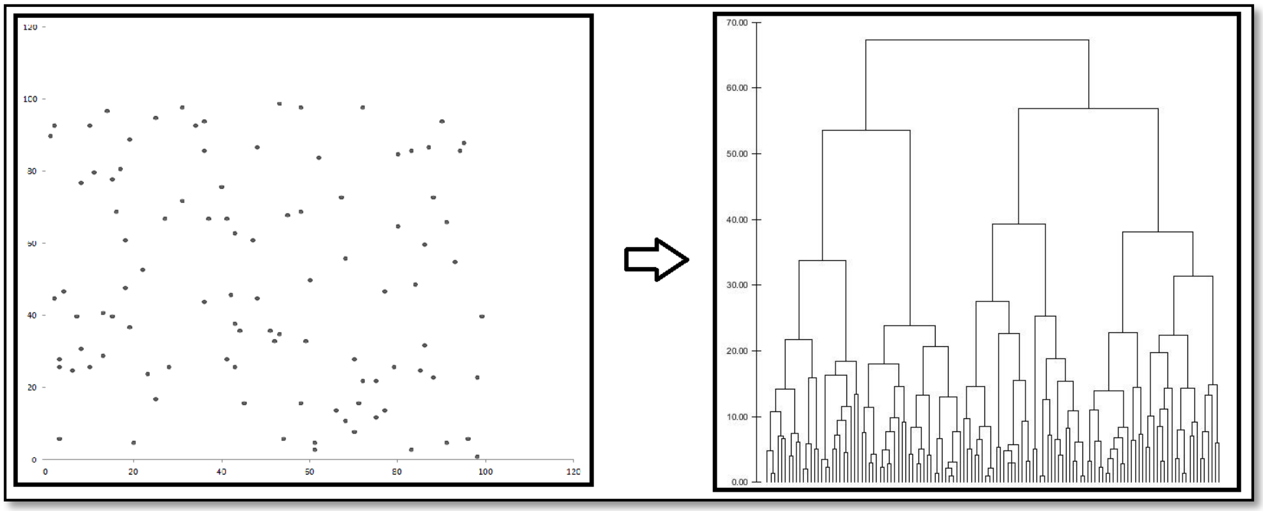


Fig 1.1: Shown above on left side is a scatter plot with random positioning of the points, where x and y are axes representative of arbitrary traits. The distance matrix from this plot was used to produce a Dendrogram plot using UPGMA clustering method, which is shown on the right. The dendrogram gives an appearance of 'clusters' when there are none. As one can see such ‘joining algorithms’ do not objectively reject the null hypothesis of there being no significant clustering in the data.

- Not human friendly in the sense of being difficult to comprehend and remember.
- A fundamental assumption of the Linnaean system is that life is hierarchically arranged: organisms belong to species, species to genera, genera to families, and so on. Although an organism belong to multiple taxa, an organism cannot belong to two taxa of the same rank. But this may not be necessary, an order in one classification may occur at a very different level of inclusiveness from an order in another classification. This feature that different levels of hierarchy across organisms may not overlap actually comes out as an obvious prediction of SSC. To illustrate, take a look at the following diagram;

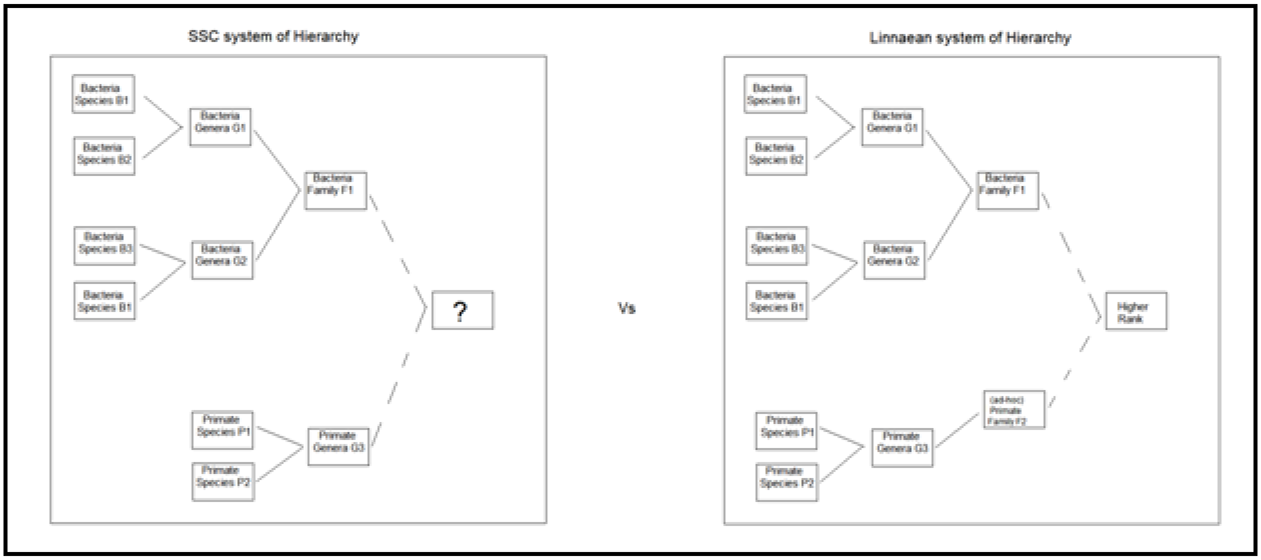


Fig 1.2; Diagrammatic representation of the conceptual difference in use of ranks in BSc vs. SSC

The continued use of the Linnaean ranks misleads (at least some) biologists to think that taxa assigned the same rank are similar regardless of their classification. This mistaken view causes biologists to enter “unfruitful debates” over the ranks of taxa (Hennig 1969, xviii; also see Griffiths 1976, and Ax 1987, Chapter K). Biologists could create more Linnaean ranks, but they are hesitant to do so because they would like to keep the framework of classification stable (Wiley 1981, 204). The interested reader may want to look at Hennig’s (1969[1981]) numerical classification of the Mecopteroidea as one example of the previous attempts at resolving this issue.

1.2 The Statistical Species Concept (SSC)

Defining species is not just about a philosophical debate but a lot of crucial scientific and as well as legislative tasks depend on it. The significance of a robust definition of species, as it is a fundamental building block of the theory of evolution, is not only necessary for communication and tagging purposes but also very important for quantifying speciation. Here, we shall not go into the details of why the need for a new concept of species but the interested reader can look into these excellent reviews (Hausdorf 2010;Isaac et al. 2004; Agapow et al.2004)

The Statistical Species Concept, or SSC for short, is a new concept which is based on looking at the frequency distribution of the distances between the individuals to statistically find a cut off for classifying them into species or other higher ranks. In the coming sections of this chapter we develop the algorithms for a delimitation criteria based on SSC. Most of the statistical procedures and parameters have been used earlier in similar contexts, but the recipe for the use of these ingredients to develop this algorithm for SSC is novel.

The primary objective of the SSC algorithm is to be able to objectively state, that for a given data, whether two individuals belong to the same or different species. The goal is to

- not assume any model of evolution
- eliminate any ad-hoc or human judgment based cut-off criteria,
- be able to attach a significance/pvalue for of predictions from SSC algorithm
- be able to reject or accept the null that there is no clustering and no hierarchy in the given data.

In this section, we first list some problems with the existing concepts of species. Then we build up the SSC algorithm by showing illustrations and using known or available statistical procedures.

1.3 Quantifying Clustering

Clustering as a pre-cursor to classification: We begin with a simple intuitive idea; consider a multi-dimensional space consisting of the all the characters/traits of an organism. A dimension in such a space is the quantified trait of the individual. Let us, for the simplicity of illustration, consider a hypothetical dataset where we have the weights and heights of a hundred individuals. We would like to know whether there are any ‘natural clusters’ in the data

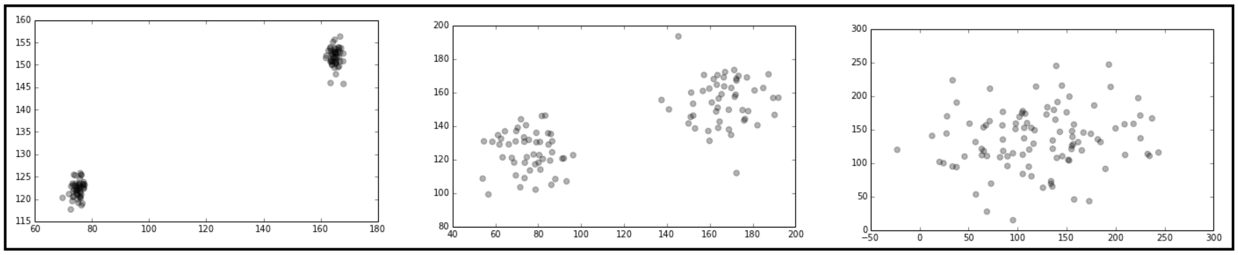


Fig1.3: Three plots for 100 individuals, the x axis and y axis are representative of two particular characters of the organisms being considered for a 2D visualization of the multi-dimensional character space. The clustering is clear in the first one, less so in the second one and absent in the third.

Two clusters can be said to be significantly different if the between-cluster distances are significantly greater than within-cluster distances. However this is not a sufficient test for the existence of natural clusters. Even if two contagions are picked up from a random set of points in the character space, they can satisfy this condition. If two clusters exist naturally in the data, then in the frequency distribution of all possible pair-wise distances we would observe a distinct peak indicating within cluster distances and another one corresponding to the between cluster distances. If there are more than two clusters of comparable spreads, there would be one leftmost peak representing within cluster distances of all clusters and then one or more peaks of between cluster distances. If there is no clustering, there will be uni-modality in the distribution. Any departure from uni-modality would be a unique signature for detecting natural clustering in a multi-modal data.

See the illustration below for three cases of clustering;

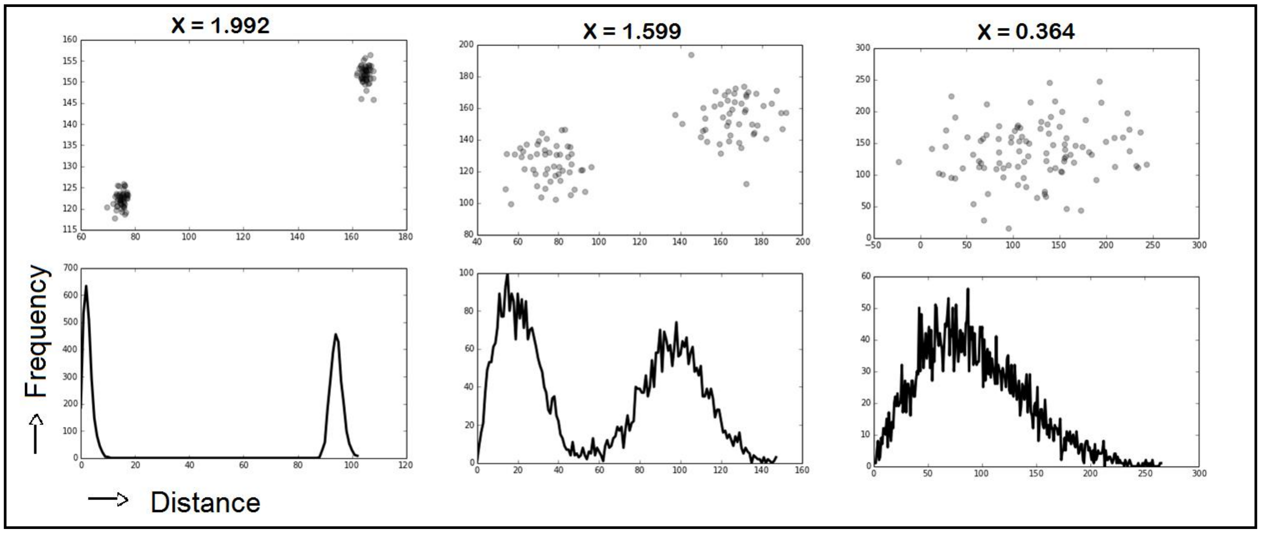


Fig1.4: The scatter plots (above) and the Frequency Distribution plots (below) are shown for three cases where the clustering is high in the first one, lower in the second one and absent in the third one. The values of clustering parameter X are displayed on the top.

One can see the corresponding deviation from uni-modality as an indicator of clustering. The significance of deviation from uni-modality can be tested using some indices/parameters available in literature. The problem of detection of bi- or multi-modality or departure from uni-modality is extensively discussed in statistical literature (refs). A clustering parameter X defined below is adopted from an easy to calculate and reproducible index of multimodality

To quantify this departure from uni-modality we define the statistic quantity, clustering parameter, X

$$X \; = \; (\text{Skewness})^2 - (\text{Kurtosis})$$

where X <= 1.488 is a sharp inequality for the class of uni-modal distributions (Klaassen et al., 2000). We tested this parameter using Monte-Carlo simulations generating one or more random clusters making1000 replicates each and found that for the cases of single cluster in the data i.e. uni-modal distributions 98.6% values were less than 0.5. On the other hand when more than one distinct clusters were randomly generated, we got 95.9% values greater than 0.5. The parameter X was large when the spread of the clusters was small as compared to the distance between clusters and approached 0.5 or less when the spread was large as compared to the distance.

X is basically a monotonic function of the clustering i.e. greater the distinctness clustering, greater the value of X. The cut-off value of 0.5 is ad-hoc but we can calculate the significance of the value of X by calculating the probability of which distribution it may belong to. Hence we can use this parameter as an indicator of the clustering present in data.

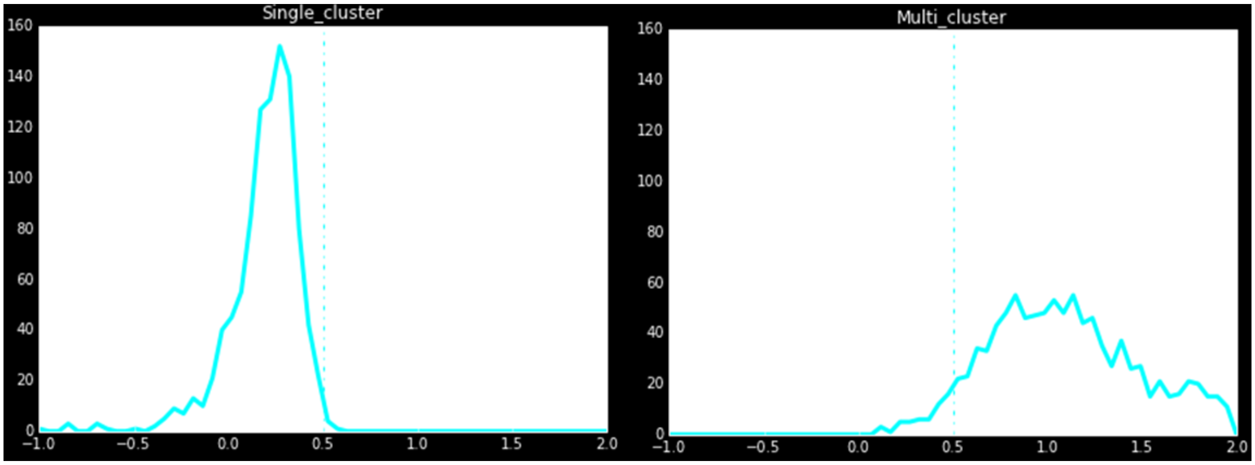


Fig1.5: The histogram for the values of clustering parameter X calculated in the Monte-Carlo simulation for 1000 replicates in case of a) Single cluster and b) Multiple clusters

A low clustering parameter indicates no clustering/single clusters/uni-modality, whereas a high value of X indicates that there is significant clustering in the data and it makes sense to go ahead and classify the individuals into different clusters. One can look again at figure1.3 to verify this. The parameter X gives a quantitative indication of how clear the clustering is and in principal we can have a pvalue corresponding to each value of X. Thus not only can clustering be tested for, but there is no need for a subjective criteria to determine when is clustering natural, one can look at the pvalue to have a statistical measure of how ‘natural’ the clustering is.

This Clustering Parameter X is the first fundamental building block of SSC. It resolves the problem of being able to accept or reject the null of hypothesis of there being no clustering. A small note here, though the clustering parameter is successful in 95.9% cases, it fails or gives false positives for 4.1% of the cases. We shall explain more about this in chapter 2 on Synthetic Sequences where we report the exact set of conditions where X fails to detect multi-modality.

Once we have a measure of significant clustering in data based on how high the value of X is, it now makes sense to go ahead and identify the clusters and see what set of common traits do they have in order to classify them as species group. Another note here, for the Frequency Distribution plot (FDp) to work properly it needs enough within-cluster distances as well as between-cluster distances. This means that for the above developed algorithm to detect a cluster, statistically there should be enough members from that group.

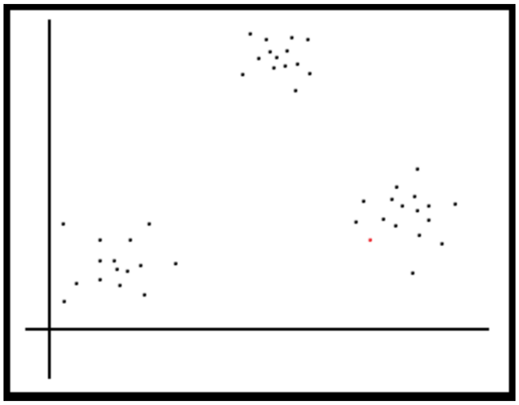
Two and multiple clusters: Finding two clusters is a simple and straightforward problem and the parameter X can be directly used. If there are multiple clusters, bimodality and parameter X are expected to work best if the spreads of different primary clusters are comparable. In that case the distribution of within cluster distances would be quite homogenous although between cluster distances will be more varied. As a result one would see a good first peak and then a broken or spread out second peak. On the other hand if individual clusters have substantially different spreads the FDp would be quite unpredictable. This is important for the species concept since we assume that species should have comparable spreads. In a given taxon one species should not be comparable to some other entire family. Although it is possible that some species have greater variability than another species, unless there is some comparability, it may not be worth calling them at the same level of clustering if the spreads are too different. If the spreads of clusters are comparable the first peak of FDp would always be nice and clear.

1.4 Making Boxes; Finding clusters

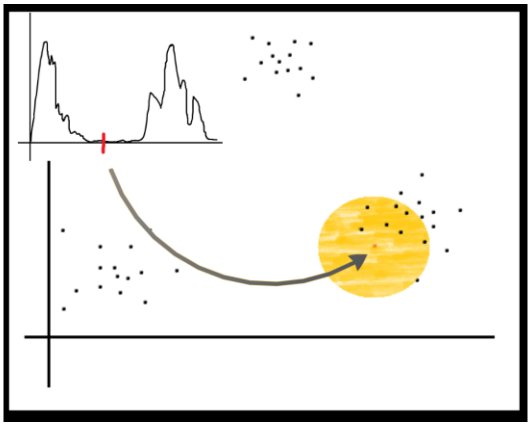
Finding Clusters: Having quantified the amount of clustering taking place in a data, the next obvious step is to find out the number of clusters, as well as which particular cluster does each individual in the data belongs to. For this we have used the idea that continues from the use of FDp; we find the radius at which the clustering takes place which corresponds to the first significant dip in the FDp. Using this radius, we can find the clusters by making circles and including every other individual that falls in that circle.

Shown below are 3 steps that illustrate this cluster finding algorithm;

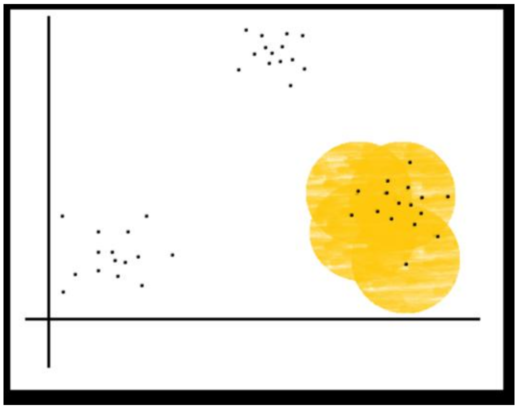
Step1: Randomly choose one individual (shown in orange)



Step2: Draw a circle of radius indicated from the FDP



Step3: For all the points in the shaded area repeat Step2



Keep repeating this procedure until no new point is covered by a circle.

Segregating clusters bridged by one or a few individuals: it is possible that natural clusters do exist but are linked by one or a few individuals that are less than the cut off distance from two clusters and therefore happen to join the two clusters into one by the above algorithm. We can detect such cases by looking for significant bimodality in the FDP within the cluster. If it exists it can be treated as a cluster complex rather than an individual cluster and further identification of individual clusters within the complex is possible using a different algorithm.

Defining a type specimen: Each cluster will have an individual from which the sum of squares of all distances within the cluster is minimum. This can be defined as the type specimen of the cluster. Once a cluster is identified one can search for a set of characters that is uniquely present in all members of the cluster and absent in all others. If such characters are found, they can be said to be the characters defining a cluster.

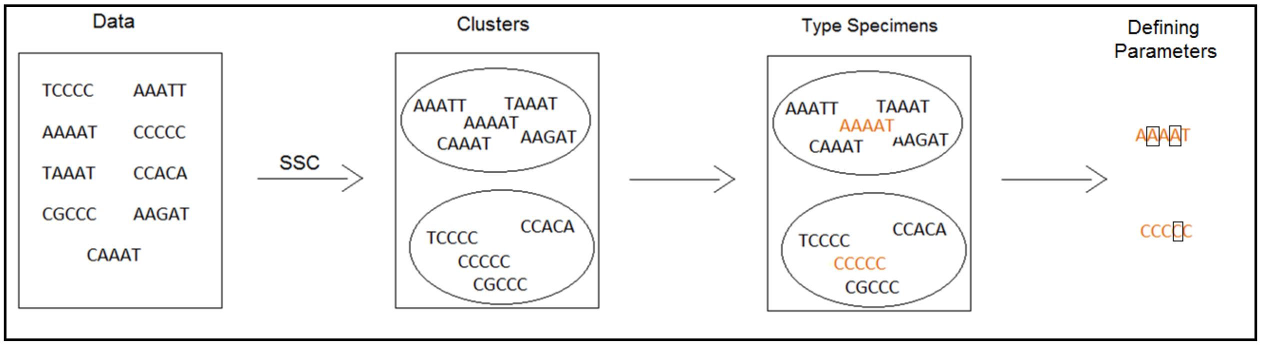


Fig1.6: Diagrammatic illustration of an example on how defining characters can be selected after identification of a cluster. 9 sequences are considered, which are clustered into two groups using the cluster finding algorithm developed above. Then the type-specimens are found as the individuals with the least sum of squared distances to every other individual in that cluster (shown in orange). The defining parameters for the corresponding clusters have been marked with black boxes. In cluster 1 A at position 2 and 4 and in cluster 2 C at position 4 are sufficient to define the clusters. Thus the clusters are made using all available data and then the data are trimmed to give simple identification criteria. This is the most likely path taken by classical taxonomy. But this might give a false impression that the groups were made using only these characters.

