REPORT ON THE DISCUSSION MEETING

ON APPLICATIONS OF STATISTICS & PROBABILITY IN SCIENCE & ENGINEERING

June 12–13, 1997

Chemical Engineering Division
National Chemical Laboratory
Pune 411008

SPONSORED BY
Department of Science & Technology
Government of India

ORGANISED BY
Dr. Probal Chaudhuri
Indian Statistical Institute, Calcutta
and
Dr. Vivek V. Ranade
National Chemical Laboratory, Pune

REPORT ON THE DISCUSSION MEETING

ON APPLICATIONS OF STATISTICS & PROBABILITY IN SCIENCE & ENGINEERING

June 12–13, 1997

Chemical Engineering Division
National Chemical Laboratory
Pune 411008

SPONSORED BY

Department of Science & Technology
Government of India

ORGANISED BY

Dr. Probal Chaudhuri
Indian Statistical Institute, Calcutta
and
Dr. Vivek V. Ranade
National Chemical Laboratory, Pune

PREFACE

A group of approximately twenty people consisting of statisticians and probabilists on one hand and scientists and engineers from different fields on the other met at the National Chemical Laboratory at Puna during June 12–13, 1997. The objective of the meeting was to identify specific problems in science and engineering where statistical and probabilistic methodology can be potentially useful. A fair amount of 'homework' was done by the group before actually meeting at Puna, and problems to be discussed in the meeting were identified well in advance. A good deal of technical exchanges took place before and during the meeting among the scientists posing the problems highlighting statistical and probabilistic components there and the statisticians and the probabilists participating in the discussion giving their inputs. It is hoped that this discussion meeting will act as a seed for growing effective multidisciplinary research projects involving scientists and engineers collaborating with statisticians and probabilists.

The meeting was organised jointly by **Dr. Probal Chaudhuri** of *Indian Statistical Institute, Calcutta* and **Dr. Vivek V. Ranade** of *National Chemical Laboratory, Pune.* It was hosted by the **Chemical Engineering Division of National Chemical Laboratory**. Funding for the meeting was provided by the **Department of Science and Technology, Government of India**.

Principal participants of the meeting were **Dr. Tathagata Banerjee** (*a statistician from Calcutta University*), **Dr. D. P. Burma** (*a molecular biologist from Banaras Hindu University*), **Dr. Probal Chaudhuri** (*a statistician from Indian Statistical Institute, Calcutta*), **Dr. J. V. Deshpande** (*a statistician from University of Puna*), **Dr. Anup Dewanji** (*a statistician from Indian Statistical Institute, Calcutta*), **Dr. A. Dharmadhikari** (*a statistician from University of Puna*), **Dr. A. P. Gore** (*a statistician from University of Puna*), **Dr. R. Jaganathan** (*a chemical engineer from National Chemical Laboratory*), **Dr. N. V. Joshi** (*an ecologist from Indian Institute of Science, Bangalore*), **Dr. Sumitra Purokayastha** (*a statistician from Indian Institute of Technology, Bombay*), **Dr. M. B. Rajarshi** (*a statistician from University*

of Puna), **Dr. B. Rajeev** (*a probabilist from Indian Statistical Institute, Calcutta*), **Dr. S. Ramasubramanian** (*a probabilist from Indian Statistical Institute, Bangalore*), **Dr. Vivek V. Ranade** (a chemical engineer from National Chemical Laboratory), **Dr. Rahul Roy** (*a probabilist from Indian Statistical Institute, Delhi*), **Dr. Debashis Sengupta** (*an atmospheric scientist from Indian Institute of Science, Bangalore*), **Dr. V. Sitaraman** (*a biologist from University of Puna*). **Dr. Mohan Delampdy** (*a statistician from Indian Statistical Institute, Bangalore*) and **Dr. P. K. Saraswati** (a geologist from Indian Institute of Technology, Bombay) contributed substantially towards discussion on two main problems before the meeting. They were unable to attend the meeting for unforeseen personal reasons. Besides the principal participants, some research scholars and scientists from University of Puna and National Chemical Laboratory attended the meeting.

The organisers of the meeting are thankful to the **Project Advisory Committee, Mathematical Sciences Section, Department of Science and Technology, Government of India** for their generous financial and moral support.

FIRST SESSION
(Morning of June 12, 1997)

Statistical Problems in the Analysis of Chaotic Dynamics of Multiphase Reactors

ANALYSIS OF CHAOTIC DYNAMICS OF MULTIPHASE REACTORS

Vivek V. Ranade
Chemical Engineering Division
National Chemical Laboratory
Pune 411 008, INDIA
Tel/Fax: + 91 212 333941
Email: ihcu@ncl.ernet.in

Prepared for the discussion meeting between Statisticians and Scientists & Engineers

Background:

Multiphase reactors are widely used in chemical and bio-chemical industries. The knowledge and adequate modelling of macro-scale fluid dynamics is essential for a proper design and scale-up of these reactors. Importance of simulation of these flow patterns has been recognized since long. However, almost all of the available information about the hydrodynamics of these reactors is from the crude, empirical models and from the experiments. Several excellent reviews are available in the literature, all of which emphasise a need for generating additional information about the details of hydrodynamics of these reactors. Only recently, there have been some attempts to develop models to simulate detailed fluid dynamics (for example Ranade, 1995) which show promising results. Unfortunately, these attempts of simulating detailed fluid dynamics use various averaging procedures. These averaging procedures blurr existing flow strucutres. While these models can predict mean velocities and turbulent kinetic energy distribution the details of inherently unsteady flow are lost. When gas is sparged at the bottom of a liquid pool in a sparged reactor, the resulting flow is characterised by many distinct flow structures of various length scales. There exists a radially non-uniform distribution of gas phase. This generates overall circulatory flow within the column. This flow interacts with several smaller vortices generated near the wall. Besides these, each bubble will shed vortices behind it with its characteristic frequency and scale. In general the flow will be turbulent and will be dominated by the interactions between flow structures of various scales. None of these complexities can be captured by the existing CFD tools. It is indeed necessary to develop appropriate methods to characterise these details of flow in multiphase reactors.

Only recently has it become clear that sparged multiphase reactors can be considered as a deterministic chaotic system. Tremendous advances have been made in the analysis of deterministic choas in dissipitive systems. These advances in chaos analysis can be fruitfully utilised to characterise dynamics of flow in bubble columns. Though chaos in dissipative systems is a much younger field, it provides a theory and tools to discover the inner workings of systems that on first viewing appear inaccessible. Perhaps nonlinear dynamical properties of whole macroscopic systems could be used to obtain information on parameters of the small and even microscopic components of the system. These methods can also help to identify and characterise coherent structures in these flow systems.

The analysis of pressure fluctuations can be fruitfully employed to understand the chaotic dynamics of these multiphase reactors. Rugged pressure sensors which can withstand severe environment occurring in these multiphase reactors are available and can be used to analyse dynamics of industrial multiphase reactors. Our idea is to develop a diagnostic tool kit based on analysis techniques applied to pressure fluctuations data from multiphase reactors. Such a tool kit along with the time averaged CFD models will be very useful for design and scale-up of multiphase reactors. There have been attempts to collect and analyse fluctuating pressure data from mulitphase reactors. However, in almost all the cases the data is corrupted by noise. The methods for estimating and characterising chaotic dynamics of multiphase reactors with the noisy data is a complex problem. The proposed discussion meeting with statisticians will be useful to clarify and improving the diagnostic tool kit to understand the dynamics of multiphase reactors.

Analysis of Noisy Signals:

The collected digitised time series of pressure fluctuations is normally low pass filtered and normalised with average absolute deviation. The filtered and normalised time series is then processed to reveal various characteristics. Following indicators are computed and used to get preliminary analysis:

+ average cycle frequency
+ average number of data points per cycle
+ cycle frequency distribution
+ probability density distribution
+ power spectra

In addition to these several indicators are computed to understand chaotic properties of the pressure fluctuations. Some of these are as follows:

+ rescaled range (R/S) analysis: it gives us a possibility to estimate Hurst exponent (H) for a given time series. The value of H obtained from R/S analysis is related to local fractal dimension, dF as dF=2-H for H in the range of zero to one.
+ correlation dimension: it measures spatial structure of the attractor in the reconstructed state space. This gives an useful indication of the number of degreed of freedom and minimum number of independent state variables required to describe the chaotic system.
+ Kolmogorov entropy: it measures degree of 'chaotic-ness' of dynamic system and represents rate of information loss or average degree of predictability of a dynamic state.
+ Principal component analysis and construction of phase space trajectories

One has to make appropriate selections of the required parameters like number of data points, bin sizes, cut-off distances while computating these indicators. The information about the influence of these selections on the quality of estimated indicators is scanty. No general guidelines are available to make judicious selections. Presence of noise furhter complicates these estimation procedures. It will be very helpful to get a feedback from statisticians about the following:

+ guidlines for objective selection of number of data points, bins etc.
+ error analysis of the estimated indicators
+ handling of noisy data
+ efficient algorithms

SUMMARY OF COMMENTS FROM STATISTICIANS AND PROBABILISTS

The topic contains a mixture of dynamical systems and stochastic process analysis. Although the gas sparging and the related "engineering" aspects of the problem suggest a dynamical system model in operation, the analysis done with the pressure data obtained seem to rely on techniques of stochastic processes.

Dynamical Systems : The introduction of the gas at the spargers starts the system, and it is assumed that "soon" the system arrives at an "equilibrium" state. Mathematically then we can model it by an appropriate measure preserving transformation T under a suitable measure μ. Then observing the pressure is equivalent to observing the state at a particular point in space (viz. the point where the pressure is being measured). However, this mathematical model seems to be inconsistent with the data because it would predict an invariant distribution of the observations, which does not seem to be true. This appears to be a major difficulty with the dynamical systems approach of modeling of the observed phenomenon here.