This can possibly resolve the hen and egg paradox in classical taxonomy. Did a classical taxonomist use body hair as a character defining mammals before knowing what mammals were? Most likely not. The human mind has an incredible capacity of pattern recognition the mechanisms of which are not completely understood. For example we are extremely good at face recognition but we often fail to define the set of characters based on which we identify a face. A verbal description of a face is highly inadequate as compared to our actual facial memory. We often are able to make classifications without being able to describe the criteria for such classification. For example one can identify a Punjabi looking face from a Maharashtrian face significantly correctly but without consciously knowing which characters were used to differentiate. We propose therefore that the early taxonomists first identified the groups at various levels based on the innate capacity of the human mind of recognising patterns from a large set of perceptible characters. Once the groups were identified they systematically and consciously searched for characters that were unique to the group and used them as a formal and objective definition of the group. However this gives a false impression that they first chose a set of characters defining a group and then the groups were identified accordingly.

Now what happens when there is clustering at more than one levels? This would lead to a Hierarchy in the clustering of individuals of the data. The next section deals with this question.

1.5 Levels of Clustering; Finding Hierarchy

To look for hierarchy, we need to know very clearly what hierarchy is, the Wikipedia page on hierarchy gives the following definition;

> *"Hierarchy is an arrangement of items (objects, names, values, categories, etc.) in which the items are represented as being above, below, or at the same level as one another"*

Unfortunately we don't understand hierarchy that well. In the Linnaean system of classification, every taxon once identified as a cluster, is assigned a rank, the rank decides the level of that particular taxon in the hierarchy. Though there was not a clear picture of evolution at the time Linnaeus, it is interesting to note how deeply embedded in this system of hierarchy is the notion of evolution that all organisms have evolved from similar ancestors. One can indeed imagine the justification for human tendency of classifying things and making boxes within boxes. Note that though it is logically obvious to imagine that there is always a bigger box that can encompass the two given boxes, the idea embedded in the minds that every individual entity must belong to a single box with defined rigid boundaries is not a necessity.

Hierarchy in other words is organization at different levels, and in biology, organization means clustering of organisms. And when we use our understanding of finding clusters in a multi-dimensional character space, it is not difficult to see that,

"*Hierarchy is clustering at more than one level*"

It translates to looking for significant clustering at more than one radius in the FDp. We already have a clustering parameter, but it only tells us if there is hierarchy in the overall data, hence we need to look for a method to find significance attached to the clustering associated with each radius.

When we were testing the algorithm to find clusters in different pilot situations, testing for different radii, finding the significance attached to each radius was pretty obvious as the next step, as in many scenarios it wasn't clear where is the dip in the FDp, what to do when there are multiple dips and so on. So, we found the number of clusters corresponding to each radii. We get the number of clusters corresponding to radius zero as the total number of individuals N. And for radius equal to the maximum distance between any two individuals, we get 1 single cluster. This is illustrated in the following graph;

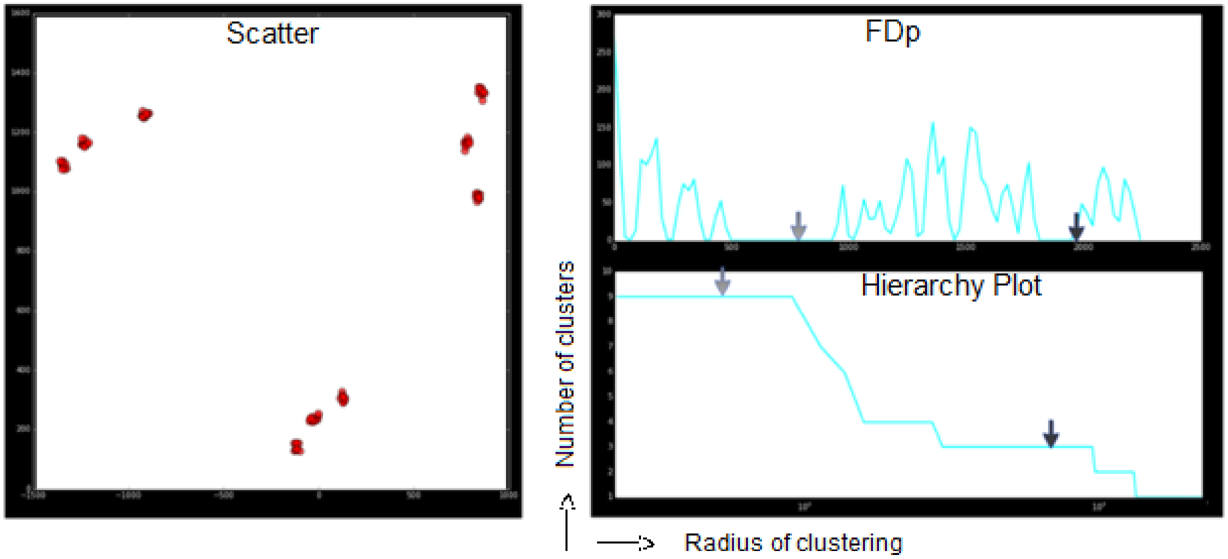


Fig1.7: On left side we have the scatter plots showing hierarchical clusters, where 100 individuals are clustered into 9 groups which are further clustered into 3 groups. On right side we have the FDp (above) and the plot for number of clusters (y axis) versus the radius (x axis). This graph clearly reflects two significant steps as a measure of two levels of hierarchy.

The idea is, if there is hierarchy i.e. clustering at more than one radius, than if we keep looking at the numbers of clusters at different radii and make a plot of number of clusters versus the radius of clustering, than the number of significant steps in this plot would tell us the hierarchy present in the data. From now on, we shall refer to this plot as the Hierarchy plot.

Significance of hierarchy: Even if we find unique steps in the hierarchy plot, it is difficult to tell by looking which of those are significant, for this we need a method to test for the significance of each step. We find this in the following way;

For the smallest radius we will get N clusters, N being the total numbers individuals used for the analysis. And for the largest there will be only one cluster. Now at each next radius after the first smallest one, the number of clusters will keep on reducing and effectively the problem boils to down to distributing N-2 clusters in R-2 steps equivalent to distributing N-2 balls in R-2 boxes. If we find the probability of each step size given the null hypothesis that N-2 clusters are distributed in R-2 steps randomly, this probability will serve as the pvalue for the significance of that step.

Now we derive the probability of finding a step of a given size in the Hierarchy plot. The probability of having a step of size at least 1 is equivalent to having at least one box with no balls $= \left(\frac{R-3}{R-2}\right)^{N-2}$ where is the resolution for the distance data used.

the probability of having a step of size at least 2 is equivalent to having at least two boxes with no balls

$$= \left(\frac{R-4}{R-2}\right)^{N-2}$$

In general, the probability of having a step of size at least S is equivalent to having at least S boxes with no balls

$$\begin{aligned} &= \left(\frac{R-(S+2)}{R-2}\right)^{N-2} \\ &= \left(\frac{(R-2)-S}{R-2}\right)^{N-2} \\ &= \left(1-\frac{S}{R-2}\right)^{N-2} \end{aligned}$$

So, given the resolution R of the distance data and the total number of individuals N, we can calculate the probability of a step of size S. This probability is based on random allotment of balls, i.e. clustering of individuals at random radii. This is the same as having no hierarchy as the null hypothesis. So if for a step of given size the probability is very low, it means that the step is significant and the null hypothesis of there being no hierarchy is rejected. Now we can find p-values for each step in the Hierarchy plot. As one can see from the formula, the probability decreases with increase in S and also with increase in N, which makes sense as greater the step size and lesser the number of individuals for a fixed value of R, lesser should be the probability. This is shown in the graph below,

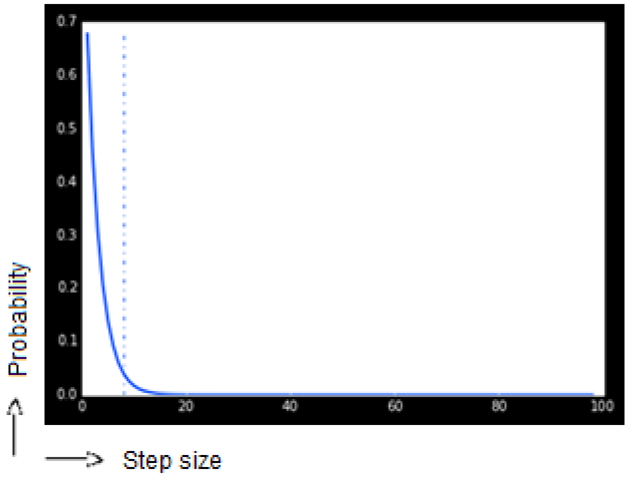


Fig1.8: The plot for decrease in probability with increase in the step size, for Resolution, R = 100 and number of individuals, N = 40. For a cut-off at 0.05, we get a minimum significant step size of 8, as indicated by the dotted line in the figure.

Reconstituting species concept: If we have a clear and objective way of finding natural clusters in a given character space, we can try to reconstitute a definition for species as

> *"According to SSC, a species is a group of individuals that when correspond to the first level of significant clustering using de novo clustering detection methods"*

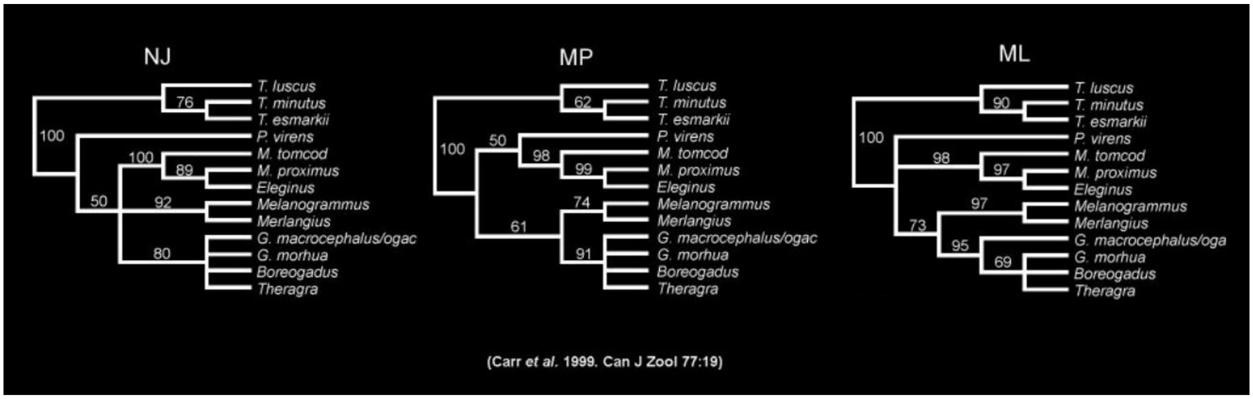
As discussed before, the first level of clustering can be picked out easily and reliably from the FDp if the spreads of primary clusters is similar to each other as compared to the spread of the entire data.

To build up a philosophy of species we should have principles and methods that can be used with any set of characterization data. Once we have such methods, the character space utilized can be expanded to cover as much data as possible. The dimensionality of the character space would be decided by the available technology which will be ever expanding but the principles and methods used to treat the data would remain essentially the same. If this is achieved one would expect the stability of clustering to increase in a saturation curve with the dimensionality and expanse of characterization data. Therefore beyond a critical amount of data the method will become practically independent of data inclusiveness. This is the ultimate desired state of taxonomy where subjectivity will be minimized. At present we may be quite short of reaching this saturation, but this can be visualized to be possible with the rapid expansion of data. The reconstituted SSC can reinterpret some of the classical concepts such as BSC. If two clusters are bimodal and therefore classified separately, crossbreeding can potentially merge the two peaks and make it uni-modal. However this will happen only if cross breeding is sufficiently frequent. Thus in this concept the frequency rather than possibility of crossbreeding would matter. It also means that species can not only diverge but also converge. This is compatible with the finding that in flowering plants many species have originated through cross population hybridization [(Abbott et al., 2013).

CHAPTER 2: Testing with Synthetic Data; a Pilot Study

2. Introduction

We want to test the usefulness of the alternative algorithms against a set of data where we know the evolutionary relations. This can be best done with synthetic data since we have a control over the relations. This cannot be done with real life data since it can easily become circular. So the idea is to compare the above developed algorithm for SSC to some of widely used phylogenetic reconstruction algorithms today. Some of these cladistic tree-joining algorithms are (i) Neighbour Joining (NJ) which represents one of the conceptually and computationally simplest, but does not consider any alternate topologies, (ii) Maximum Parsimony (MP) which is better as it considers alternate topologies but suffers from the problem of long branch attractor(Pol and Siddall 2001) and (iii) Maximum Likelihood (ML) which is currently believed to be the most sound (Russo et al., 1996) among the three, but requires heavier computational requirements. For illustration of the phylogeny trees that are made by these algorithms, consider the following figure with real data,



(Carr *et al.* 1999. Can J Zool 77:19)

Fig2.1: For a family of gadid fish, different predicted phylogenies by Neighbor Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) algorithms. Figure adopted from Carr et al. 1999

The main problem with these algorithms is that they do not consider the null hypothesis of there being no clusters and no hierarchy in the data. This means that given a set of gene sequences which do not have a hierarchical structure in reality, the algorithms may suggest a hierarchy. We will test this using synthetic data below for different cases of hierarchies.

We shall compare the results of Maximum Likelihood (ML) method to the results of our SSC algorithm. For all the figures of molecular phylogenetic analysis by ML method shown in the subsequent sections, the evolutionary history was inferred based on the Tamura-Nei model (Tamura and Nei, 1993). The bootstrap consensus tree inferred from 100 replicates (Felsenstein, 1985) is taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches (Felsenstein, 1985). Initial tree(s) for the heuristic search were obtained automatically as follows. When the number of common sites was < 100 or less than one fourth of the total number of sites, the maximum parsimony method was used; otherwise BIONJ method with MCL distance matrix was used. The analysis involved 100 nucleotide sequences. There were a total of 1000 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 (Tamura et al., 2011).

2.1 Zero levels of Hierarchy

We take a set of 100 sequences, each of size 1000 made by randomizing the positions of A, T, C and G bases with equal probability. Here there are no clusters and no hierarchies and ML satisfactorily shows no clustering with significant boot strap values. We look at 3 different sub-cases where the synthetic sequences are designed in slightly different ways

2.1.1 For the sub-case where all the 1000 sites were randomized in the 100 sequences, we get the following result;

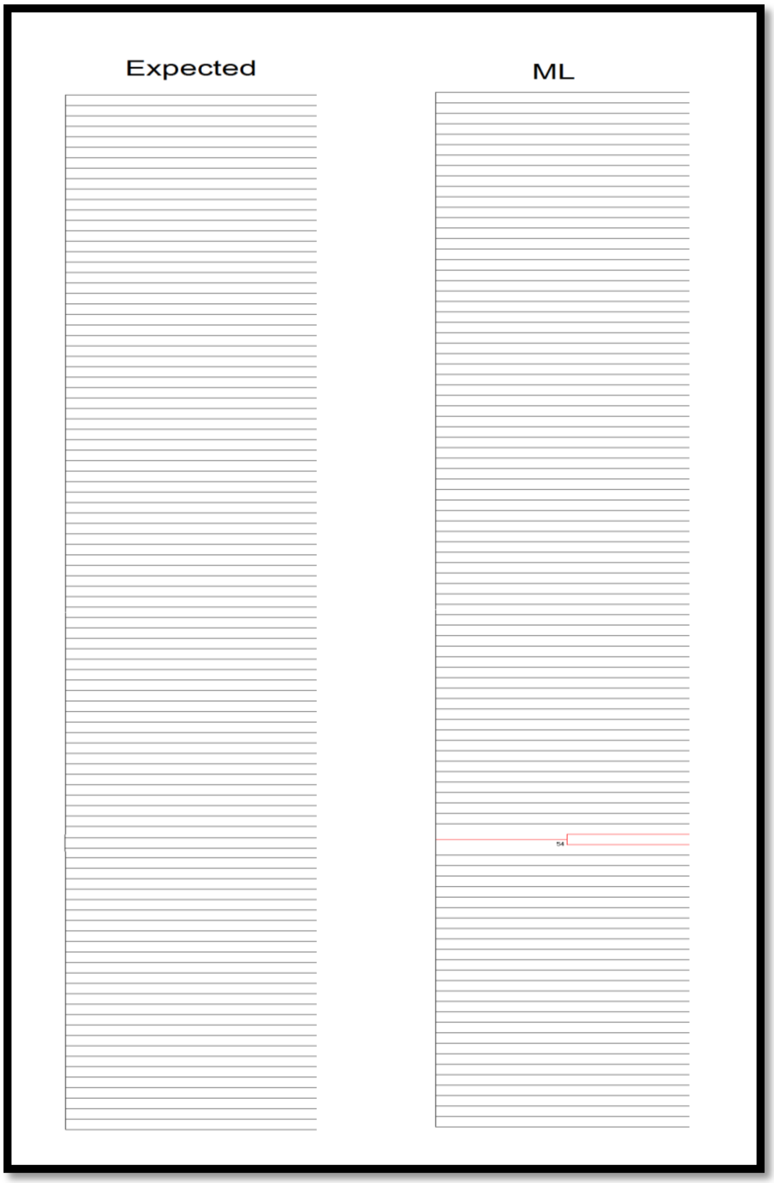


Fig2.2: Barring one branch out of 100which indicated false hierarchy(shown in red above), ML performs as expected.
We test this for 10 times and every time ML satisfactorily shows the correct trees.

Next, we show the results from our SSC algorithm in the following 10 hierarchy plots;

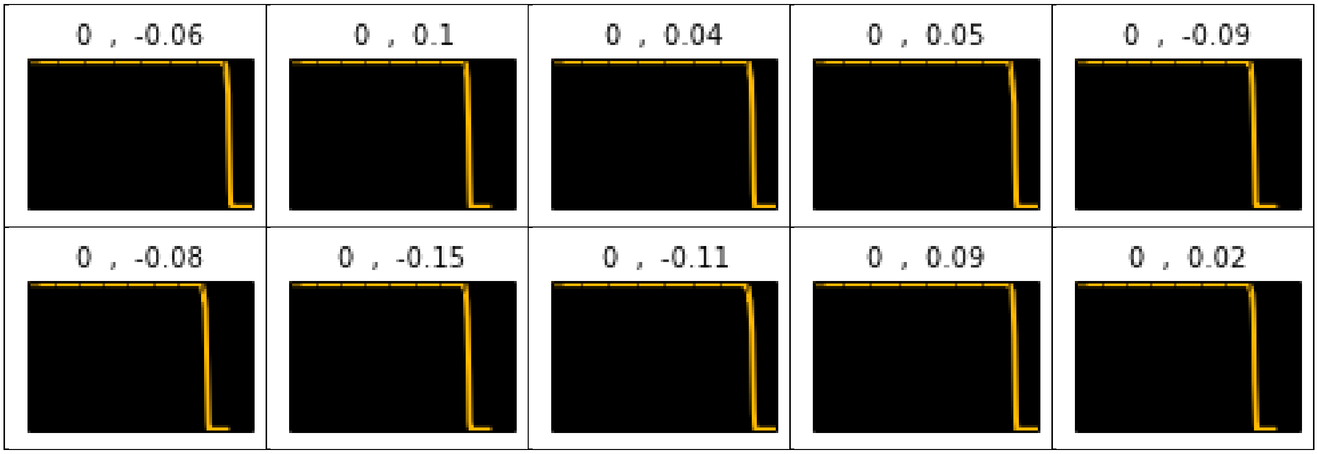


Fig2.3: Above shown are the Hierarchy plots for corresponding 10 sub-cases. The values written over the top are Number of significant steps and clustering parameter X respectively. For each plot, we get the prediction of 0 levels of hierarchy and low values of X, indicative of no significant clustering.
Hence SSC also passes 10 out of 10 times.

2.1.2 For the sub-case, where 10% of the sites were randomized and the remaining 90% sites were common for all the 100 sequences, we got same results as in 2.1.1; We tested for 5 times and barring one or two branches out of hundred where there was incorrect representation, ML performed as expected.

Results of SSC algorithm for this set of 5 sub-cases;

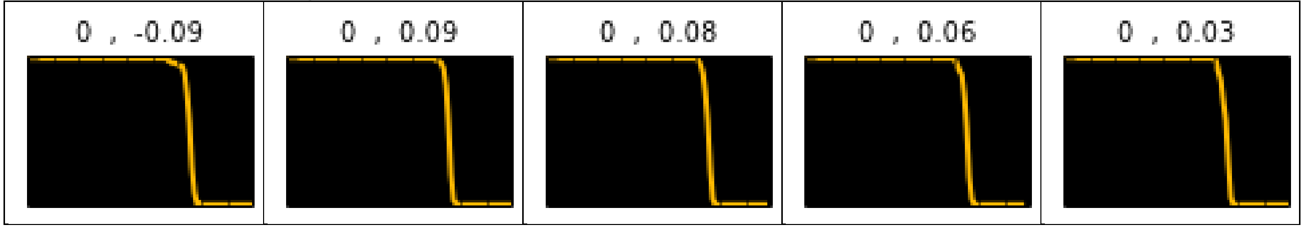


Fig2.4: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter X respectively. For each plot, we get the prediction of 0 levels of hierarchy and low values of X indicative of no significant clustering.
Hence SSC passes 5/5 times.

2.1.3 For the sub-case, where 1% of the sites were randomized and the remaining 99% sites were common for all the 100 sequences, we again got same results as in 2.1.1; ML passed 5/5 times.