Stochastic Modeling : To do a fractal analysis of the data as has already been done, one can make use of the extensive literature on the theory of stochastic processes initiated by Pierre Kahane in the 60's and presented excellently in Robert Adler's book "The Geometry of Random Fields". Suppose that $\{X_t\}$ is a stationary Gaussian process and assume that $Cov(X_t, X_{t+h}) = O(h^\alpha)$ as $h \to 0$ for some $\alpha > 0$. This α is called the **fractal index** of the process. It is known that if D_X is the (almost sure) Hausdorff dimension of the sample paths of X_t then $D_X = 2 - (1/2)\alpha$. This suggests a connection between the fractal index and the Hurst exponent that has already been used in the analysis. This relation holds for a wider class of precesses which may be obtained as functions of Gaussian processes. Moreover there are nice statistical methods based on counting upcrossings of a certain level by the processes. So if one could model the chemical phenomenon "stochastically" there could be some conclusions drawn using standard stochastic techniques. The definition of correlation dimension suggests that it is related to crossings. As such it has possible connections with the fractal dimension mentioned earlier. Thus mathemat-

ically it is worth investigating whether there is any relation between the fractal index, Hurst exponent and correlation dimension. Finally about Kolmogorv entropy mentioned here it would be useful to investigate how the three definitions suggested relate to the standard Kolmogorov-Shanon entropy studied in probability literature.

Statistical Estimation : The statistical issues of estimating the right parameters can be studied once various notions given in the paper are given firm probabilistic and mathematical footing. For various equations used to estimate correlation dimension and Hurst exponent from the data, one can use "change point" estimation techniques (which is now well developed in statistical literature) for identifying "different regimes" of the equations. Just as linear equations are fitted by the method of least squares applied to observed data points, one can also estimate the change points, where the line changes its slope using least squares techniques. Appropriate selection of "bin sizes", "cut-off distances", etc. will involve careful exploratory analysis of the observed data and proper utilization of any prior knowledge that might be available.

SECOND SESSION
(Afternoon of June 12, 1997)

Statistics in Palaentological Problems

Statistics in Palaeontological Problems

P.K. Saraswati

Department of Earth Sciences
Indian Institute of Technology, Bombay
Powai, Mumbai 400076
INDIA

Note prepared in connection with a discussion meeting between statisticians and scientists of other disciplines, to be held at NCL, Pune, in June, 1997.

1. Introduction

Fossil record is the direct evidence of the evolution of life. To read the evolutionary history from the fossil record one needs to recognize the basic biological units from which the evolutionary patterns are deduced. The biological units are recognized at various levels of hierarchy. In this system of hierarchic classification the natural groups of same status are designated by rank-names. 'Species' is the lowest taxonomic rank in this hierarchy. There are two kinds of species concepts, one is process related and the other is pattern based. The former emphasizes the biological processes that give rise to the species and the latter stresses the operational means of recognizing them. Due to various limitations of process based definitions, especially for the fossil forms, in practice the species are mostly identified on diagnosable morphological characters. The process of recognition begins with the identification of a morphological group that is consistently diagnosable on the basis of certain meaningful characters. It is a two-step process. In the first step the morphological variations within the populations of single locality or time plane are documented. The next step involves comparison of populations from different localities and different time planes to find out the existence of significant morphological difference. The closely similar populations are aggregated and those with significant differences or with unique character states are separated.

2. Nature of Morphological Characters

The morphological features may be qualitative, quantitative or dichotomous. The qualitative characters may have many states, for example, ornamentation on a shell may be radial, spiral or reticulate. The dichotomous characters are those which are either present or absent, for example, presence or absence of alar prolongation. Some of the characters can be quantified and they can be both continuous as well as discrete variables. The length and width of chambers in *Nummulites* are continuous variables while number of operculine chambers in *Heterostegina* is a discrete variable.

Traditionally, taxonomists have relied on qualitative and dichotomous characters. Very often quantifiable characters are also transformed to qualitative descriptions. The size of the shell or its component parts are mostly described as 'large' or 'small'. It is, therefore, not uncommon in palaeontologic literature that species assignment of one worker is questioned or rejected by the other.

3. Heterogeneity of Morphological Data

Morphological data seldom shows an ideal bell-shaped, normal distribution. The departure from this type of distribution in fossil population is likely to arise due to one or more of the following reasons: a. Individuals in a population may represent different growth stages. On account of this the data may be highly skewed, or it is possible that some size class may be missing. b. Different modes of reproduction may produce bimodal distribution. For example, a foraminiferal population may show two modal classes corresponding to A- and B- generations. c. Individuals from different habitats may be transported by geological agencies to produce heterogeneity in the sample.

4. Numerical Taxonomy

As stated elsewhere, it is not uncommon to have differences of opinion over the identification of taxa. In most cases it is due to recognition of taxa on qualitative criteria. To bring objectivity, 'phenetic species' concept was developed by numerical taxonomists in 1960s and 1970s. This is excellently reviewed by Sneath and Sokal (1973). It suggests use of all kinds of characters including morphological, physiological, behavioral, chemical and ecological. The process involves measurements of various characters for a set of taxa, computation of similarity coefficients and cluster analysis of the similarity matrix. For estimating taxonomic resemblance either distance coefficients or the general similarity coefficient of Gower (1971) are used. The distance coefficient (d_{jk}) is calculated as follows:

$$d_{jk} = \sqrt{\frac{\sum_{i=1}^{m}(X_{ij} - X_{ik})^2}{m}},$$

where, X_{ij} and X_{ik} denote the value of i-th variable of j-th and k-th object, respectively and m is the total number of variables. The advantage of Gower's similarity coefficient is that it accepts both quantitative and qualitative data. It is expressed as follows:

$$S_{jk} = \frac{\sum_{i=1}^{m} W_{ijk} S_{ijk}}{\sum_{i=1}^{n} W_{ijk}},$$

when comparison between two individuals is valid, the weight W_{ijk} is set to 1 and when the value of character i is unknown for one or both the individuals it is set to 0. S_{ijk} takes different values for different types of characters. For two state and

multistate characters $S_{ijk} = 1$ for matches and $= 0$ for mismatches. For continuous variables it is computed as follows:

$$S_{ijk} = 1 - \frac{|X_{ij} - X_{ik}|}{R_i},$$

where X_{ij} and X_{ik} are the values of character i for the j-th and k-th individuals (to be read from the data matrix) and R_i is the range of variable in the sample.

After obtaining a similarity matrix, mutually similar individuals are grouped together by cluster analysis. In taxonomic studies Q-mode cluster analysis is used which shows interrelationships between individuals. There are large number of possible methods of cluster analysis which are described by Everitt (1980) and Sneath and Sokal (1973). The results of clustering is displayed by constructing a dendrogram, customarily known as phenogram in numerical taxonomic work.

As stated above, numerical taxonomists advise a large number of characters for natural groupings. In fossil materials it is not possible to obtain physiological, behavioral or ecologic data. Moreover, the morphological characters are also limited compared with the living forms. Due to these limitations, conventional taxonomy continues to be practiced routinely. Statistical methods are, however, useful to differentiate closely similar taxa and to ascertain the taxonomic status of those large number of forms which are created on the evidence of minor variations. To illustrate the application in this area, statistical analysis of two species of a foraminiferal genus is discussed below.

Nummulites is a protistan microfossil, characterised by a calcareous test which is planispirally coiled and is internally divided into chambers. Two of its species, *Nummulites stamineus* and *Nummulites neglectus,* recorded from Middle Eocene successions, are morphologically very similar. Only an expert's eyes, which are accustomed to the amount of variation typical within each group, would be able to differentiate the two species. To examine the validity of the two closely similar species and to make the identification easy, statistical approach is followed here. Seven morphological features are measured in the equatorial sections of 35 individuals of the B-forms of the two species. The Euclidean distance coefficients are calculated and the data matrix is subject to Q-mode cluster analysis. The sequential agglomerative hierarchial non-overlapping (acronymed as SAHN) method of cluster analysis is followed. The groupings (fig. 1) in general correspond to the conventional classification.

To allocate an unknown or disputed specimen to one of the two species, linear discriminant analysis is carried out. Discriminant function based on two samples is a linear function that has greatest variance between samples relative to the variance within samples. For discriminant analysis, the data should have multivariate normal distribution. In the present example only marginals are checked. From the histogram (fig. 2) some of the data do not seem to be normal. The variables are therefore transformed to their log values, and the skewness and kurtosis of the (log-) transformed data are found. It is observed that for certain variables skewness and kurtosis values are improved (approaching zero) after log transformation while for the others it deteriorates. Because some of the highly skewed distributions are improved by transformation, the log values are used for discriminant analysis. The following discriminant function is found to seperate the two species:

$$R = -195.9D + 335.5T - 22.9LL - 406.7HL + 17.0M.$$

The discriminant score R_0, which sets the dividing line between the two species is 174.6. Specimens having $R > R_0$, will be classified as *Nummulites stamineus* and those $< R_0$ as *Nummulites neglectus*. The discriminant score plot (fig. 3) of the two species shows that the function clearly discriminates the two species.

5. The Problems

The statistical techniques are often based on the theoretical assumption of a normal distribution. In view of this the morphological data needs to be normalised prior to statistical analysis. Taxonomists' interest lies in knowing the following:

1. How to find whether data is normally distributed? Is it sufficient to check the skewness and kurtosis values or is it necessary to derive mathematically the nature of distribution ?

2. How much deviation from normal distribution is admissible ?

3. How to check multivariate normality ?

4. Can a "user friendly" approach be devised to detect both marginal and multivariate normality of the data and to normalize a non-normal distribution ?

Most of the palaeontological problems involve classification and discrimination. Cluster analysis, discriminant analysis and principal component analysis are the commonly used methods in taxonomy. Because the palaeontological data often deviate from normal distribution it is required to explore the methods which are least likely to be affected by limited violation of the theoretical assumption of the multivariate methods.

6. Conclusion

The palaeontological problems amenable to statistical analysis are related to classification and discrimination of ancient biological forms. The input data may be combination of quantitative, dichotomous and multistate characters. The quantitative data is likely to be non-normal due to various biological and abiological factors. It is required to detect the nature of distribution and transform the data to multivariate normal distribution. The robust methods of statistical analysis should be explored which would be least affected by limited violation of the normal distribution.