Results of SSC algorithm for this set of 5 sub-cases;

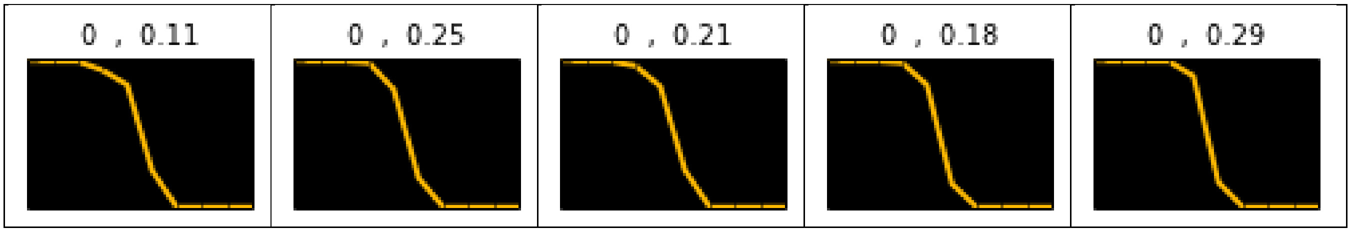


Fig2.5: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter X respectively. For each plot, we get the prediction of 0 levels of hierarchy and low values of X indicative of no significant clustering.
SSC passes 5/5 times.

Conclusion for zero level of hierarchy: This was the case where there was no hierarchy and no clustering, both ML and SSC make correct inferences in each of the 20 sub-cases.

2.2 One level of Hierarchy

2.2.1 We take 10 sequences with all their sites randomized and then derive 10 daughter sequences from each of them by making 10% random substitutions at random sites. The following figure is an illustration of this procedure;

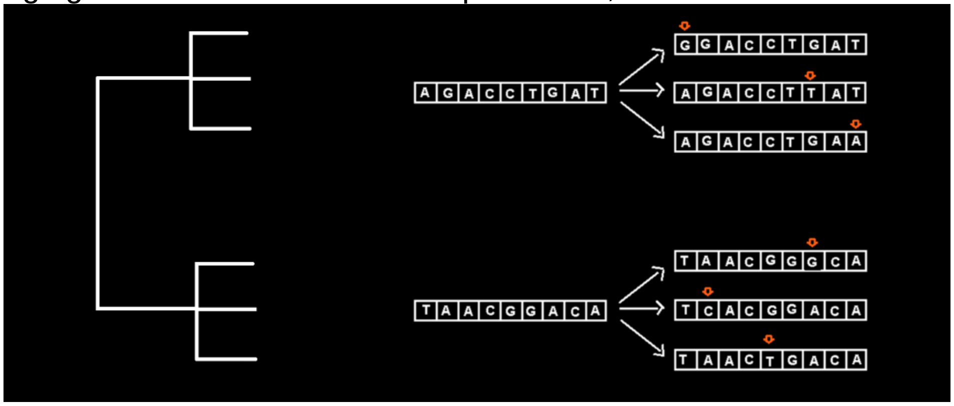


Fig2.6: Shown on left is the expected hierarchy and on right is how the synthetic sequences are constructed. Starting with the two random sequences, the daughter sequences are derived by making 1 mutation each.

We get the following result;

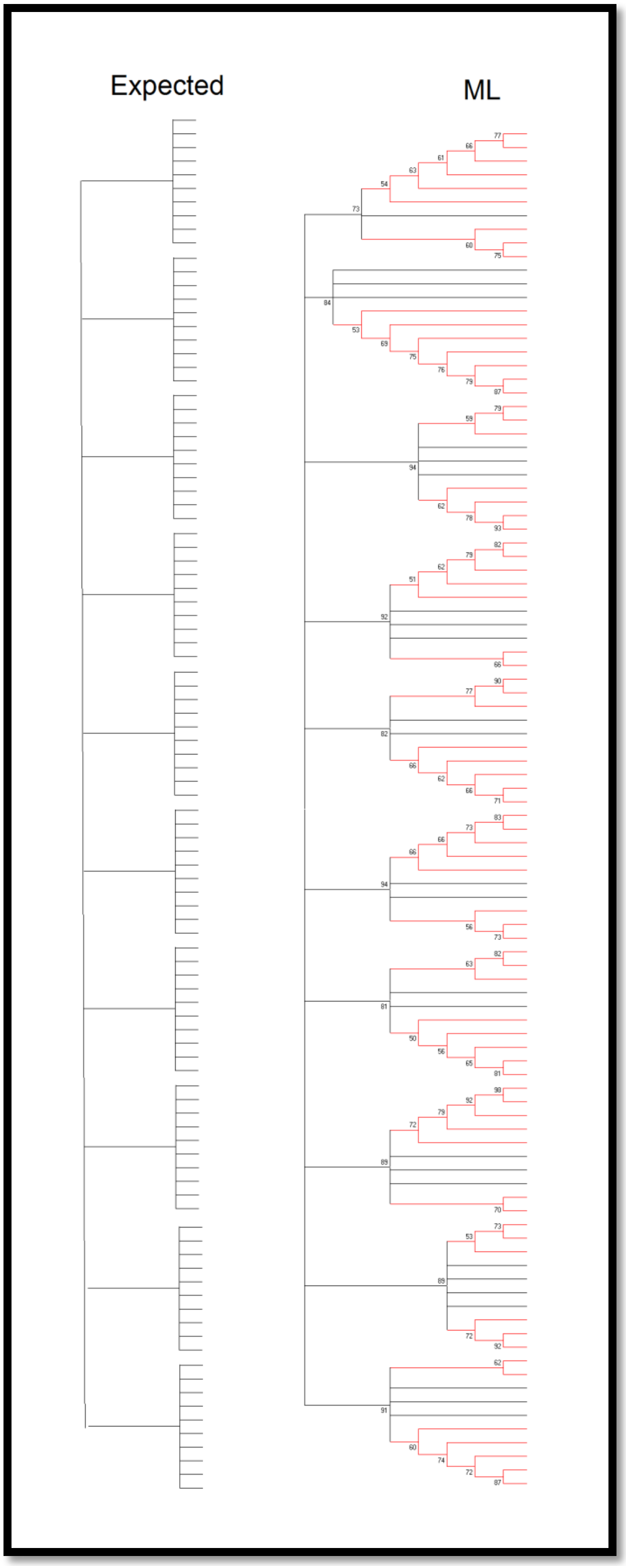


Fig2.7: Shown in red are the branches which do not match. ML predicts more than 1 levels of hierarchy with significant bootstrap values.

We test this for 5 times, and ML falsely reports clustering where there was none expected in each of the 5 sub-cases.

Results of SSC algorithm for this set of 5 sub-cases;

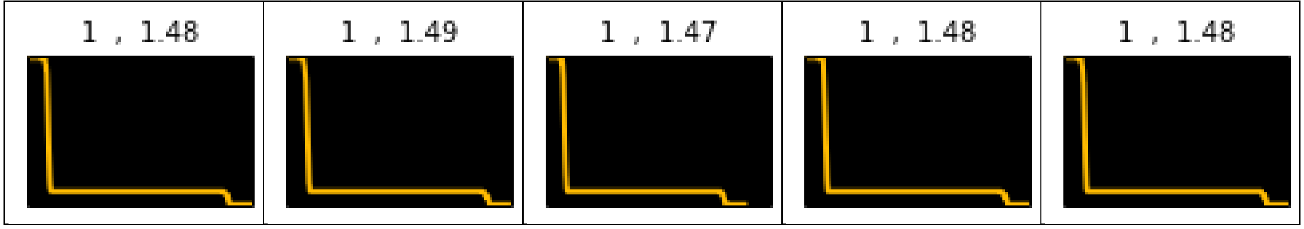


Fig2.8: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter X respectively. For each plot SSC predicts 1 level of hierarchy and high clustering parameter values indicative of significant clustering.

2.2.2 Next we take 10 sequences which have 90% common regions and are randomized with respect to one particular region of size 10%.10 daughter sequences are derived from each of them by making 60% substitutions at random sites in a different region. We get the following result;

(shown in red), ML also resulted in a false hierarchy at a higher level with significant bootstrap value.

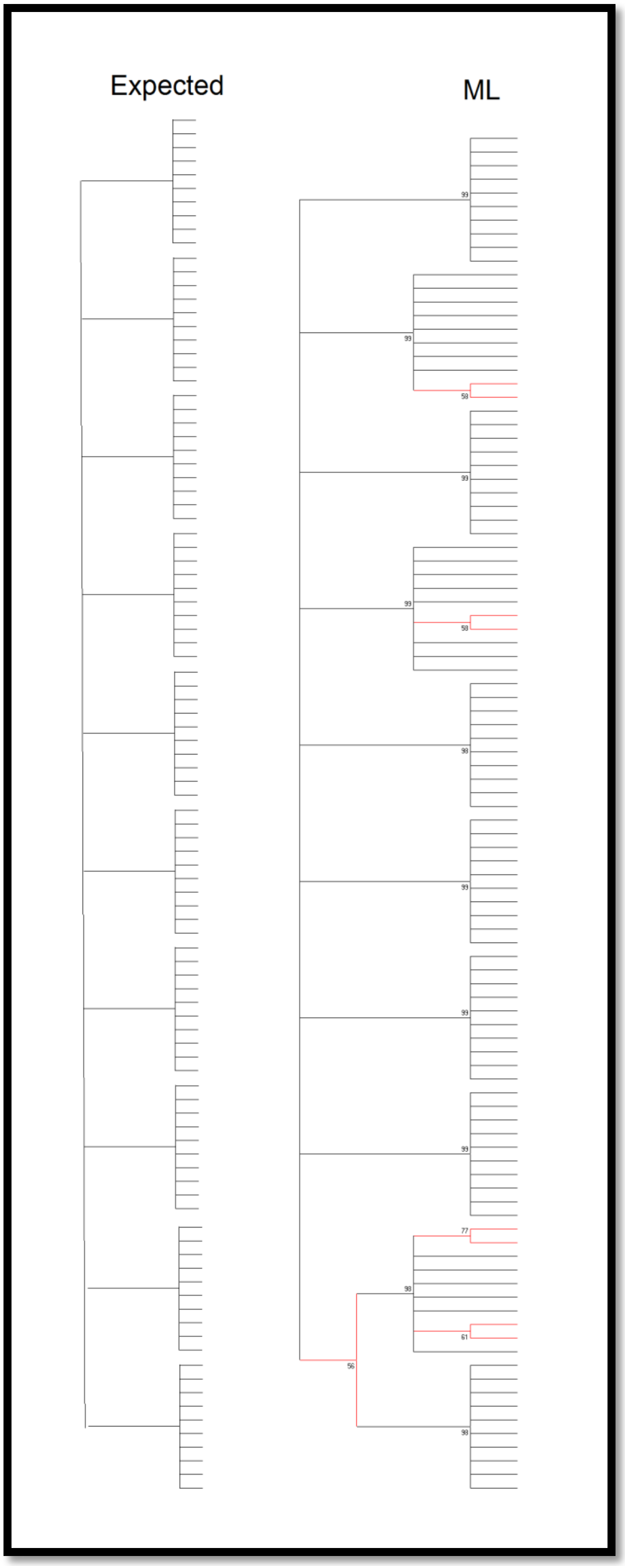


Fig2.9; Along with a few branches indicating clustering where there was none expected (shown in red), ML also resulted in a false hierarchy at a higher level with significant bootstrap value.

We test this with 5 sub-cases and 2 out of 5 times ML predicts (approximately) correct phylogenies.

Results of SSC algorithm for this set of 5 sub-cases;

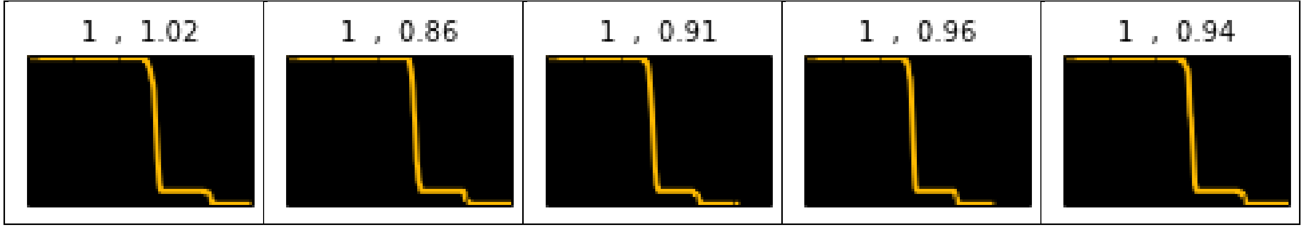


Fig2.10: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter X respectively. For each plot, SSC predicts 1 level of hierarchy and high clustering parameter values around 1 indicative of significant clustering.

2.2.3 Next, we take 10 sequences which have 90% common regions and are randomized with respect to one particular region of size 10%.10 daughter sequences are derived from each of them by making only 10% random substitutions at random sites, in a different region. We do this for 5 times and the results were similar to the (2.2.2) case; 3 out of 5 times ML predicted (approximately) correct phylogenies.

Results of SSC algorithm for this set of sub-cases;

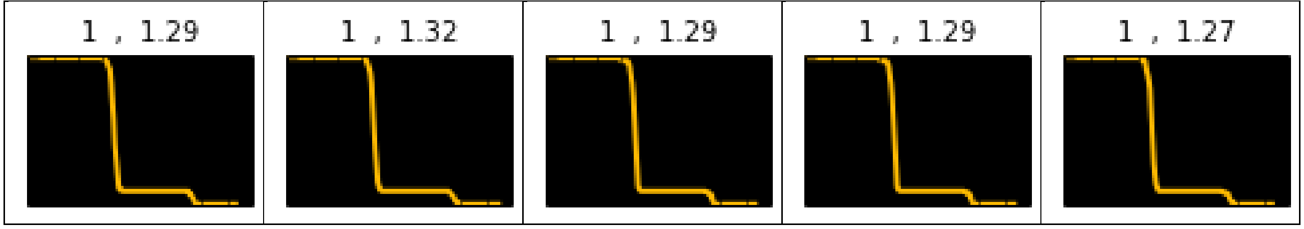


Fig2.11: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter X respectively. For each plot, SSC predicts 1 level of hierarchy and high clustering parameter values around 1.3 indicative of significant clustering.

2.2.4 Next we take 10 sequences which have 90% common regions and are randomised with respect to one particular region of size 10%. 10 daughter sequences are derived from each of them by making 1% random substitutions at random sites, in a different region. We do this for 5 times and the results were same as (2.2.3) case; 3 out of 5 times ML predicted (approximately) correct phylogenies.

Results of SSC algorithm for this set of sub-cases;

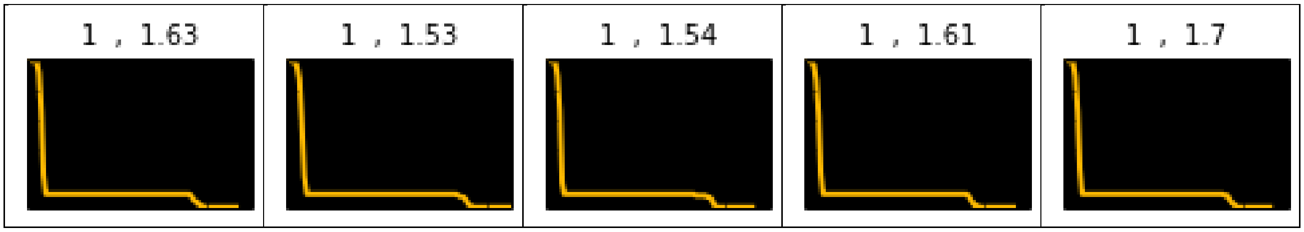


Fig2.12: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter X respectively. For each plot, SSC predicts 1 level of hierarchy and high clustering parameter values around 1.3 indicative of significant clustering.

Conclusions for one level of hierarchy: ML shows clustering in cases where there was none expected i.e. too many false positives. It passes 8/20 times while SSC passes 20/20 times.

2.3 Two levels of Hierarchy

2.3.1 We take 5 sequences with all sites randomized, and then derive 4 daughter sequences from each of them, by making 10% random substitutions at random sites, and then further derive 5 sequences from each of them, by again making 1% random substitutions at random sites. We get the following result,

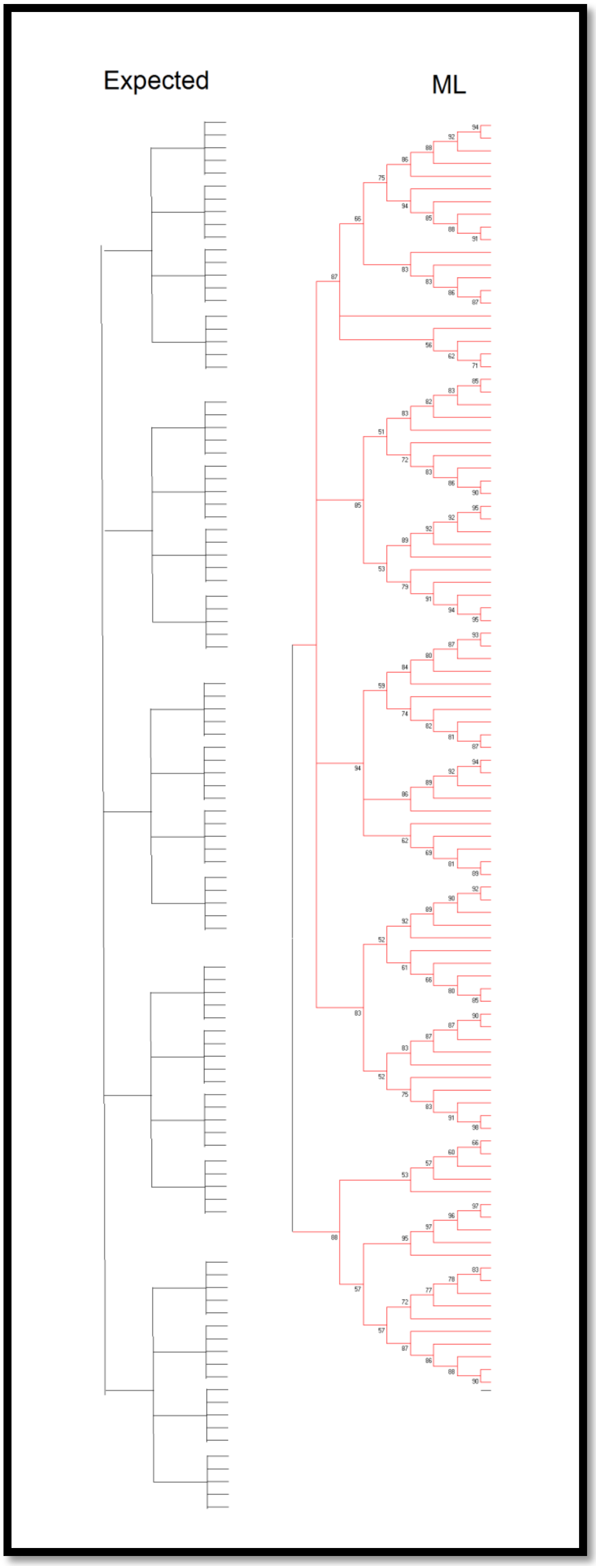


Fig2.13: The predicted tree by ML does not match to the expected. Ml predicts false hierarchies as well clustering where none is expected.

We do this for 5 times, and ML predicts phylogenies with different hierarchies each time, none of which match with the expected tree.

Results of SSC algorithm for this set of 5 sub-cases;

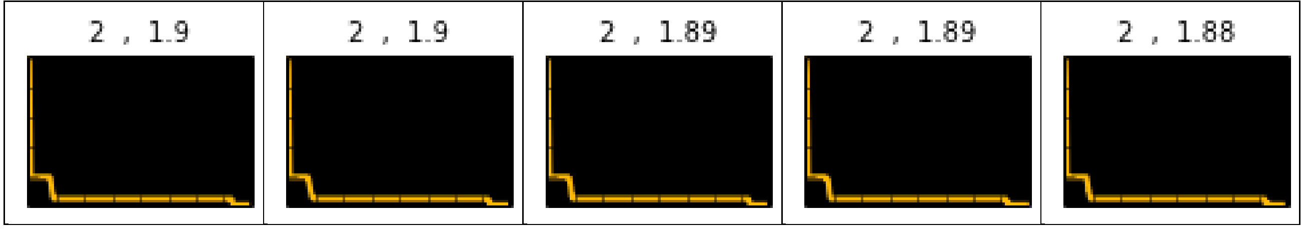


Fig2.14: Above shown are the Hierarchy plots for the 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter respectively. For each plot, SSC predicts 2 levels of hierarchy and high clustering parameter values around 1.9, indicative of significant clustering.

2.3.2 Next, we take 5sequences with all sites randomized, and then derive 4 daughter sequences from each of them by making 10% random substitutions at random sites. Further for each of the daughter sequences so obtained, we derive 5 from each by making 50% random substitutions at random sites. For all of the 5 sub-cases, ML failed to predict the expected hierarchy.

Results of SSC algorithm;

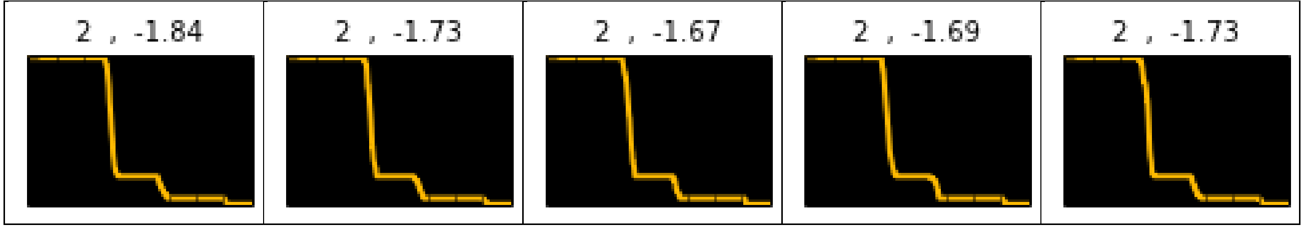


Fig2.15: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter respectively. For each plot, SSC predicts 2 levels of hierarchy and low clustering parameter values around -1.7 indicative of no significant clustering.