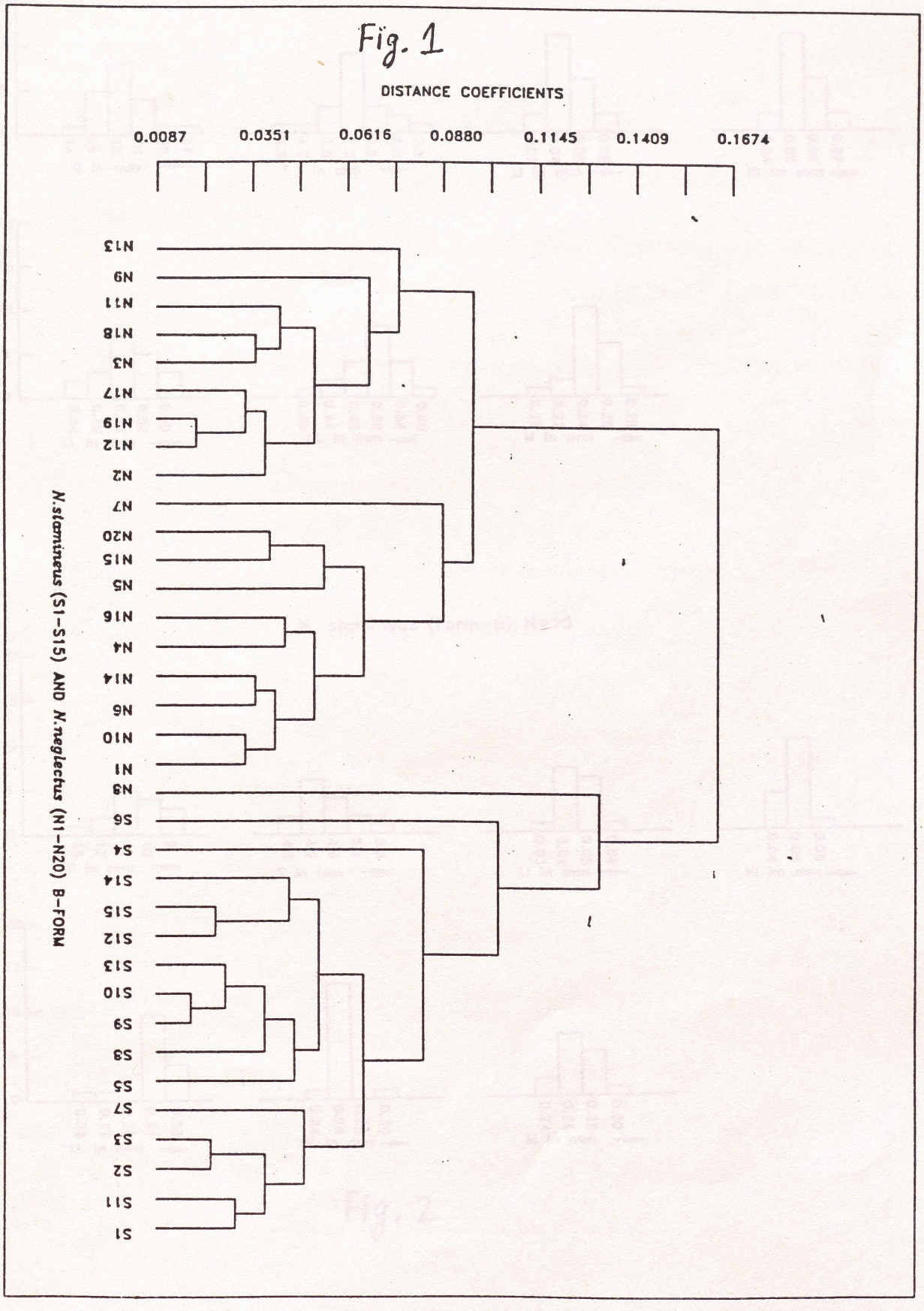
References

Everitt, B. (1980). *Cluster Analysis.* 2nd Ed., Halsted Press, New York, 135 pp.

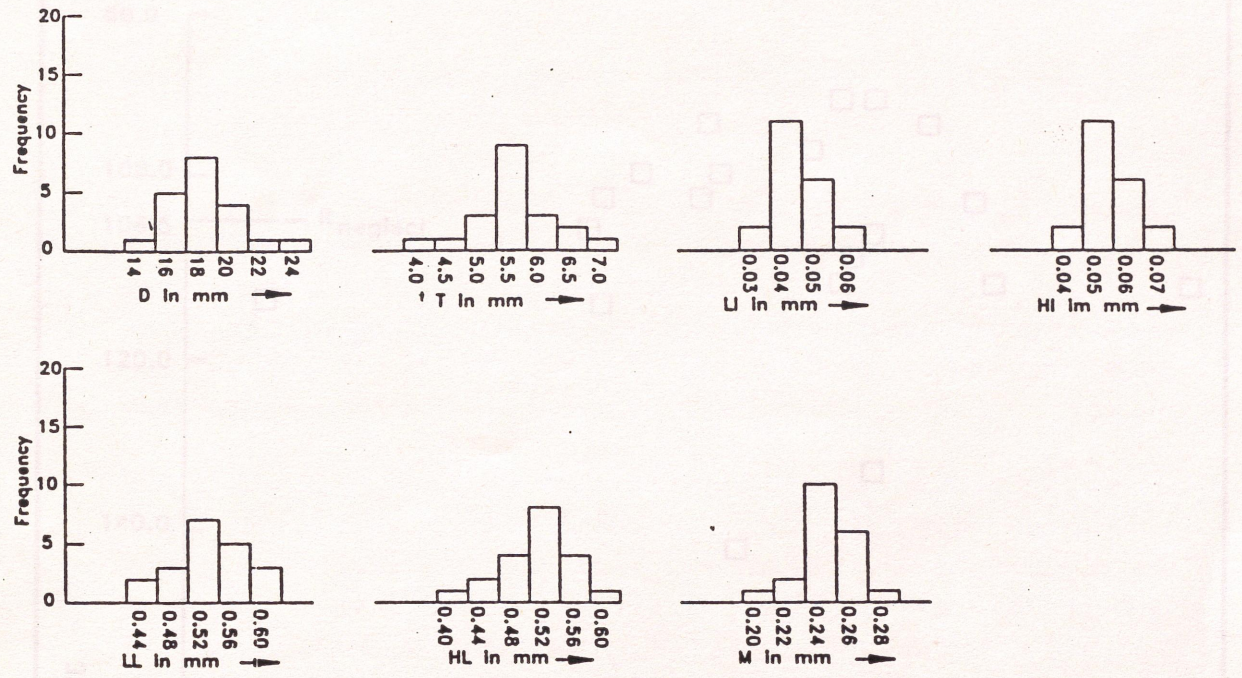
Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857-872.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy.* W.H. Freeman and Co., San Francisco, 573 pp.