The SSC algorithm correctly predicts the hierarchy 5 out of 5 times. But an interesting point is raised about the low values of clustering parameter, and there are two ways one can interpret the result. The first obvious remark would be that SSC and in particular clustering parameter X failed to predict the clustering in this case. We have seen this before, X is not a perfect measure of clustering, it only predicts correctly in 95.9% of the cases. But it is also important to note that while designing the sequences for these 5 sub-cases, there were a greater percentage of random substitutions in the daughter sequences than the parent sequences from which they were derived. The final sequences which were used, were derived by making 50% substitutions in the sequences that were derived by making only 10% substitutions in completely random sequences.

Intuitively one can makes sense of this observation by imagining that clustering in the parent/ancestor sequences could get overshadowed by greater substitutions in the daughter sequences. But the question is whether this should destroy the hierarchy or not? SSC still predicts the expected hierarchy, but then intuitively, it does not make sense to find hierarchy in clustering given that there is no clustering!

2.3.3 We take 5 sequences having a common region of size 70% and the remaining 30% randomized. Then we derive 4 daughter sequences from each of them by making 10% random substitutions at random sites, and then further from each of the daughter sequences so obtained, we derive 5 from each by making 50% random substitutions at random sites. We get the following result;

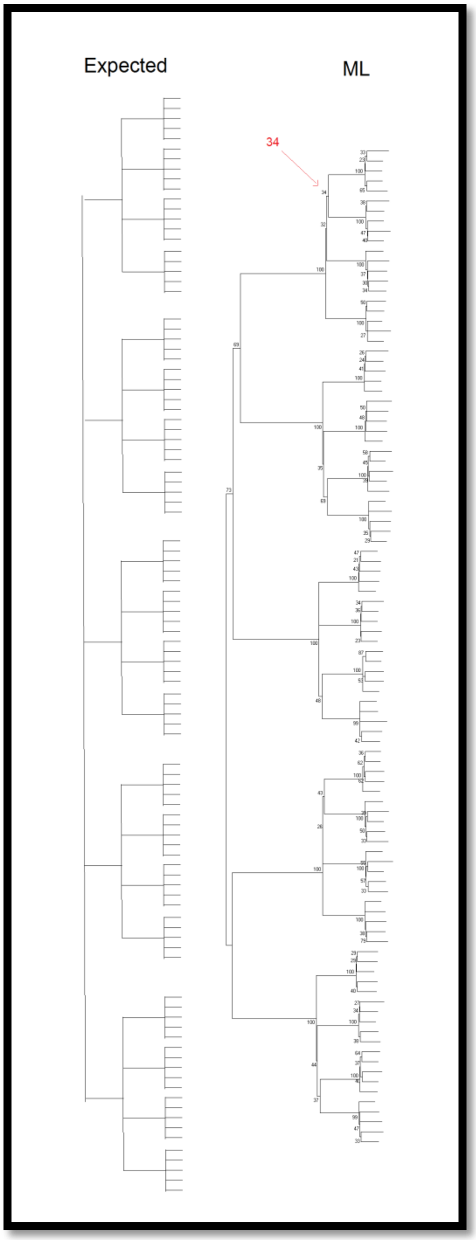


Fig2.16: Shown above are the expected and the bootstrap consensus tree predicted by ML. Note this is not the condensed tree as shown in all the previous figures. This is to show, that since the distances scale is maintained in a consensus tree, we can see that

two levels of hierarchy are apparent, but upon bootstrapping we get false representations, as many clusters get condensed into one or are attached to higher levels than expected. One such value is shown in red, indicating a low bootstrap value of 34.

This is a case of false negative; ML fails to represent hierarchies that are expected. We test this for 5 times, and each time ML fails to predict the expected hierarchy with significant bootstrap values.

Results of SSC algorithm for this set of 5 sub-cases;

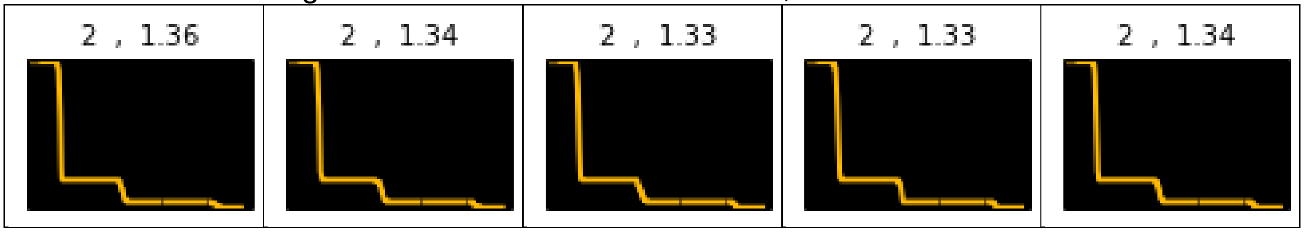


Fig2.17: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter respectively. For each plot, SSC predicts 2 levels of hierarchy and high clustering parameter values around 1.3 indicative of significant clustering.

2.3.4 We take 5 sequences having a common region of size 90% and remaining 10% sites randomized. Then we derive 4 daughter sequences from each of them by making 20% substitutions in a particular region, and then further from we derive 5 from each of the daughter sequences so obtained, by making 30% substitutions in the same region. Results of ML did not match with the expected.

Results of SSC algorithm for this set of sub-cases;

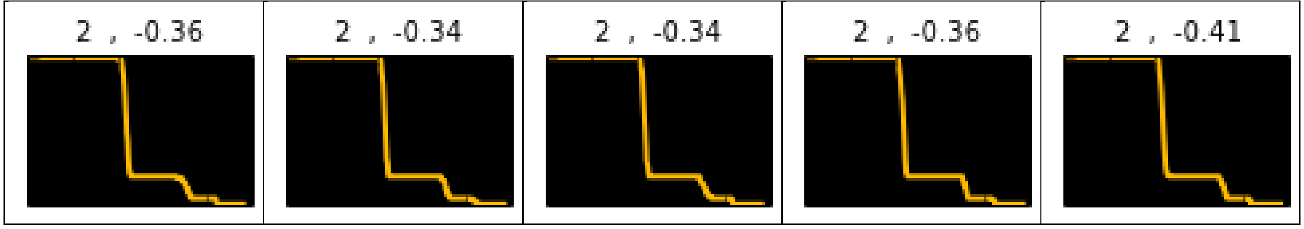


Fig2.18: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter respectively. For each plot, SSC predicts 2 levels of hierarchy and low clustering parameter values around -0.3 indicative of no significant clustering.

The SSC algorithm correctly predicts the hierarchy 5 out of 5 times but the same point regarding the low values of clustering parameter is raised again. While designing the sequences for these 5 sub-cases, there were greater percentage of substitutions in the daughter sequences than the parent sequences, the final sequences were derived by making 30% substitutions in the sequences that were derived by making only 20% substitutions in the sequences that were derived by making only 10% substitutions in completely random sequences. Along with this 'overshadowing of ancestry' hypothesis, we also note that the number of individuals for the first level of clustering is only 5 now as compared to 10 in the case of 1 level of hierarchy, this results in a lower probability of detection of those clustering at that level.

Conclusions for two levels of hierarchy: ML gives false negatives and is unable to predict the correct tree for any of the 20sub-cases. SSC, though it fails on predicting significant clustering in 10/20 sub-cases, predicts the hierarchy correctly in all the 20/20 sub-cases.

2.4 Three levels of Hierarchy

2.4.1 We take 3 sequences with all sites randomized, and then we derive 3 from each by making 30% random substitutions at random sites. We again derive 3 from each of these 9 sequences by making 10% random substitutions at random sites. And then finally we derive 4 from each of the so obtained 27 sequences by making 1% random substitutions at random sites. The following are the results for the 108 sequences made synthetically to reflect 3 levels of hierarchy;

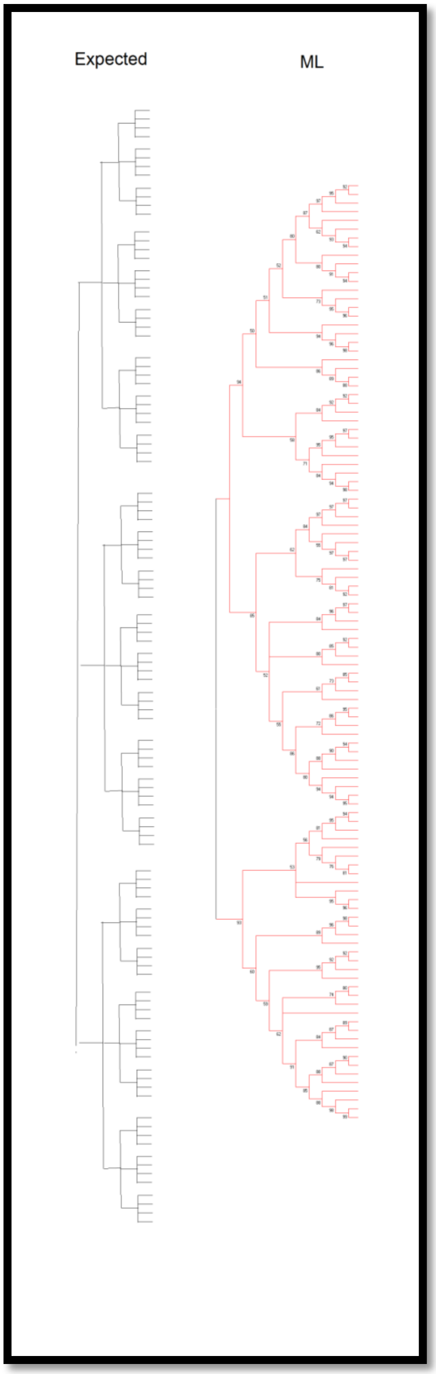


Fig2.19: Shown in red is the result of ML as compared to the expected tree on the left.
The predictions of ML do not match to the expected.
We do this for 5 sub-cases and ML does not predict the expected tree for any of them.

Results of SSC algorithm for this set of sub-cases;

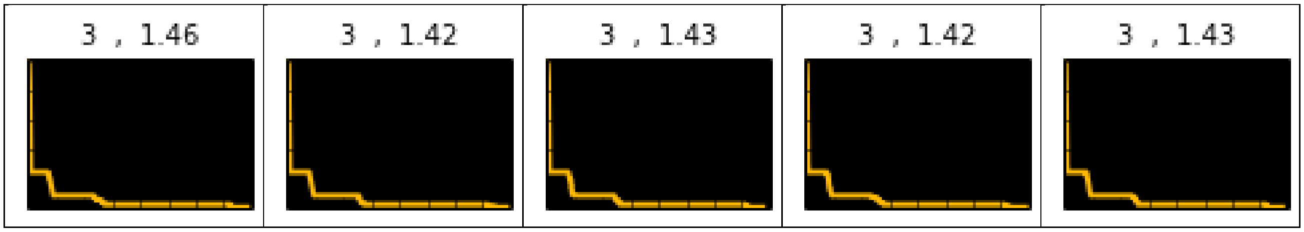


Fig2.20: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter respectively. For each plot, SSC predicts 3 levels of hierarchy and high clustering parameter values around 1.4, indicative of significant clustering.

2.4.2 We take 2 sequences with all sites randomized, and then we derive 2 from each by making 30% random substitutions at random sites. We again derive 5 from each of these 4 sequences by making 10% random substitutions at random sites. And then finally we derive 5 from each of the so obtained 20 sequences by making 1% random substitutions at random sites. For the 5 sub-cases we tested for, ML failed to predict the expected tree every time.

Results of SSC algorithm for this set of 5 sub-cases;

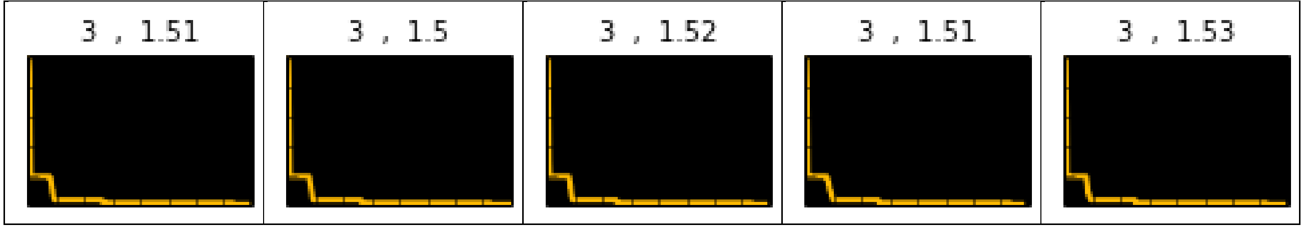


Fig2.21: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter respectively. For each plot, SSC predicts 3 levels of hierarchy and high clustering parameter values around 1.5, indicative of significant clustering.

2.4.3 We take 2 sequences with all sites randomized, and then we derive 2 from each by making 30% random substitutions at random sites. We again derive 5 from each of these 4 sequences by making 10% random substitutions at random sites. And then finally we derive 5 from each of the so obtained 20 sequences by making 1% random substitutions at random sites. For the 5 sub-cases we tested for, ML failed to predict the expected tree every time.

Results of SSC algorithm for this set of 5 sub-cases;

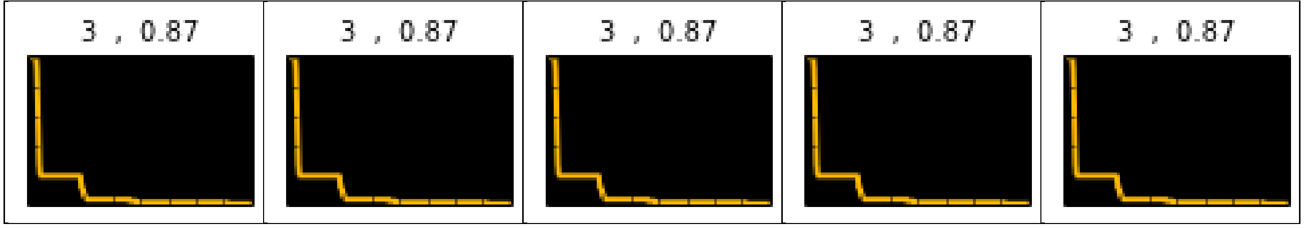


Fig2.22: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter

respectively. For each plot, SSC predicts 3 levels of hierarchy and high clustering parameter values around 1.5, indicative of significant clustering.

2.4.4 We take 2 sequences with 90% region common and the remaining 10% randomized. Then we derive 2 sequences from both of these sequences by making 30% substitutions in a particular segment. Then we derive 5 sequences from each of these 4 sequences by making 10% substitutions a different region. And then finally we derive 5 sequences from each of the so obtained20 sequences by making 1% substitutions in the same region. The following are the results of ML;

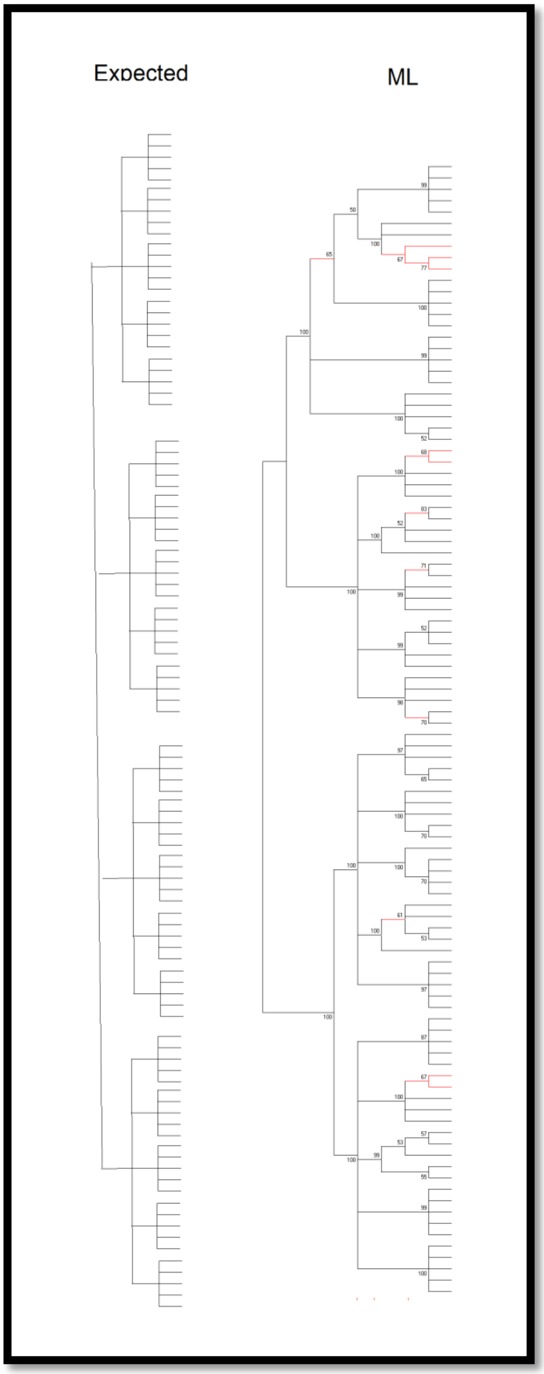


Fig2.23: Shown on left is the expected tree and on right, is the tree predicted by ML. Barring a few branches which are not as expected, ML satisfactorily predicts the expected tree.

We test this for 5 sub-cases and ML performs satisfactorily 3/5 times.

Results of SSC algorithm for this set of 5 sub-cases;

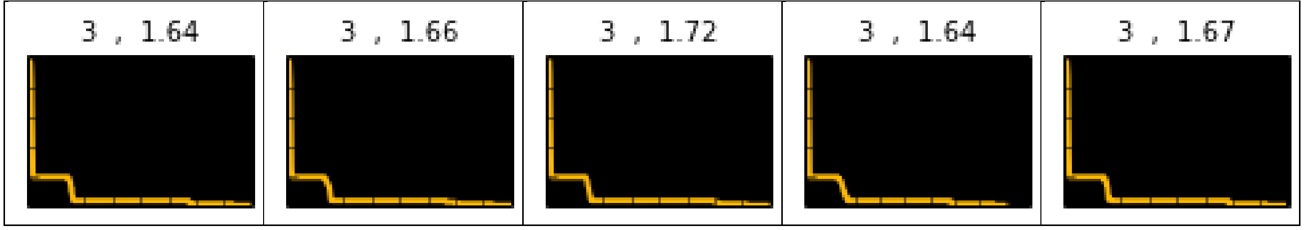


Fig 2.24: Above shown are the Hierarchy plots for corresponding 5 sub-cases. The values written over the top are Number of significant steps and clustering parameter respectively. For each plot, SSC predicts 3 levels of hierarchy and high clustering parameter values around 1.5, indicative of significant clustering.

Conclusions for three levels of hierarchy: ML gives mixed results and is able to predict the correct tree only 3 out of 20 times, while SSC predicts correctly 20/20 times.

2.5 No Significant level of Hierarchy (Random branching)

This is a different case from 2.1, where we expect the ideal algorithm to predict number of significant levels of hierarchy as zero. We designed the synthetic sequences by the following algorithm; take one random sequence and derive another from it by making random number of substitutions at random sites. Now take these two sequences, choose one randomly and derive another sequence in the similar way by making random number of substitutions at random sites. Keep doing this until you get 100 sequences in all. This should give an expected tree with random branching structure.

ML gave the following result;

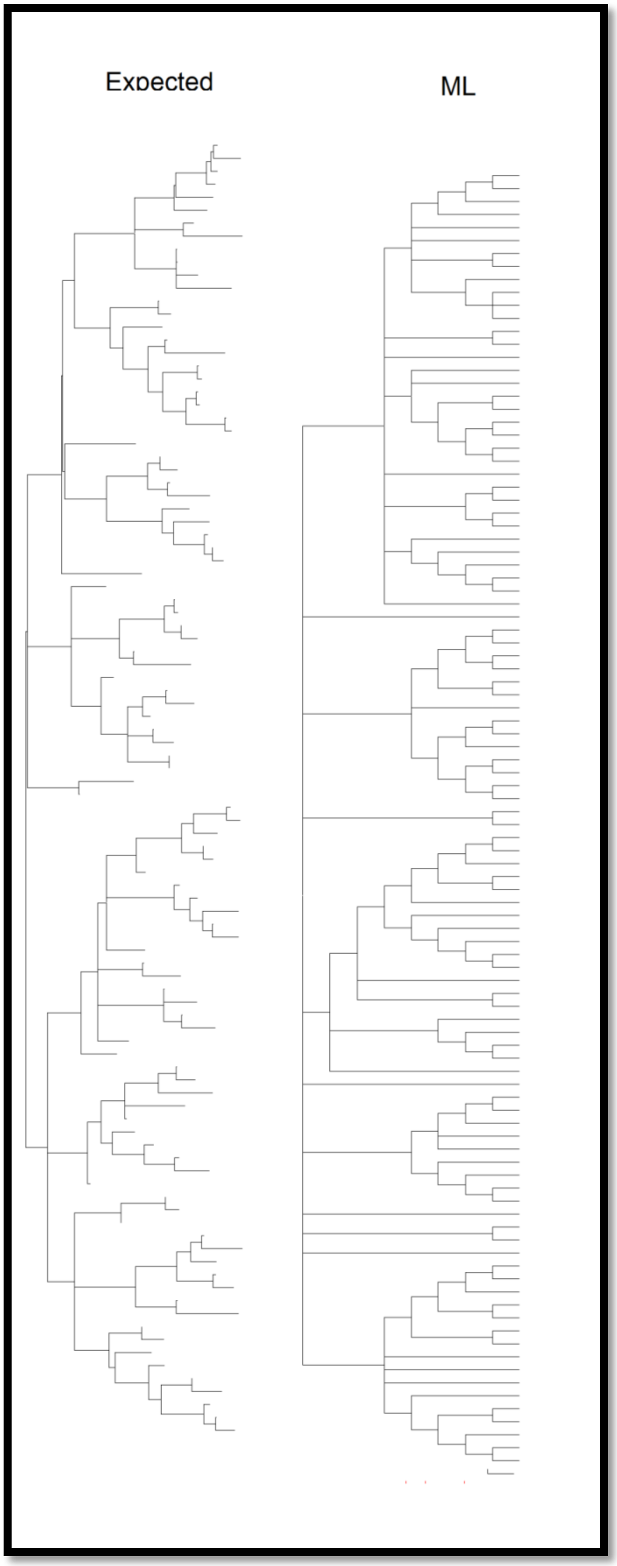


Fig2.25: Shown on left is the expected tree and on right, is the prediction by ML. It can be seen that ML captures the branching faithfully. However it does not give any idea as to whether there are discrete levels of hierarchy or not.

Results of SSC algorithm for the set of 20 sub-cases;

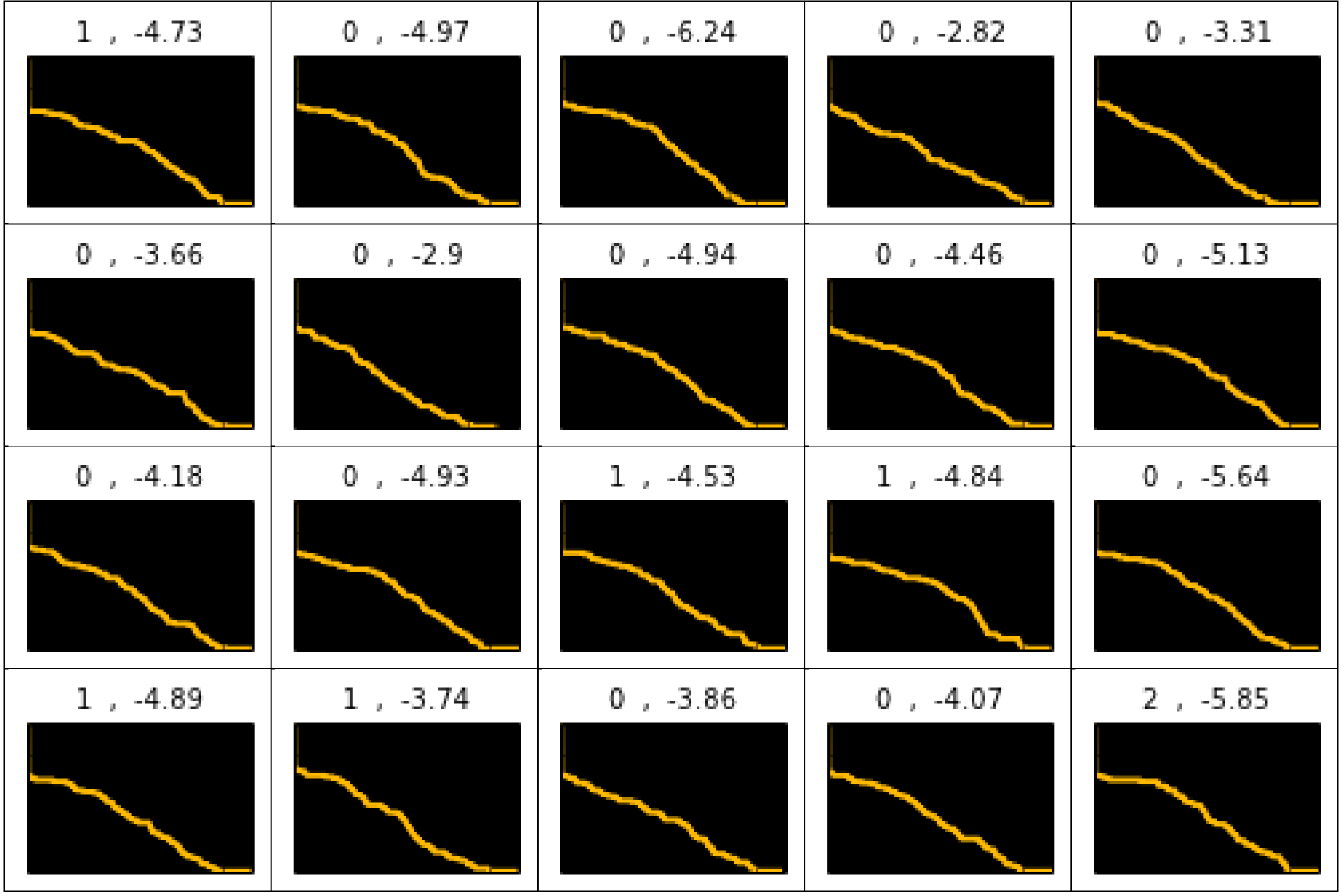


Fig2.26: Above shown are the Hierarchy plots for corresponding 20 sub-cases. The values written over the top are Number of significant steps and clustering parameter respectively. For 14 out of 20 plots, SSC predicts 0 levels of hierarchy and very low clustering parameter values below -2, which are indicative of no significant clustering. In 6 of the sub-cases SSC predicted 1 or 2 levels of hierarchy which was a false positive error.

Conclusion for random branching: When the data had random branching with had no clusters and hierarchy ML gave no indication as to presence or absence of hierarchy. SSC predicted no hierarchy and no clustering 14/20 times.

2.6 ML Vs SSC; Evaluating Performance

In all there are 5 different cases for 0, 1, 2, 3 and no levels of hierarchy. Each case has 20 sub-cases, shown below is the summary of the results;

Case (levels of hierarchy)	**ML**	**FDp**
0	20/20	20/20
1	8/20	20/20
2	0/20	20/20
3	3/20	20/20
None	0/20	14/20
Total	31/100	94/100

Some points to be summarized;

- SSC performs significantly better than ML in identifying clusters and hierarchies as it makes almost no errors of the type false negative. This is important if one wants to have a system independent of the accepted model of evolution, for rejecting the null hypothesis that there is no clustering (species) and no hierarchy (higher ranks).
- Much more work is needed to corroborate these findings, especially in the direction of understanding hierarchy and to make sense of the SSC's prediction of finding hierarchy when there is no clustering as observed in 2.3.2 and 2.3.4
- Having fortified the validity of this as pilot project and the use of synthetic sequences, we would like to test in future another important class of alternative algorithms that are widely used namely the Principal component analysis (PCA) and Bayesian Inference Methods.

The performance of ML and SSC in synthetic data needs to be weighed in the light of evolution and phylogenies. If we believe that there are natural hierarchies in taxonomy, i.e. evolutionary divergence was not continuous but in spurts that left more than one levels of hierarchy, SSC will be able to test and identify the steps. On the other hand if evolution took a random branching path with no detectable hierarchies, ML would be a good tool after SSC is used to test that there are no hierarchies. In the latter case we will have to modify the structure of taxonomy or accept that the hierarchical structure of taxonomy is not natural but imposed artificially.

CHAPTER 3: Applying to Real Genetic Data

3. Introduction

Having tested with synthetic sequences, the next thing that we would like to do is test the SSC algorithm on real sequence data. But what kind of data would really be required? There are a few criteria that such a data must fulfill. The first and the foremost condition is the appropriate model of distances to be used. Ideally the distances should exhaustively consider all the differences between the individuals under study. This means along with the whole genome, we would also need distances that incorporate epigenetic, morphometric, osteological and even behavioral aspects of every individual under study. This is impossible to achieve and is not even going to happen in the near future. But statistically speaking, all these differences, with the advancement in our understanding and technology, adding more parameters to the distance model would not lead to significant changes in the clustering results of SSC algorithm. This point of saturation is inevitable but at point will it be reached is a question that only time will tell. This does not mean that single genes have not been used to delineate species; in fact the 16s mitochondrial gene is widely used today in microbial ecology as a 'barcode' gene to reconstruct phylogenies [Coenye and Vandamme 2003]

Coming to the pragmatic execution of the task of testing SSC on real data, we would need datasets with at-least enough number of individuals, so that we can get statistically significant answers. Also we must ensure that there is sufficient number of individuals from each species, but this can only happen if we start with the knowledge based on the Biological Species Concept, that which individual belongs to which species. Otherwise, it would become a circular argument and there is no way of knowing whether SSC functions properly or is it just that the conventional naming of the individuals has been wrong. The data that has been sampled or has been uploaded on the internet will have an inherent bias as the species categorization for each individual has been done already. As a result, the within cluster distances are grossly underrepresented and therefore one may not get a clear FDp.

We use a few example data sets where many individuals from the same species are characterized and therefore we hope to get the within and between cluster distinction. SSC algorithm is tested on three sets of data; i) mtDNA and HVS-I sequences of Homo sapiens, ii)16s gene of 3 species belonging to the family Homonidae and iii) cox1 gene of catfishes belonging to the family Bagridae

3.1 Homo Sapiens

Based on Prof. K. Thangaraj's work on genetic sampling of human DNA from the different geographical and ethnic human populations of India, we looked at two datasets. The data was clubbed together based on the sequences available; complete mtDNA for one set and HVS-I for the other. If geographical and ethnic differences amount to calling different groups as different species, then we should expect SSC to make predictions of significant clustering from such a data.

3.1.1 Dataset-I

Data Used:
Complete mtDNA genome sequences for 86 individuals;

i. 12 individuals belonging to Indian Muslim populations from different states (GenBank accession number FJ157838-FJ157849) (Eaaswarkhanth et al., 2009)

ii. 54 individuals belonging to a range of ethnic populations form 17 states of India (GenBank accession number FJ467940–FJ467993)(Thangaraj et al., 2009)

iii. 20 individuals from 3 tribal populations; Bhil, Bharia and Sahariya (GenBank accession number GU480001- GU480020)(Sharma et al., 2012)

Results:

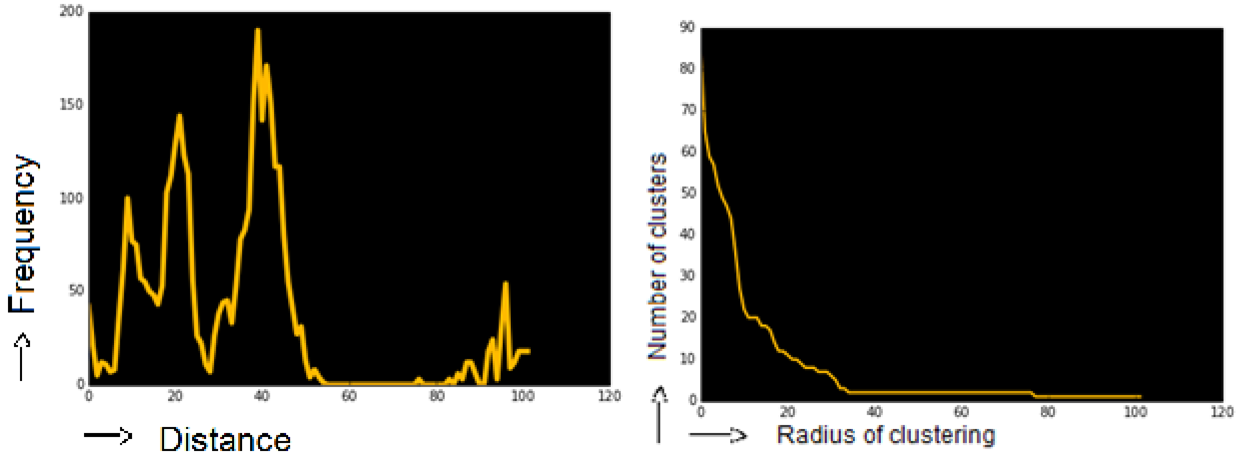


Fig3.1: Shown above on the left is the FDp which is multi-modal and on the right is the hierarchy plot which gives one significant step.

Clustering parameter, X = -0.564
Since the clustering parameter is very low, it implies that all individuals belong to one single cluster. Hierarchy plot gives one significant step corresponding to two clusters; one of size 83 and the other of size 3. These 3 individuals do not correspond to any distinct geographical region, caste or ancestry and can be safely consider as outliers rather than forming a distinct cluster.

Conclusion: Based on a low value of X and no significant level of hierarchy, SSC predicts that all individuals belonging to the same species. This is particularly important because the data come from distinct groups by cast/region that have been shown to differ in certain haplotypes (Reich et al., 2009). But our analysis shows that the difference does not indicate different species.

3.1.2 Dataset-II

Data Used:
The sequences of first hypervariable segment (HVS-I) regions for 625 individuals
i) 472 individuals belonging to Indian Muslim populations (GenBankaccession number FJ157366-FJ157837) (Eaaswarkhanth et al., 2009)
ii) 153 Siddis and 269 from individuals from the nearby Indian populations (GenBankaccession number JN022021–JN022442) (Shah et al., 2011)

Results:

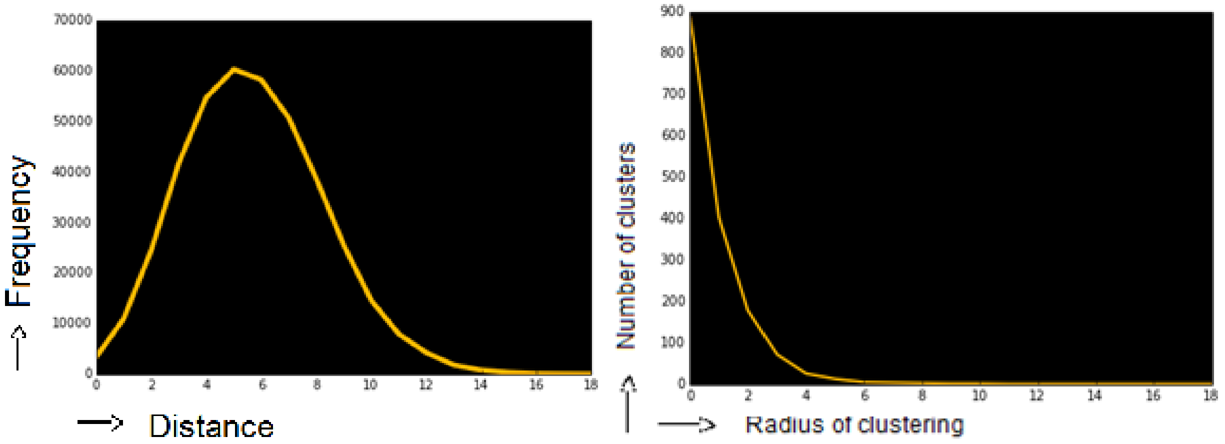


Fig3.2: Shown above on the left is the FDp which is uni-modal and on the right hierarchy plot gives no significant steps with pvalue less than 0.05

Clustering parameter, X = 0.066

Conclusion: Clearly, SSC predicts the individuals of this particular dataset of humans, taken from different parts of India, as belonging to one single species. Here again people of different ethnicities are represented by substantial numbers in the data but the FDp does not deviate from uni-modality. This is compatible with our concept that the first level of distinct clustering should define species.

3.2 Homonidae

After having dealt with some real data at the species level, we look at some data across genera level, taken from the freely accessible NCBI website. We wanted to test whether such a data corroborates with SSC to give two significant levels of hierarchies reflecting the clustering at the species and at the genera level.

Data Used: Ribosomal RNA genome sequences for 40 individuals;

i) 10 individuals marked as Pan paniscus
ii) 10 individuals marked as Pan troglodytes
iii) 20 individuals marked as Homo sapiens

Results:

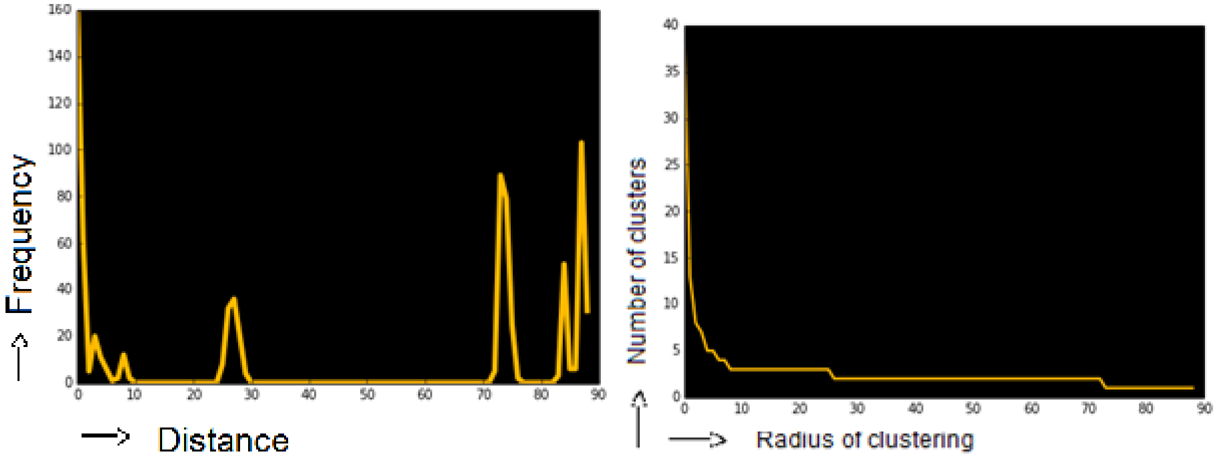


Fig3.3: On the left is shown the FDp, which is clearly mutli-modal. On the right we have the hierarchy plot which gives two significant steps.

Clustering Parameter, X = 1.824
Details of the significant steps with pvalue less than 0.05 are;

Step number = 3 Corresponding radius = 8 Number of clusters = 3 Pvalue = 0.0002			

Details of 3 clusters	Size = 20
	Spread = 2
		Size = 10
		Spread = 4
			Size = 10
			Spread = 9

List of Members	DQ834558.1|_Homo_sapiens DQ834559.1|_Homo_sapiens DQ834560.1|_Homo_sapiens DQ834561.1|_Homo_sapiens DQ834562.1|_Homo_sapiens DQ834563.1|_Homo_sapiens DQ834564.1|_Homo_sapiens DQ834565.1|_Homo_sapiens DQ834566.1|_Homo_sapiens DQ834567.1|_Homo_sapiens DQ834580.1|_Homo_sapiens DQ834581.1|_Homo_sapiens DQ834582.1|_Homo_sapiens	AB062538.1|_Pan_troglodytes AB062539.1|_Pan_troglodytes AB062540.1|_Pan_troglodytes AB062541.1|_Pan_troglodytes AB062542.1|_Pan_troglodytes AB062543.1|_Pan_troglodytes AB062544.1|_Pan_troglodytes AB062545.1|_Pan_troglodytes AB062546.1|_Pan_troglodytes AB062547.1|_Pan_troglodytes	AB050150.1|_Pan_paniscus AB050151.1|_Pan_paniscus AB065137.1|_Pan_paniscus AB065138.1|_Pan_paniscus AB065139.1|_Pan_paniscus AB065140.1|_Pan_paniscus AB065141.1|_Pan_paniscus AB065142.1|_Pan_paniscus AB065143.1|_Pan_paniscus AB065144.1|_Pan_paniscus

	DQ834585.1|_Homo_sapiens DQ834586.1|_Homo_sapiens DQ834587.1|_Homo_sapiens DQ834589.1|_Homo_sapiens DQ834590.1|_Homo_sapiens DQ834591.1|_Homo_sapiens DQ834592.1|_Homo_sapiens		
Step number = 4 Corresponding radius = 26 Number of clusters = 2 Pvalue = 3.837e-13			

Details of 2 clusters	Size = 20
	Spread = 2
		Size = 20
		Spread = 29	

List of members	DQ834558.1|_Homo_sapiens DQ834559.1|_Homo_sapiens DQ834560.1|_Homo_sapiens DQ834561.1|_Homo_sapiens DQ834562.1|_Homo_sapiens DQ834563.1|_Homo_sapiens DQ834564.1|_Homo_sapiens DQ834565.1|_Homo_sapiens DQ834566.1|_Homo_sapiens DQ834567.1|_Homo_sapiens DQ834580.1|_Homo_sapiens DQ834581.1|_Homo_sapiens DQ834582.1|_Homo_sapiens DQ834585.1|_Homo_sapiens DQ834586.1|_Homo_sapiens DQ834587.1|_Homo_sapiens DQ834589.1|_Homo_sapiens DQ834590.1|_Homo_sapiens DQ834591.1|_Homo_sapiens DQ834592.1|_Homo_sapiens	AB062538.1|_Pan_troglodytes AB062539.1|_Pan_troglodytes AB062540.1|_Pan_troglodytes AB062541.1|_Pan_troglodytes AB062542.1|_Pan_troglodytes AB062543.1|_Pan_troglodytes AB062544.1|_Pan_troglodytes AB062545.1|_Pan_troglodytes AB062546.1|_Pan_troglodytes AB062547.1|_Pan_troglodytes AB050150.1|_Pan_paniscus AB050151.1|_Pan_paniscus AB065137.1|_Pan_paniscus AB065138.1|_Pan_paniscus AB065139.1|_Pan_paniscus AB065140.1|_Pan_paniscus AB065141.1|_Pan_paniscus AB065142.1|_Pan_paniscus AB065143.1|_Pan_paniscus AB065144.1|_Pan_paniscus	
Conclusion: SSC predicts two significant levels of hierarchy, in accordance with the accepted ranks for the given taxa, showing clustering at the species and the genera level. This is clear demonstration that in this case not only species are natural clusters, at a higher level genera are also natural clusters.			

3.3 Bagridae

We look at three datasets from the family Bagridae (fresh water catfish); i) Bleekeri, ii) Cavasius and Sengtee, and iii) Malbaricus. These are tested against the predictions made by Dr. Neelesh Dhanukar based on morphometric, phylogenetic and osteometric analysis.

3.3.1 Bleekeri:

Data used: Cox1 gene sequences from 29 individuals belonging to the species Mystus Bleekeri. Neelesh predicts that Wai and Ambodi samples form a distinct cluster.