Fig. 1


DISTANCE COEFFICIENTS
0.0087
0.0351
0.0616
0.0880
0.1145
0.1409
0.1674
N13
N9
N11
N18
N3
N17
N19
N12
N2
N7
N20
N15
N5
N16
N4
N14
N6
N10
N1
N8
S6
S4
S14
S15
S12
S13
S10
S9
S8
S5
S7
S3
S2
S11
S1
N. stamineus (S1-S15) AND N. neglectus (N1-N20) B-FORM

N. neglectus (From-B) N=20



N. stcmineus (Form-B) N=15

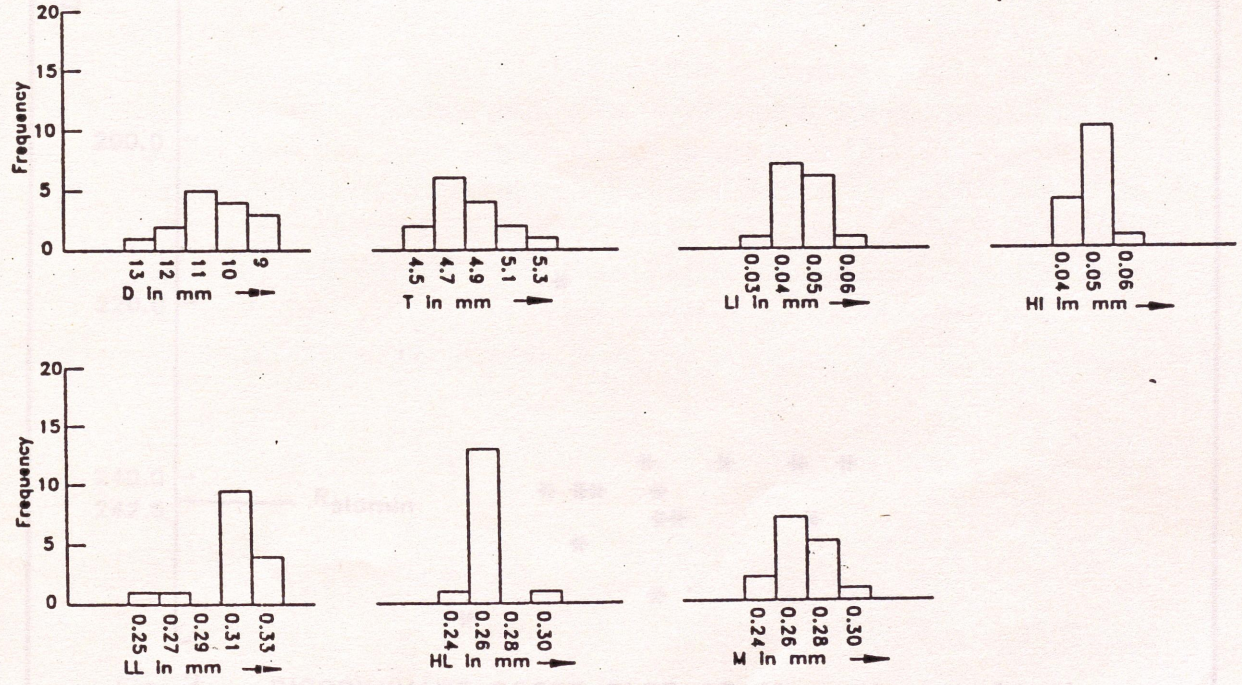


Fig. 2

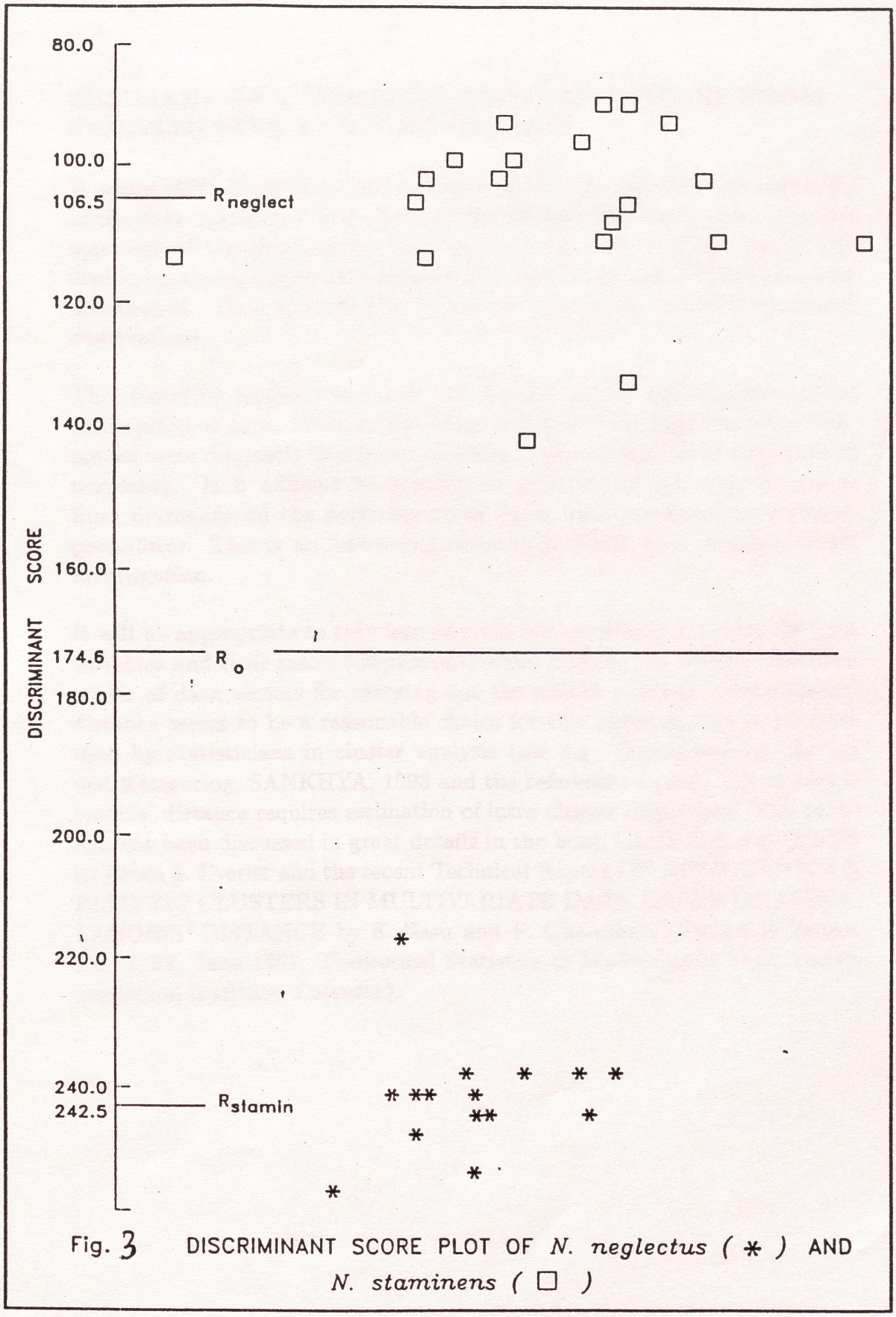


Fig. 3 DISCRIMINANT SCORE PLOT OF *N. neglectus* (*) AND *N. staminens* (□)

SUMMARY OF COMMENTS AND SUGGESTIONS FROM STATISTICIANS AND PROBABILISTS

It seems difficult to carry out a statistical test for multivariate normality of the data based on so few observations. Instead one can try the Box-Cox approach of transforming the multivariate data with the hope that it will lead to an appropriate transformation making the data reasonably normally distributed. Then analysis can be carried out based on the transformed observations.

The clustering methodology does not depend on the multivariate normal assumption as such. However the linear and quadratic discrimination techniques were originally developed based on the assumption of multivariate normality. It is difficult to quantify in general the effect of deviations from normality on the performance of linear and quadratic discriminant procedures. This is an interesting research problem that requires careful investigation.

It will be appropriate to take into account the correlations among different variables and their scales (dispersions) while forming the distance between a pair of data vectors for carrying out the cluster analysis. Mahalanobis' distance seems to be a reasonable choice for this purpose, and it has been used by statisticians in cluster analysis (see e.g. Gnanadesikan, Harvey and Kettenring, SANKHYA, 1993 and the references there). Use of Mahalanobis' distance requires estimation of intra-cluster dispersion. This problem has been discussed in great details in the book CLUSTER ANALYSIS by Brian S. Everitt and the recent Technical Report ON HOMOGENEOUS ELLIPTIC CLUSTERS IN MULTIVARIATE DATA BASED ON MAHALANOBIS' DISTANCE by S. Basu and P. Chaudhuri (Technical Report No. 6/97, June 1997, Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, Calcutta).