Results:

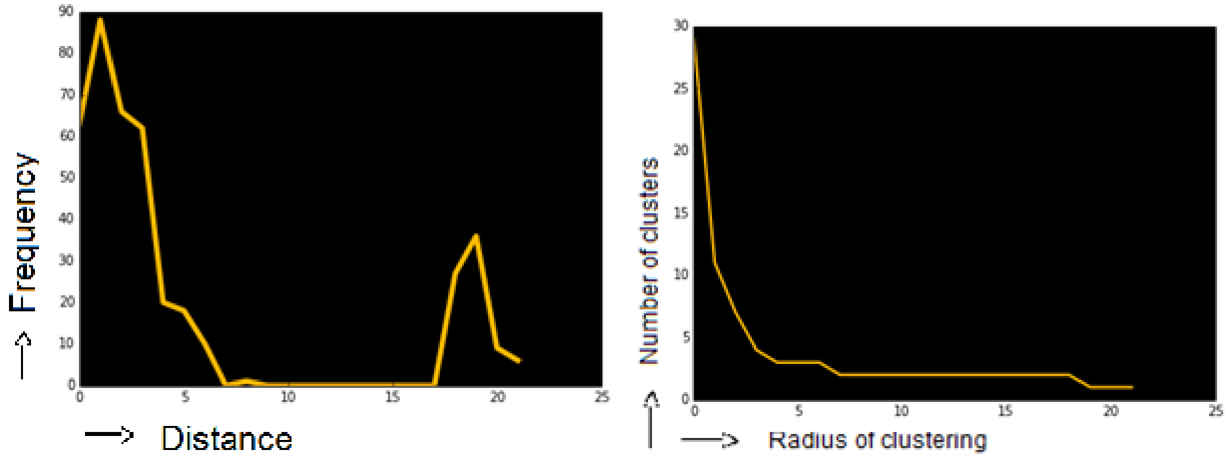


Fig3.3: Shown on left is the FDp which is distinctly bi-modal. On the right, the Hierarchy plot gives one distinct step corresponding to two clusters.

Clustering parameter, X = 1.741
Only 1 significant steps (with pvalue> 0.05);

Step number = 2 Corresponding radius = 7 Number of clusters = 2 Pvalue = 4.332e-10		

Cluster Details	Cluster number = 1
	Size = 26
	Spread = 8
		Cluster number = 2
		Size = 3
		Spread = 6

List of members	#Mystus_Bleekeri_Shirur #Mystus_Bleekeri_Shirus #Mystus_Bleekeri_MULA #JN228943_Mystus_bleekeri	#Mystus_Bleekeri_Wai #Mystus_Bleekeri_Amboli_Ajra #Mystus_Bleekeri_Wai

	#JN228944_Mystus_bleekeri #JN228945_Mystus_bleekeri #JN628898_Mystus_bleekeri #JN628899_Mystus_bleekeri #JN628901_Mystus_bleekeri #JN628904_Mystus_bleekeri #JN628928_Mystus_bleekeri #JX260916_Mystus_bleekeri #JX260917_Mystus_bleekeri #JX260918_Mystus_bleekeri #JX983370_Mystus_bleekeri #JX983371_Mystus_bleekeri #JX983372_Mystus_bleekeri #JX983373_Mystus_bleekeri #JX983374_Mystus_bleekeri #JX983375_Mystus_bleekeri #JX983376_Mystus_bleekeri #KF824794_Mystus_bleekeri #KF824795_Mystus_bleekeri #KF824796_Mystus_bleekeri #KF824797_Mystus_bleekeri	
Conclusion: The results of our analysis are an exact match to the predictions by Neelesh and Aniket; Wai and Ambodi clusters form a separate distinct cluster.		

3.3.2 Cavasius and Sengtee:

Data used: Cox1 gene sequences from 36 individuals belonging to the species Mystus cavasius and Mystus seengtee. Neelesh predicts that cavasius and seengtee are actually the same species

Result:

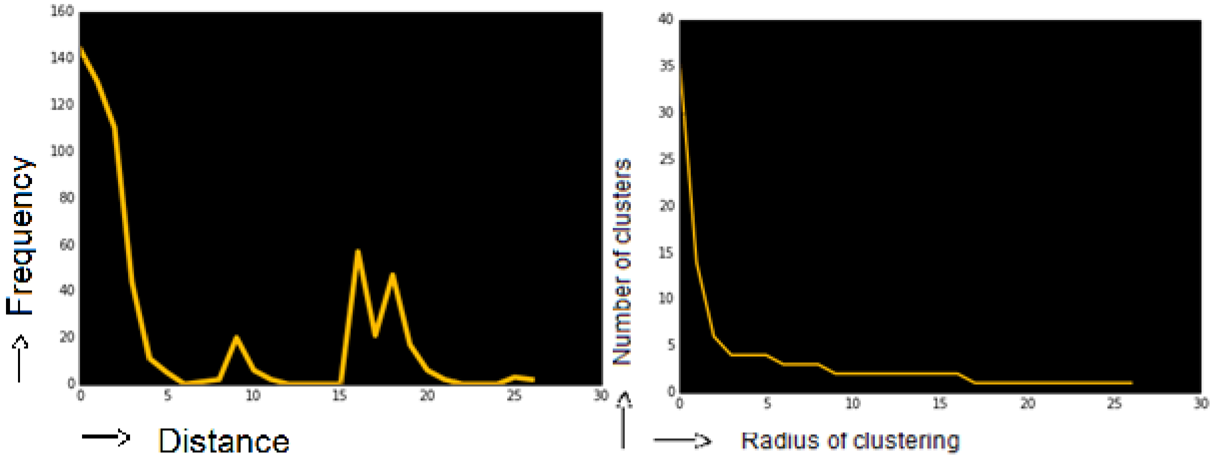


Fig3.4: On the left is shown FDp which is bi-modal. On the right is the Hierarchy plot which gives one significant step.

Clustering parameter, X = 1.475

Details of the significant steps (with pvalue> 0.05);

Step number = 3 Corresponding radius = 9 Number of clusters = 2 Pvalue = 1.410e-05		

Cluster Details	Size = 31
	Spread = 11
		Size = 5
		Spread = 7

List of members	#Mystus_seengtee_Bhima(MBS1a)
	#Mystus_Seengtee_Bhima(MBS2)
	#Mystus_cavacious_Kerala(MSKa1)
	#Mystus_Seengtee_Bhima(MBS6)
	#Mystus_Seengtee_Bhima(MBS5)
	#Mystus_Seengtee_Yerwada(Msy2)
	#Mystus_Cavacious_Kerala(Dadar)
	#Mystus_seengtee_Bhigwan(Msu1)
	#Mustus_Seengtee_bhigwan(msb2)
	#Mystus_Seengtee_Yerwada(Msy3)
	#Mystus_cavacious_Ganga(MCg)
	#Mystus_Seengtee_Mangaon
	#Mystus_cavcious_Kerala(MC1)
	#Mystus_Seengtee_Panvel(MSPS)
	#Mystus_Seengtee_Bhigwan(Msb1)
	#Mystus_Seengtee_Ujjani(Msu2)
	#Mystus_Cavacious_Solapur(Dadar)
	#Mystus_seengtee_Bhima(MBS3)
	#Mystus_Cavacious_Solapur(dadar)
	#Mystus_Seengtee_Bhima(MBS8)
	#Mystus_Seengtee_Yerwada(MSY)
	#Mystus_cavacious_Solapur(Dadar)
	#Mystus_Seengtee-Bhima(MBS7)
	#JX260919_Mystus_cavasius
	#JX983377_Mystus_cavasius
	#JX983378_Mystus_cavasius
	#JX983379_Mystus_cavasius
	#JX983380_Mystus_cavasius
	#JX983381_Mystus_cavasius
	#JX983382_Mystus_cavasius
	#JX983383_Mystus_cavasius
		#JN228946_Mystus_cavasius
		#JN228947_Mystus_cavasius
		#JN228948_Mystus_cavasius
		#JN628905_Mystus_cavasius
		#KF742435_Mystus_cavasius

Conclusion: The result of our analysis puts most of the individuals belonging to cavasius and seengtee as the same species. Analysis shows only one significant levels of hierarchy corresponding to a group of 5 cavasius individuals as a separate cluster.

3.3.3 Malbaricus:

Data used: Cox1 gene sequences from 28 individuals belonging to the species Mystus malbaricus. Neelesh predicts that malbaricus is a species complex having 4 sub-clusters

Results:

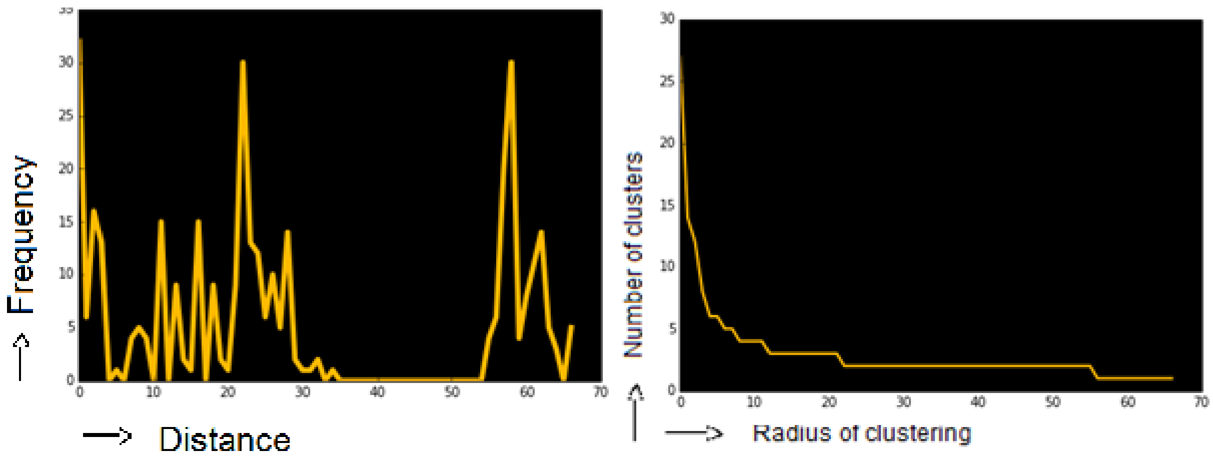


Fig3.5: FDp, on the left, is multi-modal while the Hierarchy plot on the right gives 2 distinct steps in the cluster plot

Clustering parameter, X = 1.675
2 significant step details (with pvalue<0.05) ;

Step number = 4 Corresponding radius = 12 Number of clusters = 3 Pvalue = 0.024			

Details of 3 clusters	Size = 5
	Spread = 2
		Size = 15
		Spread = 21
			Size = 7
			Spread = 8

List of Members	#HQ219109_Mystus_malabaricus #HQ219110_Mystus_malabaricus #HQ219111_Mystus_malabaricus	#Mystus_Malbaricus_Patan(MSPP2) #Mystus_Malbaricus_Amboli_Ajra #Mystus_Malbaricus_Barapole	#Mystus_malabaricus_Bhima #Mystus_Malbaricus_Bhima #Mystus_Malbaricus_BHIMA

	#HQ219112_Mystus_malabaricus
	#HQ219113_Mystus_malabaricus
	
	(Corresponds to Neelesh's prediction of cluster M1)
		#Mystus_malbaricus_Tunga(MF4)
		#Mystus_Malbaricus_Amboli_Ajra
		#Mystus_Malbaricus_Bhagamandala
		#Mystus_malbaricus_Chandechiwadi
		#Mystus_Malbaricus_Patan(MSPP1)
		#Mystus_Malbaricus_Patan
		#Mystus_malbaricus_Satara(mss1)
		#Mystus_Malbaricus_Wai(Msw3)
		#Mystus_Malbaricus_Satara(mss4)
		#Mystus_Malbaricus_Satara(MSS3)
		#Mystus_Malabaricus_Satara(MSS2)
		#DQ508092_Mystus_malabaricus
		
		(Corresponds to Neelesh's prediction of cluster M3 and M4 combined)
			#Mystus_Malbaricus_Lonavala
			#Mystus_Malbaricus_Mangaon
			#Mystus_Malbaricus_Mangaon
			#Mystus_Malbaricus_Phansad
			
			(Corresponds to Neelesh's prediction of cluster M2)

Step number = 5
Corresponding radius = 22
Number of clusters = 2
Pvalue = 2.022e-08

Details of 2 clusters	Size = 5
	Spread = 2
		Size = 22
		Spread = 34

List of members	#Mystus_malabaricus_Bhima(MMPH1)
	#Mystus_Malbaricus_Patan(MSPP2)
	#Mystus_Malbaricus_Amboli_Ajra(MAB)
	#Mystus_Malbaricus_Barapole(MBS)
	#Mystus_malbaricus_Tunga(MF4)
	#Mystus_Malbaricus_Amboli_Ajra(MAA)
	#Mystus_Malbaricus_Bhima(MMPH2)
	#Mystus_Malbaricus_Bhagamandala(MBB)
	#Mystus_malbaricus_Chandechiwadi(MCI)
	#Mystus_Malbaricus_Patan(MSPP1)
	#Mystus_Malbaricus_Patan(MSPP2a)
	#Mystus_Malbaricus_BHIMA(MMPH2a)
	#Mystus_Malbaricus_Lonavala(MM1)
	#Mystus_malbaricus_Satara(mss1)
	#Mystus_Malbaricus_Wai(Msw3)
	#Mystus_Malbaricus_Satara(mss4)
	#Mystus_Malbaricus_Mangaon(MMM1)
	#Mystus_Malbaricus_Mangaon(MMM2)
	#Mystus_Malbaricus_Phansad(MMPS)
	#Mystus_Malbaricus_Satara(MSS3)
	#Mystus_Malabaricus_Satara(MSS2)
	#DQ508092_Mystus_malabaricus
		#HQ219110_Mystus_malabaricus
		#HQ219111_Mystus_malabaricus
		#HQ219112_Mystus_malabaricus
		#HQ219113_Mystus_malabaricus

Conclusion: Based on the high value of clustering parameter X, SSC predicts that Malbaricus is indeed a species complex. From the Hierarchy plot, we get the first significant level of clustering corresponding to 3 clusters, SSC predicts these clusters should be called as species. We get another significant level of clustering corresponding to 2 clusters, based on SSC, this two clusters should be considered for the classification at the genera level.

The above examples demonstrate that more than one levels of significant clustering does exist in the living world and can be tested with sound statistical methods. In order to apply this analysis more widely there are two limitations. One is that greater computational power will be needed to handle wider groups and test the significance of family and other higher hierarchical levels. The other limitation is that unless there is multiple representation of every “species” or cluster the within and between cluster distances cannot be segregated. For many species only one are few sequences are available in the databases. There is an uploading bias currently in that if a sequence turns out to be new, the researchers are keen to upload it but if it turns out to be identical with an existing sequence, it is least likely to be uploaded. If the uploading bias can be avoided in future, the SSC can be applied widely across the living world.

CHAPTER 4: Simulation Speciation

4. Introduction

Quoting from the Wikipedia page on Speciation;

> "Speciation is the evolutionary process by which new biological species arise..
> .. Whether genetic drift is a minor or major contributor to speciation is the subject matter of much ongoing discussion"

Neutral drift as a hypothesis lacks much empirical support in terms of experiments performed in laboratory or in nature [Coyne & Orr 2004]. We still don't know, if speciation is a very complex phenomenon involving interplay of many factors or are there a few fundamental factors like the neutral drift that essentially drive the process. This is because of two reasons majorly

i) Up till now, we did not have a proper definition for species i.e. if two populations keep drifting, can we say at any point they belong to two different species and,

ii) Experimental evolving populations of bacteria for several thousand generations have not reached what we can call distinct speciation (Lenski et al., 1991; Lenski and Travisano, 1994; Barrick et al., 2009)

By making use of computational models to simulate speciation, we can overcome some of these difficulties, since simulations, in general can be easily scaled to larger landscapes and longer time periods. Though there will always be a trade-off, as simulations at best present an approximation of the reality and often suffer from being too simplistic or biased.

In the following section, we first develop the computational models of

i. Neutral Drift,
ii. Patch-dynamics (discrete niche overlap)
iii. Competition (continuous niche overlap)

These will be used to simulate speciation in the case of asexually reproducing populations to understand some very fundamental questions;

I. Can speciation be explained by a Neutral Drift model alone?
II. In microbial environments, does Patch-dynamics facilitate speciation?
III. Is competition *necessary* or a *sufficient* condition for speciation to occur?
IV. If the answer to the question III is positive, how the relationship between genetic distance and competitive forces affects speciation?

The SSC algorithm developed in chapter 1 is used for the objective quantification of the clustering/diversification events observed in the simulations. We also made a visual simulation for the three models mentioned above (not a part of the thesis).

4.1 Building the Computational Model

For all the simulations, we consider an asexually reproducing population with a fixed population size, N_{Tot}. Each individual is constructed as a binary array, which is a combination of zeros and ones. These zeros and ones represent the absence or presence of Traits. The zeros and ones are just used to code for the differences between individuals and can represent the genetic code or any other characters.

Distance between two individuals is the number of differences between the two arrays, which is a same as Hamming distance. So for example if we have three individuals with 5 traits each,
A: 00001
B: 00101
C: 10000
So the distance matrix would look like,

	A	B	C
A	0		
B	1	0	
C	2	2	0

Each of the three models that follow, incorporate number of Generations G as a parameter, where each generation comprises of three fundamental processes; Reproduction, Mutation and Dying. After each generation, the population size comes back to N_{Tot} and is ready to go for another generation. After each generation, we can check the population for its composition by looking at the value of the clustering parameter X. The graph of X versus time would be used as the observation for each simulation. Like for example,

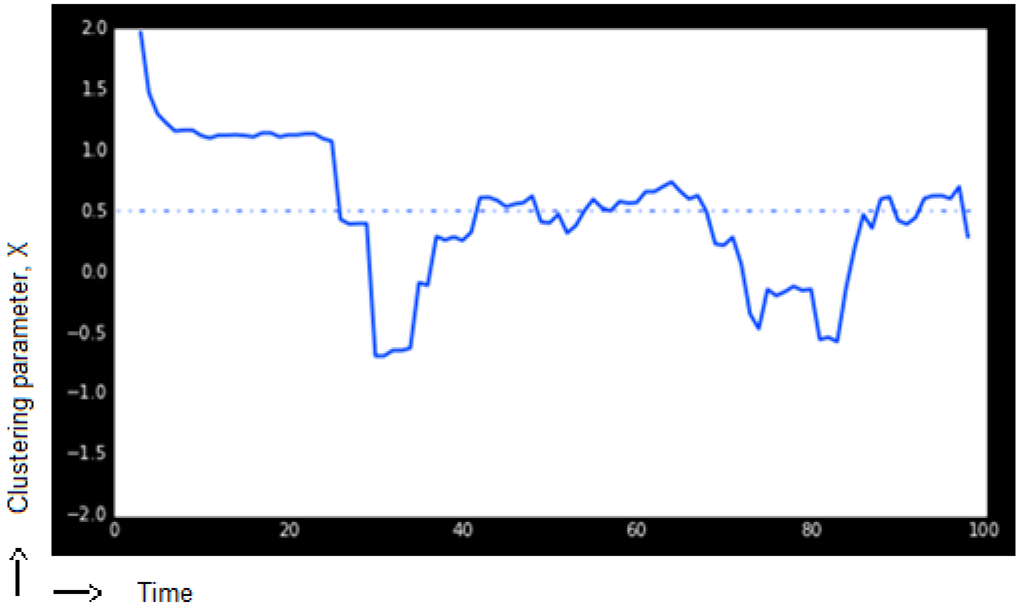


Fig4.1: Shown above is the graph for the value of clustering parameter X on the y axis versus the number of generations on the x axis. Above shown is a classic case of neutral drift, the simulation starts with all identical individuals, which diverge as is shown by the decrease in the value of X.

One problem is that, though by means of SSC we have a criterion for delimiting species, we still don't have a definition of Speciation in terms of time. It is very much engrained in the minds of biologists, that an even t of speciation also means that the separated individuals remain 'sufficiently' separated for 'sufficiently' long times. SSC tells us what is 'sufficiently' separated, but there exists no objective definition of what is 'sufficiently' long. By-passing this need for a cut-off criteria for what is 'sufficiently' long, In the following simulations we measure the how long is a clustering stable. This can allow us to compare between simulations and hence understand which factors facilitate speciation and which don't.

To summarize, while looking for a signature of a speciation event, in the graph for clustering parameter we look for,

- Higher values of clustering parameter X, well above the value of 0.5, preferably between 1 and 2.
- The stability of the value of X obtained. Like for example, getting a high value of X for a very short span of generations indicates a group of individuals diverging, followed by either converging back to main population or dying. Both of these cases in the absence of stability do not indicate a speciation event.

The following are two graphs to show two contrasting pictures, the first one with no signature of a speciation event, and the second one with clear indications of speciation events.

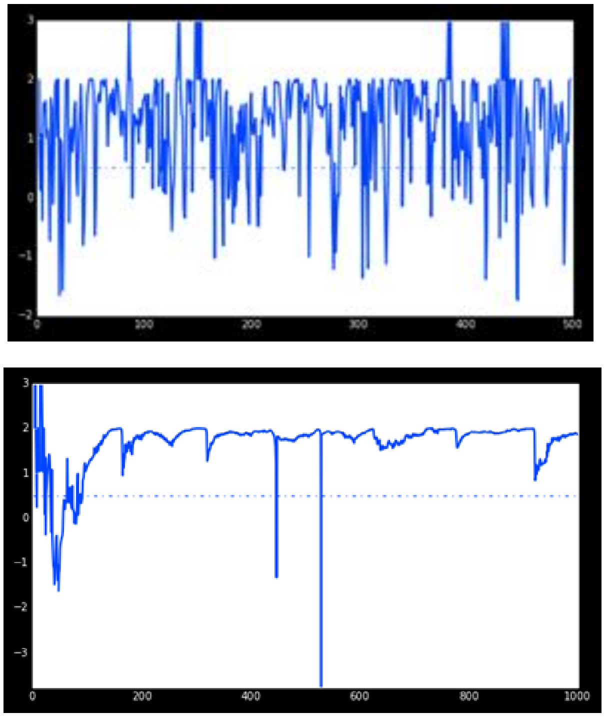


Fig4.2: The graphs are for the X versus number of generations in different simulations. The above one shows a highly fluctuating unstable value of X and while the one below shows a stable value of X around higher values between 1 and 2.

To compare for the extent of speciation between simulations, we look at the maximum number of generations for which value of X is stable around higher values. So for example in the first graph above the maximum number of generations with high values of X is 100 while of the graph below is 400. When considered for statistically significant timescale, this value reflects the probability of occurrence of speciation.

Some of the acronyms that will be used are;
T = number of traits for each individual, size of the array comprised of zeros/ones
N_{Tot}= total number of individuals, fixed size of the population
G = number of generations
R_R = rate of reproduction
M_P = mutation probability
T-f = number of timeframes, which is the number of times a population is checked for its composition.

For example, if G = 10 and T-f = 100, then the population is checked for its composition after every 10 generations. So in the graphs presented for each simulation, the value of X will be calculated at the end of G number of generations for T-f number of times.

Also to mention, since in our model population size is kept fixed i.e. the number of individuals dying is equal to the number of new individuals being born, a lower rate of reproduction also implies a lower rate of extinction.

4.2 Neutral Drift

The following model is used for simulating Neutral Drift. In each generation, we start with a fixed number of individuals N_{Tot}, some individuals reproduce, out of which some mutate and then at last some individuals are killed/deleted randomly so that we get back the fixed population size of N_{Tot} . One assumption that we use is that the number of mutations are proportionate to the number of traits each individual has.

The Model:

Three key processes are simulated;

I. Reproduction

- Let R_0 individuals reproduce.
- $R_0$is a value chosen as a number from a random normal distribution around the value $R_R x N_{Tot}$, where R_R is the reproduction probability and N_{Tot} is the total number of individuals
- From the N_{Tot} individuals choose R_0 and make their copies such the population size is now increased to $N_{Tot} + R_0$

II. Mutation

- Let M_0 individuals mutate.
- $M_0 = M_P \times T \times R_0$, where R_0 is the number of individuals reproducing, T is the number of traits for each individual and M_P is the mutation probability. These factors are multiplied as M_0should be proportionate to the number of individuals reproducing and dependence on T is an assumption of this model.
- Out of the newly made R_0 copies choose M_0 of them. Randomly choose a particular trait for each and mutate that trait i.e. change it to 1 if it was 0 and 0 if it was 1.

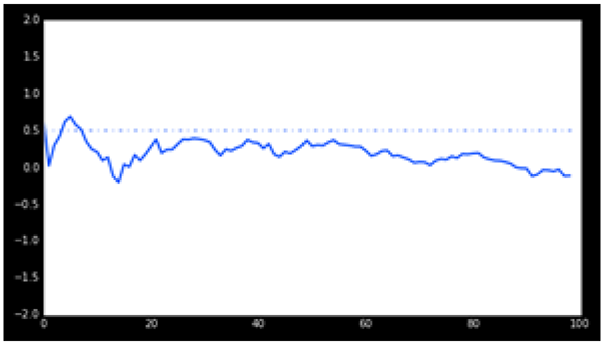
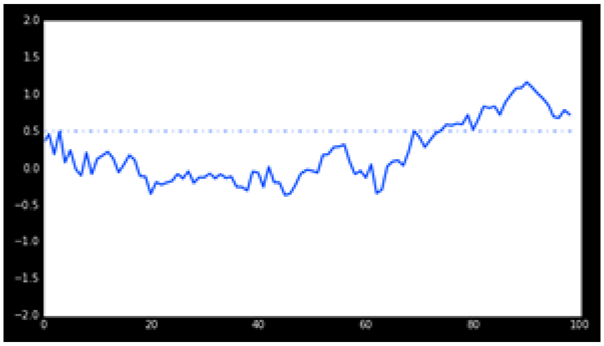
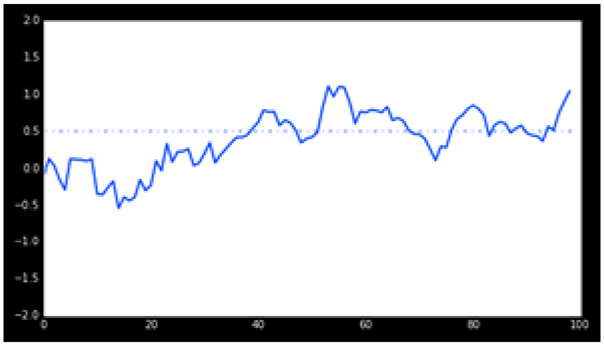
III. Dying

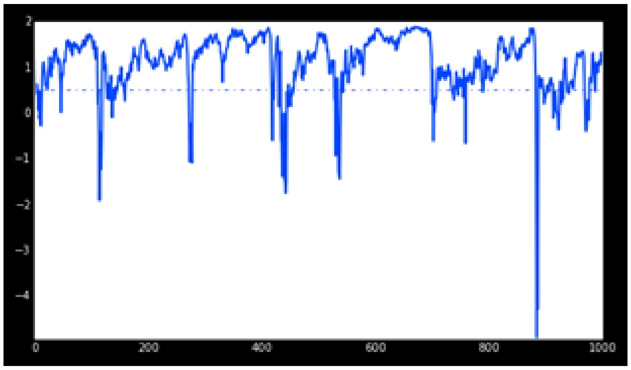
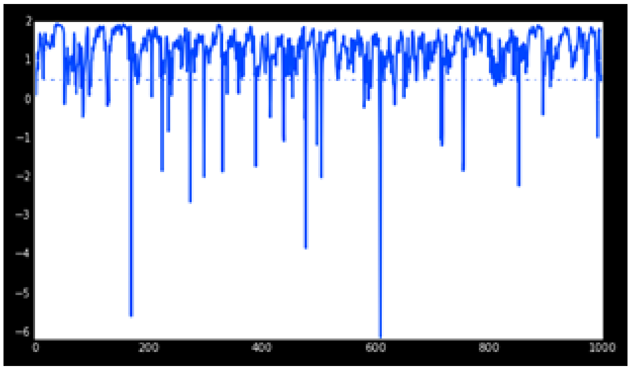
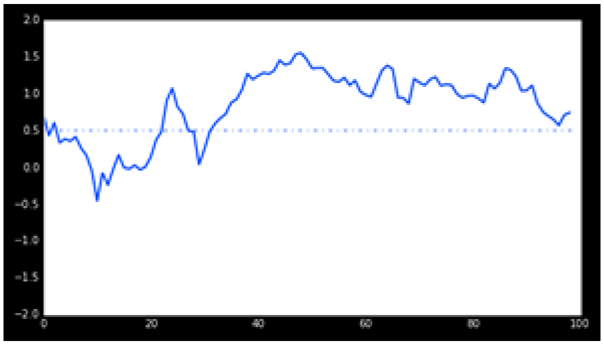
- Randomly choose R_0 out of $N_{Tot} + R_0$ and delete/kill them, such that the total population size remains fixed at N_{Tot}.

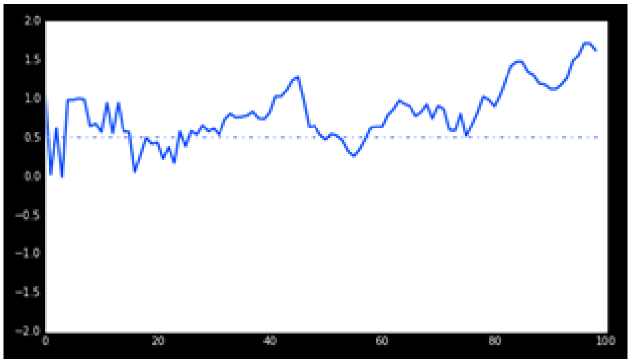
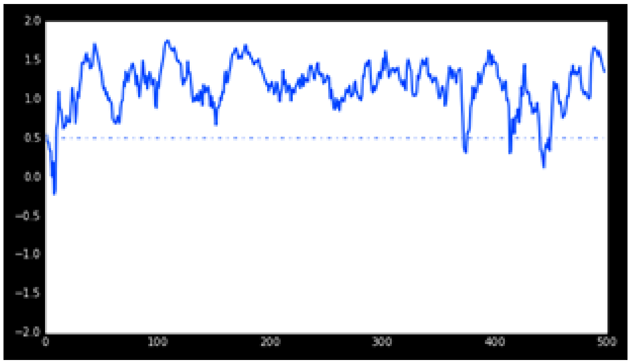
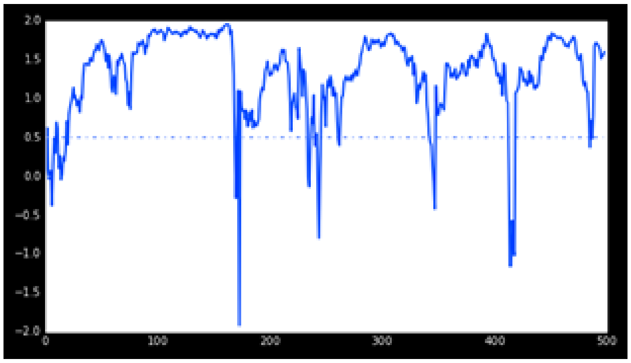
Do this cycle (I-II-III) for each generation, G number of times

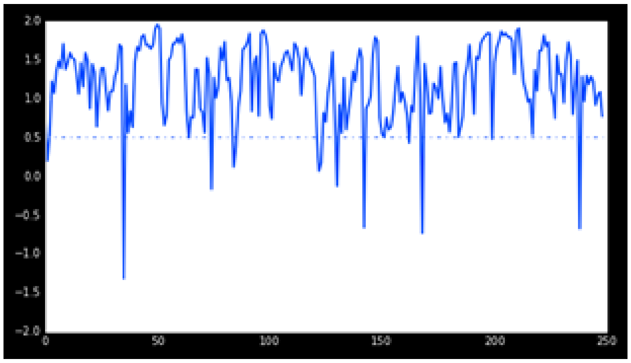
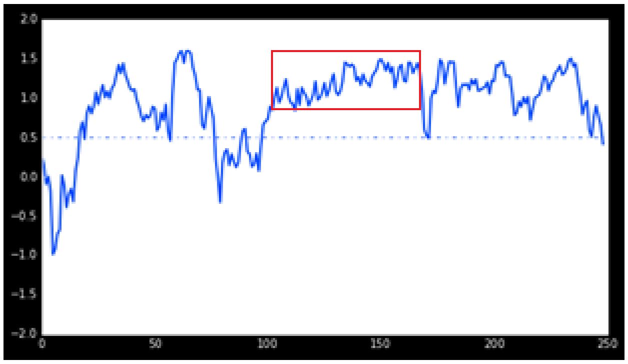
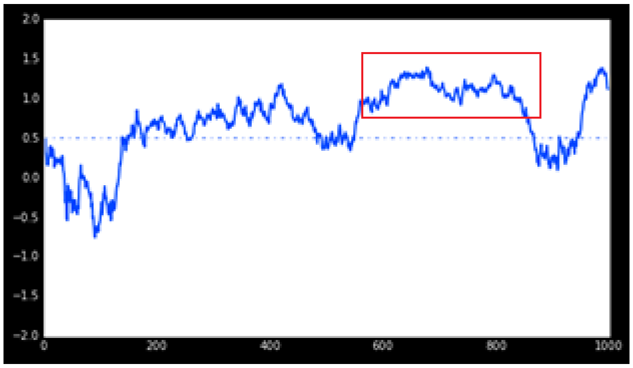
Results of the Simulations for Neutral Drift:

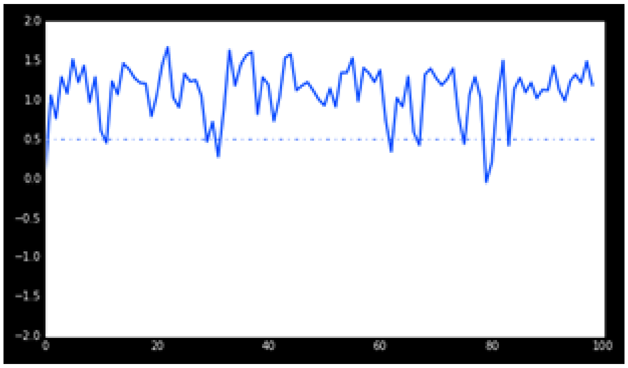
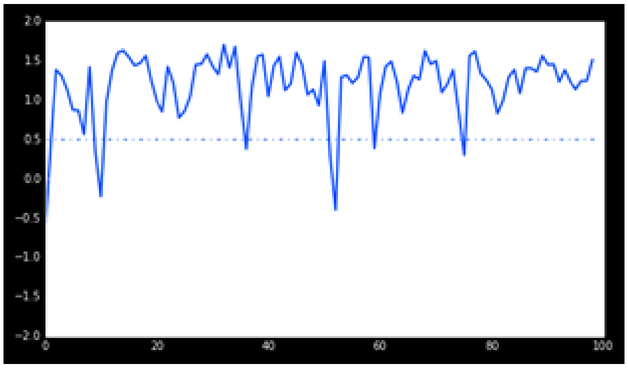
(Parameters that are varied and compared between simulations are shown highlighted in orange)

Parameters	Clustering parameter plot	Remarks/discussion
1) Testing for the effect of change in the rate of Reproduction R_R		
1a) T = 100 N_{Tot} = 100 G = 1 R_R = 0.1 M_P = 0.01 T-f = 100		
1b) T = 100 N_{Tot} = 100 G = 1 R_R = 0.5 M_P = 0.01 T-f = 100		Compared to 1a) no significant change observed at this timescale
1c) T = 100 N_{Tot} = 100 G = 1 R_R = 2.5 M_P = 0.01 T-f = 100		Comparing with 1a) this confirms that atleast on shorter time-scales there is no significant effect of greater reproduction probability

1d) T = 100 N_{Tot} = 100 G = 10 R_R = 0.1 M_P = 0.001 T-f = 1000		
1e) T = 100 N_{Tot} = 100 G = 10 R_R = 2.5 M_P = 0.001 T-f = 1000		Compared to 1d) For greater time-scales, the effect of a greater rate of reproduction rate does increase the fluctuations in the value of X and hence reduces the probability of a stable clustering.
Conclusion: For longer timescales, the effect of a large R_R is to reduce the stability of speciation.		
2) Testing for the effect of change in the mutation probability M_P		
2a) T = 100 N_{Tot} = 100 G = 10 R_R = 0.1 M_P = 0.01 T-f = 100		

2b) T = 100 N_{Tot} = 100 G = 10 R_R = 0.1 M_P = 0.001 T-f = 100		Compared to 2a) no significant change
2e) T = 500 N_{Tot} = 100 G = 10 R_R = 0.1 M_P = 0.01 T-f = 500		
2f) T = 500 N_{Tot} = 100 G = 10 R_R = 0.1 M_P = 0.001 T-f = 500		Compared to 2e) Slightly greater stability of clustering.
Conclusion: Upon reducing the mutation probability we see an increased stability of clustering values. This is consistent to what is expected of a neutral drift model.		
3) Testing for the effect of change in the Population size N_{Tot}		

3a) T = 100 N_{Tot} = 100 G = 100 R_R = 0.1 M_P = 0.001 T-f = 250		
3b) T = 100 N_{Tot} = 1000 G = 100 R_R = 0.1 M_P = 0.001 T-f = 250		Compared to 3a) *Expected:* By increasing the number of individuals, the probability of extinction of a cluster of individuals is reduced. Hence speciation is expected to be more stable *Result:* Stable clustering as a unique signature for speciation is observed. Shown with a red box, is the maximum number of generations for a stable value of higher values of X = 5000 generations
3c) T = 100 N_{Tot} = 1000 G = 10 R_R = 0.1 M_P = 0.001 T-f = 1000		Compared to 3a) Another simulation to confirm the above result. As shown inside the red box, this time the maximum number of generations for a stable value of higher values of X = 2500 generations
Conclusion: For an asexually reproducing population, with large population size, neutral drift alone is sufficient to cause speciation.		
4) Testing for the effect of change in number of Traits T		

4a) T = 100 N_{Tot} = 100 G = 100 R_R = 0.1 M_P = 0.01 T-f = 100		
4b) T = 5000 N_{Tot} = 100 G = 100 R_R = 0.1 M_P = 0.01 T-f = 100		Compared to 4a) *Result:* No significant difference observed.
Conclusion: As long as the number of traits are not too less, the number of traits do not affect the probability of speciation for this model of Neutral drift.		

To summarize the main point from the above simulations, under most of the scenarios Neutral Drift model does not give stable clustering, but with

- large population size N_{Tot}
- low mutation probability M_P,
- low reproduction/extinction rate R_R

the Neutral Drift model can give rise to stable clustering. This implies that the neutral drift model is, without the help of any natural selection, sexual selection or competition and other complex interactive forces, is sufficient to cause speciation.

4.3 Patch-Dynamics

Absence of the plausible candidate of reproductive isolation as a causal factor for speciation in asexually reproducing populations leads us to testing the next most likely candidate; Patch dynamics model (Lindström and Langenheder, 2011). In our simulations we seek to address the question; does patch-dynamics facilitate speciation?

The Model: This model basically explores the scenario of discrete niche separation. Imagine a species that exists in a region A in the multi-dimensional trait space. There are N_{Tot} number of individuals that reproduce, mutate and die. Now because of the drift, some of the individuals start moving towards region A_1, some towards region A_2, some towards A_3. . .and some to A_P, where P is the total number of patches. It is often seen in microbial species that individuals mostly survive and reproduce in patches, come close to each other, interact and then disperse back to different patches. We want to explore can under such conditions do we see a signature of stable clustering. Each of these regions A_1, A_2, A_3. . .A_P are simulations of patches. We start with a patch A comprised of one single cluster. Then allow this cluster to form patches $A \rightarrow A_1+A_2+. .+A_P$. From this set of P patches, we randomly sample N_{Tot} number of individuals and consider this set of individuals as the starting patch A` for the next iteration. Consider the following illustration;

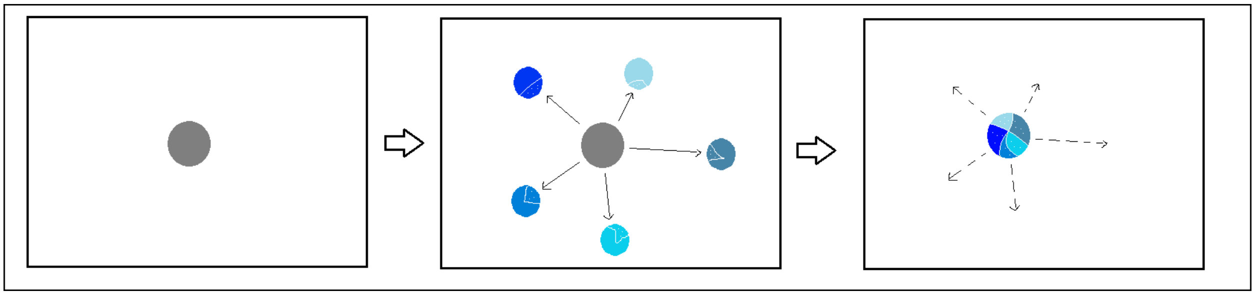


Fig: Shown above is an illustration for the Patch-dynamics model. We start with the main patch which is shown in grey in the first box. Then this patch drifts to 5 patches shown in different colors. Finally individuals from the colored patches are sampled and pooled back to form the main patch for the next cycle.

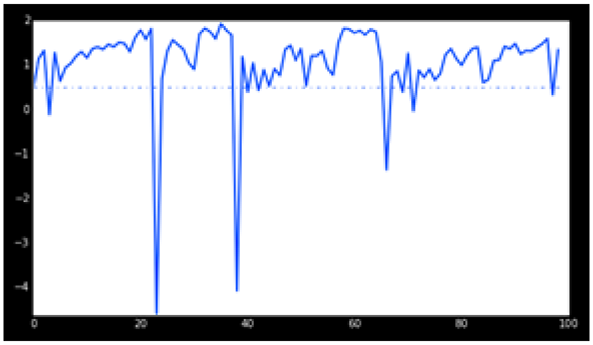
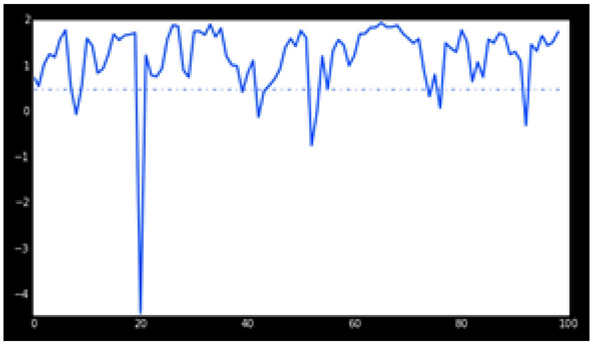
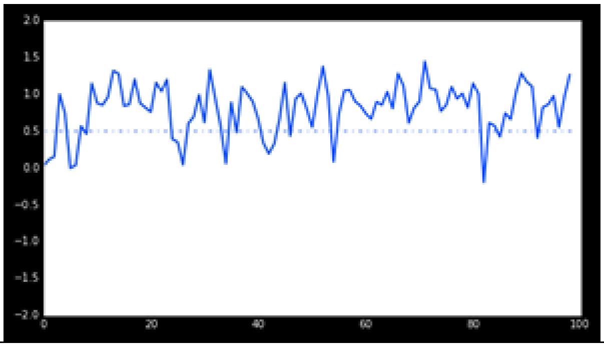
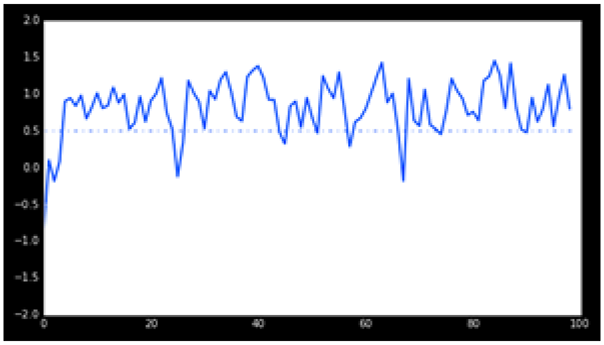
The algorithm used for the code is as follows;

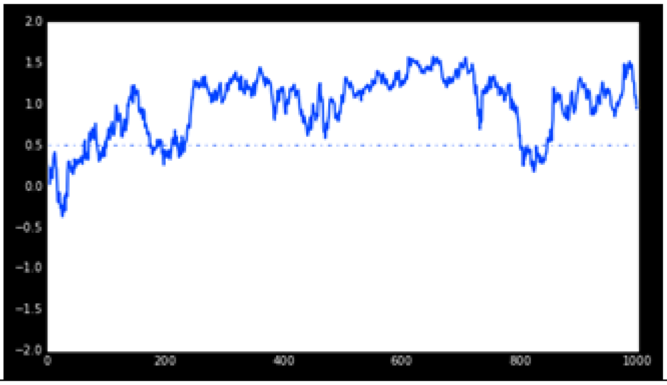
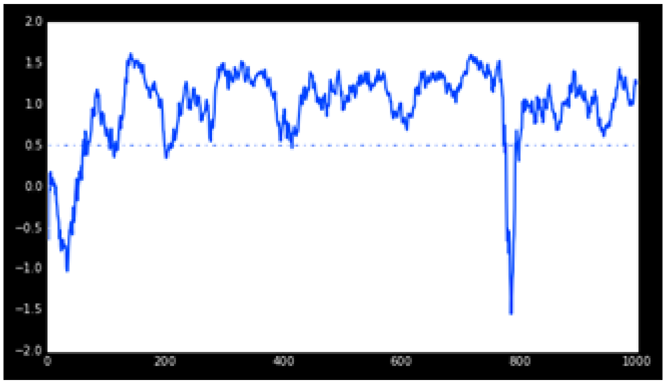
1. Start with N_{Tot} identical individuals.
2. Choose a random number between 0 and P_{max}, = p for the number of patches
3. Like the Neutral drift model, do the three steps of Reproduction, Mutation and Dying for G_{Patch} number of generations. At the end call these set of individuals as A_1
4. Repeat 3) for A_2, A_3. . .A_P
5. From A_1, A_2, A_3. . .A_P randomly sample N_{Tot} number of individuals and with these repeat procedure 1-5 for $G/G_{Patches}$, where G is the total number of generations per time-frame.

There are two new parameters for the Patch-dynamics simulations;
P_{Max} = is the maximum number of patches A_1, A_2, A_3. . .A_P
G_{Patch} = is the number of generations each patch survives for before it is pooled back into the main patch.

Results of the Simulations for Patch-dynamics:

Parameters	Clustering parameter plot	Remarks/discussion
Comparison to Neutral Drift		

5a) Neutral Drift T = 100 N_{Tot} = 100 G = 100 R_R = 0.1 M_P = 0.001 T-f = 100		
5b) Patch Dynamics T = 100 N_{Tot} = 100 G = 100 R_R = 0.1 M_P = 0.001 T-f = 100 P_{Max} = 5 G_{Patch} = 10		Compared to 5a) At this and at all lower time-scales, we found no statistical difference between the results of Neutral drift and Patch-dynamics model
5e) Neutral Drift T = 200 N_{Tot} = 200 G = 100 R_R = 0.1 M_P = 0.01 T-f = 100		
5f) Patch dynamics T = 200 N_{Tot} = 200 G = 100 R_R = 0.1 M_P = 0.01 T-f = 100 P_{Max} = 3 G_{Patch} = 25		Compared to 5e) No significant change.

5i) Neutral Drift T = 100 N_{Tot} = 1000 G = 10 R_R = 0.1 M_P = 0.001 T-f = 1000		
5j) Patch dynamics T = 100 N_{Tot} = 1000 G = 10 R_R = 0.1 M_P = 0.001 T-f = 100 P_{Max} = 5 G_{Patch} = 20		Compared to 5i) Still no significant change observed.
Conclusion: After looking at different permutation and combination of number of patches P_{Max} and the numberof generations per patch, G_{Patch}, we did not find any significant change in the probability of speciation caused by Patch-dynamics model when compared to the Neutral drift model. Patch-dynamics model in asexually reproducing organisms does Not facilitate speciation.		

4.4 Competition

Competition as theoretical model has been a strong contender amongst the factors that cause speciation in asexually reproducing populations (Rosenweig, 2008; Polechova and Barton, 2005). From the results of the simulations of the Neutral drift model, we now know that it is not necessary to cause speciation, but does it facilitate speciation? Competition decreases as a function of distance between the individuals, but what is the functional form of this dependence? These are the questions that we seek to address by means of computational simulations.

The Model:

The essential model is the same as Neutral Drift for the Reproduction and Mutation part but for the Dying part, individuals are not chosen randomly for Dying but are chosen based on a model of competition. Those individuals who are the farthest from everyone else are chosen. For this we take the sum of each individuals distance to every other individual and construct the distance sum array in the following way; For example if A, B and C are 3 individuals with the distance matrix;

	A	B	C
A	0		
B	1	0	
C	3	2	0

Distance sums	4	3	5

The distance sums for A, B, C are 4,3 and 5 respectively. This 4,3,5 is the distance array which is sorted to give the individuals which are most distant from others; C followed by A then B. In the simulations we check for three types of functional dependence on distance;

a) Concave

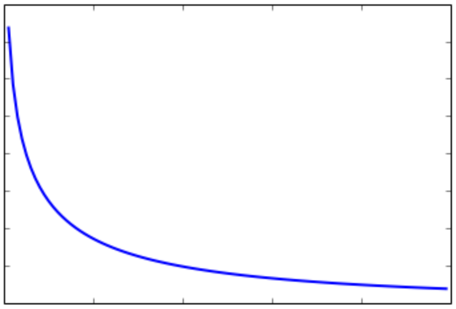


Fig4.3: Where x axis is representative of the genetic distances and on y is the strength of the competitive forces. We have used the exponentially decreasing function, $Y = -e^{-x^{0.3}},\ 0 < x < \infty$

b) Linear

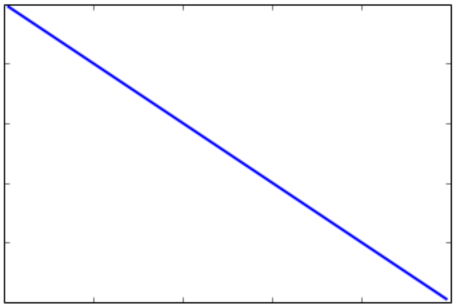


Fig4.4: Where x axis is representative of the genetic distances and on y is the strength of the competitive forces. We have used the linearly decreasing function, $Y = -\mathrm{x}$

c) Convex

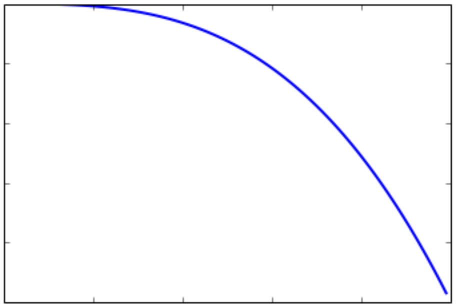
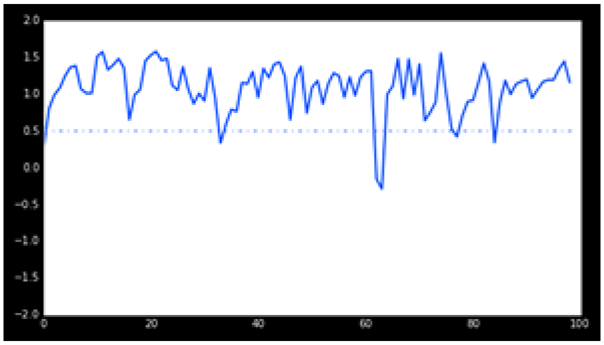
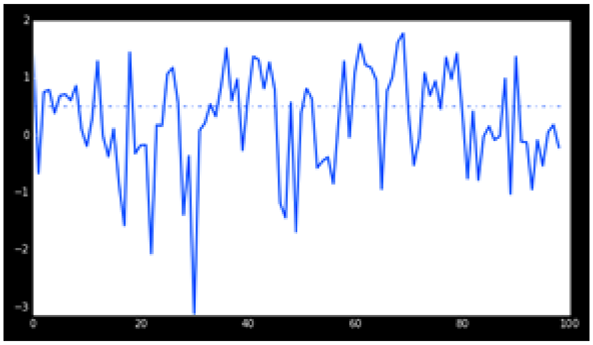


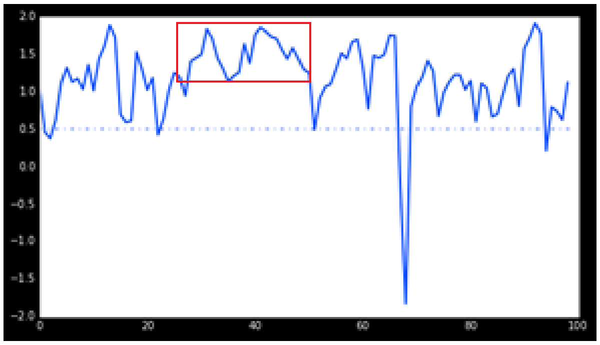
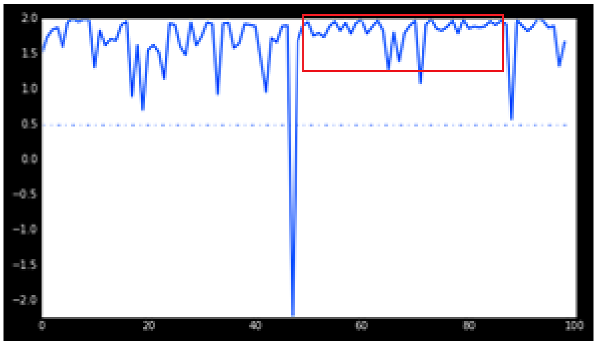
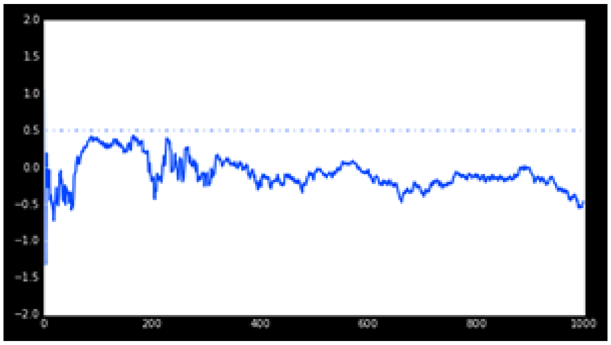
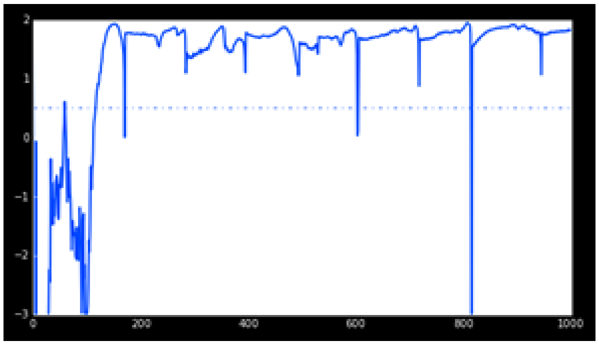
Fig4.5: Where x axis is representative of the genetic distances and on y is the strength of the competitive forces. We have used the decreasing function, $Y = -x^3$

Finally the distance array is sorted so it does not matter what the actual numbers are. Then the first N_{Tot} individuals with the least distances to other individuals are chosen and the remaining die.

Now, we first compare the results of a competition model using the concave decreasing function against the neutral drift model

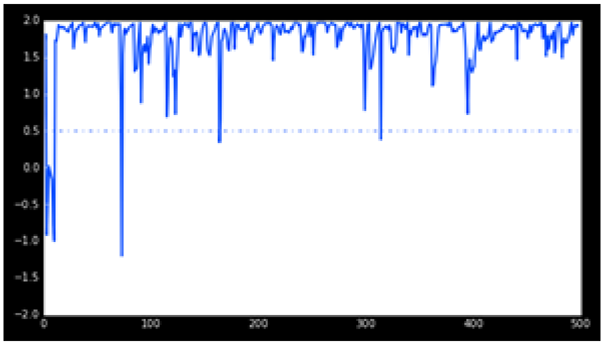
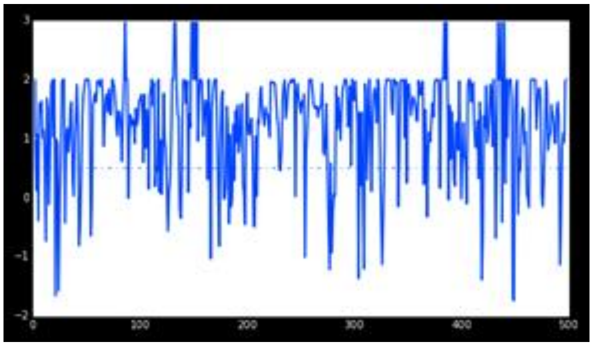
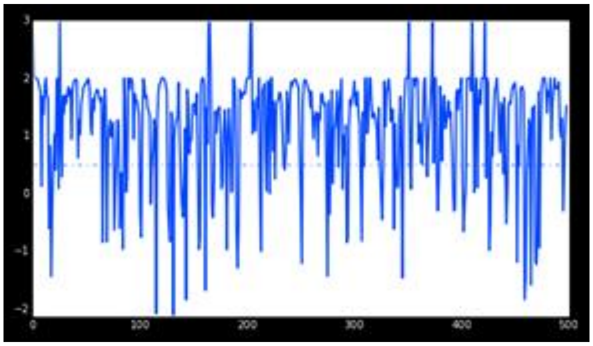
Results of the Simulations for Competition Model:

Parameters	Clustering parameter plot for Neutral drift	Clustering parameter plot for Competition
6a) $T = 100$ $N_{Tot} = 100$ $G = 100$ $R_R = 0.1$ $M_P = 0.01$ T-f = 100		 No significant clustering with high mutation rate of 0.01

6b) T = 100 N_{Tot} = 100 G = 100 R_R = 0.1 M_P = 0.001 T-f = 100	 The maximum number of values of X for which it is stable in the range 1-2 = 2500	 Upon reducing the mutation rate , we can see clustering is more stable in the competition model. The maximum number of values of X for which it is stable in the range 1-2 = 4000
6c) T = 100 N_{Tot} = 1000 G = 1 R_R = 0.1 M_P = 0.001 T-f = 1000	 A stark contrast to the competition model, high values of X were never reached.	 Though initially it takes some time to settle down, but once a high value of X is reached, it is stable for very long times.
Conclusion: The competition model of speciation, though not necessary as we have seen from the results of the Neutral drift model, adds substantially to the stability of speciation.		

Next, we compare the results of the competition model for Concave, Linear and Convex decreasing functions of distance.

Parameters	Clustering parameter plot
Comparison between different functions of distance	

7a) With a Concave function of distance T = 100 N_{Tot} = 100 G = 10 R_R = 0.1 M_P = 0.001 T-f = 500	
7b) With a Linear function of distance T = 100 N_{Tot} = 100 G = 10 R_R = 0.1 M_P = 0.001 T-f = 500	
7c) With a Convex function of distance T = 100 N_{Tot} = 100 G = 10 R_R = 0.1 M_P = 0.001 T-f = 500	

Discussion: As is clearly seen in the graphs above, stable clustering is possible only with a convex function of distance. To present a hypothesis on why this might be happening, I would like to draw analogy from physics; competitive forces in ecology are, like the strong nuclear force, a short-ranged force. By short-ranged it is meant that after some short distance, the affects of competition on fitness/survival probability of an individual fall exponentially. For example lions of the Gir forest region in Gujarat face competition of the jackals, hyenas and leopards of that region but there is hardly any competition with the birds of that region, birds being a group that is farther from the lions in the multi-dimensional space of characters.

Results from the Thesis:

- We propose a novel Statistical Species Concept (SSC) which is based on the Frequency Distribution plot (FDp) of the distances between the individuals under study.
- Using synthetic sequences, we compare the results of SSC to the phylogenetic reconstruction method - Maximum Likelihood (ML). We find,
  - Out of the 100 cases we analysed, ML performed satisfactorily 31 times and SSC 94 times.
  - SSC is shown to perform significantly better than ML with regard to false negatives and is capable of accepting or rejecting the null hypotheses of there being no clustering and no hierarchy in the data.

- Applying the SSC algorithm to real genetic data we find,
  - For data on Homo sapiens, the genetic distances between people of different ethnicities and geographical regions give a distinct uni-modal FDp. This is compatible with our concept that the first level of distinct clustering should define species.
  - For data from the family Homonidae, there is clear demonstration that not only species are natural clusters but higher level genera are also natural clusters.
  - Data on fish also reveals two significant levels of hierarchy indicating that not only species but genera also could be natural and not imposed by a taxonomist.

- We created a platform for understanding Speciation by the means of simulations on computational models. Some of the preliminary results in case of asexually reproducing populations are
  - For large population sizes, we find that a Neutral Drift model alone is sufficient to cause speciation.
  - Patch-dynamics model does not facilitate speciation.
  - Competition model adds substantially to the stability of speciation.
  - In the competition model, the dependence of the competitive forces as a function of distance should be 'concave decreasing'. The convex and the linear decreasing functions do not give any stability of clustering.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., … Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*(2), 229–246. doi:10.1111/j.1420-9101.2012.02599.x

Agapow, P.-M., O. R. P. Bininda-Edmonds, K. A. Crandall, J. L. Gittleman, G. M. Mace, J. C. Marshall, and A. Purvis. (2004). The impact of species concept on biodiversity studies. Q. Rev. Biol. 79:161–179.

Ax, P. (1987).*The Phylogenetic System*. Wiley and Sons, New York.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., … Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, *461*(7268), 1243–7. doi:10.1038/nature08480

Carr, S. M., Kivlichan, D. S., Pepin, P., & Crutcher, D. C. (1999). Molecular systematics of gadid fishes : implications for the biogeographic origins of Pacific species, *26*, 19–26.

Coenye T, Vandamme P (2003). "Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes". *FEMS Microbiol.Lett.* **228** (1): 45–49. doi:10.1016/S0378-1097(03)00717-1. PMID 14612235

Coyne, J. A., and H. A. Orr. (2004). Speciation. Sinauer, Sunderland, MA.

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology*, *56*(6), 879–86. doi:10.1080/10635150701701083

Eaaswarkhanth, M., Haque, I., Ravesh, Z., Romero, I. G., Meganathan, P. R., Dubey, B., … Thangaraj, K. (2010). Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. *European Journal of Human Genetics : EJHG*, *18*(3), 354–63. doi:10.1038/ejhg.2009.168

Felsenstein J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution***39**:783-791.

Griffiths,G. (1976).“The Future of Linnaean Nomenclature,” *Systematic Zoology* 25:168–73.

Hausdorf, B. (2011). Progress toward a general species concept. *Evolution; International Journal of Organic Evolution*, *65*(4), 923–31. doi:10.1111/j.1558-5646.2011.01231.x

Hennig, W. (1966).*Phylogenetic Systematics*. University of Chicago Press, Chicago.

Hennig,W. (1969[1981]). *Insect Phylogeny*.Translated by A. C. Pont. John Wiley Press, New York. Originally published as *Die Stammesgeschichte der Insekten*, Waldemar Kramer, Frankfurt.

Isaac, N. J. B., J. Mallet, and G. M. Mace. 2004. Taxonomic inflation: its influence on macroecology and conservation. Trends Ecol. Evol. 19:464–469

Klaassen, C. a. J., Mokveld, P. J., & van Es, B. (2000). Squared skewness minus kurtosis bounded by 186/125 for unimodal distributions. *Statistics & Probability Letters*, *50*(2), 131–135.

Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations Richard E. Lenski, Michael R. Rose, Suzanne C. Simpson and Scott C. Tadler *The American Naturalist*, Vol. 138, No. 6 (Dec., 1991), pp. 1315-1341

Lenski, R. E., & Travisano, M. (1994). Dynamics of adaptation and diversification : A 10 ,000-generation experiment with bacterial populations, *91*(July), 6808–6814.

Lindström, E. S., & Langenheder, S. (2012). Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports*, *4*(1), 1–9. doi:10.1111/j.1758-2229.2011.00257.x

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., … Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, *23*(11), 1817–28. doi:10.1101/gr.159426.113

Mallet, J. *et al*. Space, sympatry and speciation. *Journal of Evolutionary Biology* **22**, 2332–2341 (2009)]

Mayden, R. L. 1997. A hierarchy of species concepts: the denouement in the saga of the species problem. Pp. 381–424 *in* M. F. Claridge, H. A. Dawah, and M. R. Wilson, eds. Species: the units of biodiversity. Chapman and Hall, London.

Mayr, E. 1942.Systematics and the origin of species. Columbia Univ. Press, New York.

Pol, D. Siddall, M.E., (2001). Biases in Maximum Likelihood and Parsimony: A Simulation Approach to a 10-Taxon Case. *Cladistics*, *17*(3), 266–281. doi:10.1006/clad.2001.0172

Queiroz, K. De, & Queiroz, K. De. (2005). A Unified Concept of Species and Its Consequences for the Future of Taxonomy A Unified Concept of Species and Its Consequences for the Future of Taxonomy, *56*(June).

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*(7263), 489–94. doi:10.1038/nature08365

Road, W. M., & Kingdom, U. (2005). SPECIATION THROUGH COMPETITION : A CRITICAL REVIEW, *59*(6), 1194–1210.

ROSENZWEIG, M. L. (2008), Competitive speciation. Biological Journal of the Linnean Society, 10: 275–289. doi: 10.1111/j.1095-8312.1978.tb00016.x

Rundle, H. D., Breden, F., Griswold, C., Mooers, A., Vos, R. A., & Whitton, J. (2001). Hybridization without guilt : gene flow and the biological species concept, *14*, 2000–2001.

Russo, C. A. M., & Takezaki, N. (1996) Efficiencies of Different Genes and Different Tree-building in Recovering a Known Vertebrate Phylogeny Methods. Mol. Biol. Evol. 13:525-536.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science (New York, N.Y.)*, *323*(5915), 737–41. doi:10.1126/science.1160006

Shah, A. M., Tamang, R., Moorjani, P., Rani, D. S., Govindaraj, P., Kulkarni, G., … Thangaraj, K. (2011). Indian Siddis: African descendants with Indian admixture. *American Journal of Human Genetics*, *89*(1), 154–61. doi:10.1016/j.ajhg.2011.05.030

Sharma, G., Tamang, R., Chaudhary, R., Singh, V. K., Shah, A. M., Anugula, S., … Thangaraj, K. (2012). Genetic affinities of the central Indian tribal populations. *PloS One*, *7*(2), e32546. doi:10.1371/journal.pone.0032546

Tamura K. and Nei M. (**1993**).Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.*Molecular Biology and Evolution***10**:512-526.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S. (**2011**). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution*

Wiley, E. (1981).*Phylogenetics: The Theory and Practice of Phylogenetic Systematics*.Wiley and Sons, New York.