THIRD SESSION
(Afternoon of June 12, 1997)

Statistical Problems in Evolutionary Analysis of DNA sequence Data

Evolution of DNA

D.P. Burma

Molecular Biology Unit,
IMS, BHU,
Varanasi

As soon as DNA was established as the genetic material and there were glimpses about its structure it started becoming clear that the 'Evolution of species' was dependent on the 'Evolution of DNA'. That one species evolved from another was proposed by Darwin on the basis of geographic separation, habitats, phenotypic characters etc. of the various species. When the method of protein sequencing came into existence proteins became the molecular markers of evolution. Extensive work was done by various workers to build the phylogenetic trees with the help of proteins like globin, cytochrome C, fibrinopeptide, insulin and various enzymes like carbonic anhydrase. When the DNA sequencing came into vogue due to the pioneering work of Sanger (dideoxy technique) and Maxam and Gilbert (chemical method) phylogenetic approaches switched over to DNA as the molecular marker and there was no doubt left that evolution of species can be equated to the evolution of DNA.

Before the technique of sequencing of DNA was introduced attempt was made to characterise DNAs by various methods, one of the earliest being that of Chargaff who first observed from his meticulous work on base analysis of DNA that A = T and G = C. The significance of this important observation was realised by Watson and Crick and on that basis they built the double-stranded structure of DNA. The melting temperature of DNA was one of the criteria established by Doty and Marmur and used extensively for characterisation of DNA. Simultaneously Spiegelman introduced the DNA-

DNA hybridisation technique which became a very important tool not only for characterisation of DNA but also for comparison of different DNA sequences. Another important work was carried out in 1961 by Kornberg, from the dinucleotide frequencies of the newly synthesised radioactive DNA strand on a single-stranded DNA template. Dinucleotide frequencies are being used in some laboratories including ours for characterisation of genes and genomes. Recombinant DNA techniques ushered in following the discovery of restriction-modification phenomenon by Arber and extensive characterisation of restriction enzymes by Smith and Nathans. But the real breakthrough came with the introduction of DNA sequencing techniques, the chemical method as well as the enzymatic method. At present Sanger's method is the method of choice. As the DNA sequences of genes and genomes of various organisms piled up and became available through the data bank, the molecular approaches to the studies on 'evolution of DNA' became a wide-spread practice. Extensive work started and several computer programmes were developed in search of sequence homologies for various types of DNA. It also become rather easy to establish phylogenetic trees from the sequence homology. Thus the study of evolution started truly on the molecular basis.

Originally the living world was partitioned into two broad classifications, prokaryotes (eubacteria) and eukaryotes. Using ribosomal DNA as marker it was deduced that there are actually three different kingdoms, archaebacteria, being the third one. The latter being intermediate between pro and eukaryotes have characters common to both of them.

Various processes have played their individual as well collaborative roles in the different steps of evolution of DNA. Mutation is naturally one of the key steps. Several natural phenomena exerted selection pressures which have been the primary causes behind the changes of bases. Both chance and

necessity played their role. It is obvious from the fact that mutation is comparatively less in the functional areas which would have been detrimental to the survival of the species. Alternatively, some mutations will have very little effect on the functional abilities. That led Kimura to suggest the 'neutral theory of evolution' which has great significance in evolution. Recombination is another process which played important and extensive role in the evolution of DNA. Third phenomenon (transposition as well as retrotransposition) was overlooked by the geneticists for quite sometime. We are thankful to Barbara McClintock for challenging the classical concept. It should be pointed out at this stage that the maximum number of species evolved during the Cambrian period (about 500-550 million years ago). Somewhat convincing extrapolations show that the transposition of genes might have been initiated during this period.

Origin of life is still a mystery although Oparin's concept, Miller's prebiotic simulation, Orgel's hypothesis etc. contributed significantly to understand some aspects of the origin of life. It can however, be visualised that encasement of some very very primitive molecules of life within a membrane type of material gave birth to a cell (the basic unit of life) which learnt in due course of time how to divide and perpetuate itself. The next crucial event is the generation of nuclear membrane encasing the genetic material. The formation of nucleus is actually the start of eukaryotes (specially the metazoans) and its subsequent duplication along with the division of the cell was the new way devised by nature for the perpetuation of life. It was established earlier that some of the organelles like mitochondria, chloroplasts etc. are derived from prokaryotic organisms invading the eukaryotes. The DNAs that started to reside in might have been fragments of the prokaryotic DNAs. There are plenty of evidences in favour of that. Ribosmomal RNAs of the organelles are rather close in sequence homology to

prokaryotes than eukaryotes. Finally the fragments of DNA started to live in symbiosis with the host. Even some of the very recent evidences are in favour of the concept that the nucleus is also derived from invading prokaryotes. Although there is lot of controversy the theory is slowly gaining ground. It can, however, be surmised that symbiosis played a very significant role in evolution.

The epigenesis which is responsible for differentiation in the multicellular organisms has been another key step in the development of various organisms. Even in the most primitive organisms the device of expression of some specific genes and silencing of others became a rampant practice. As the term signifies epigenesis is the development involving gradual diversification and differentiation of an intially undifferentiated entity. This is one of the important mechanisms for biodiversity as well. Parental genomic imprinting mechanism may provide some clue in the area.

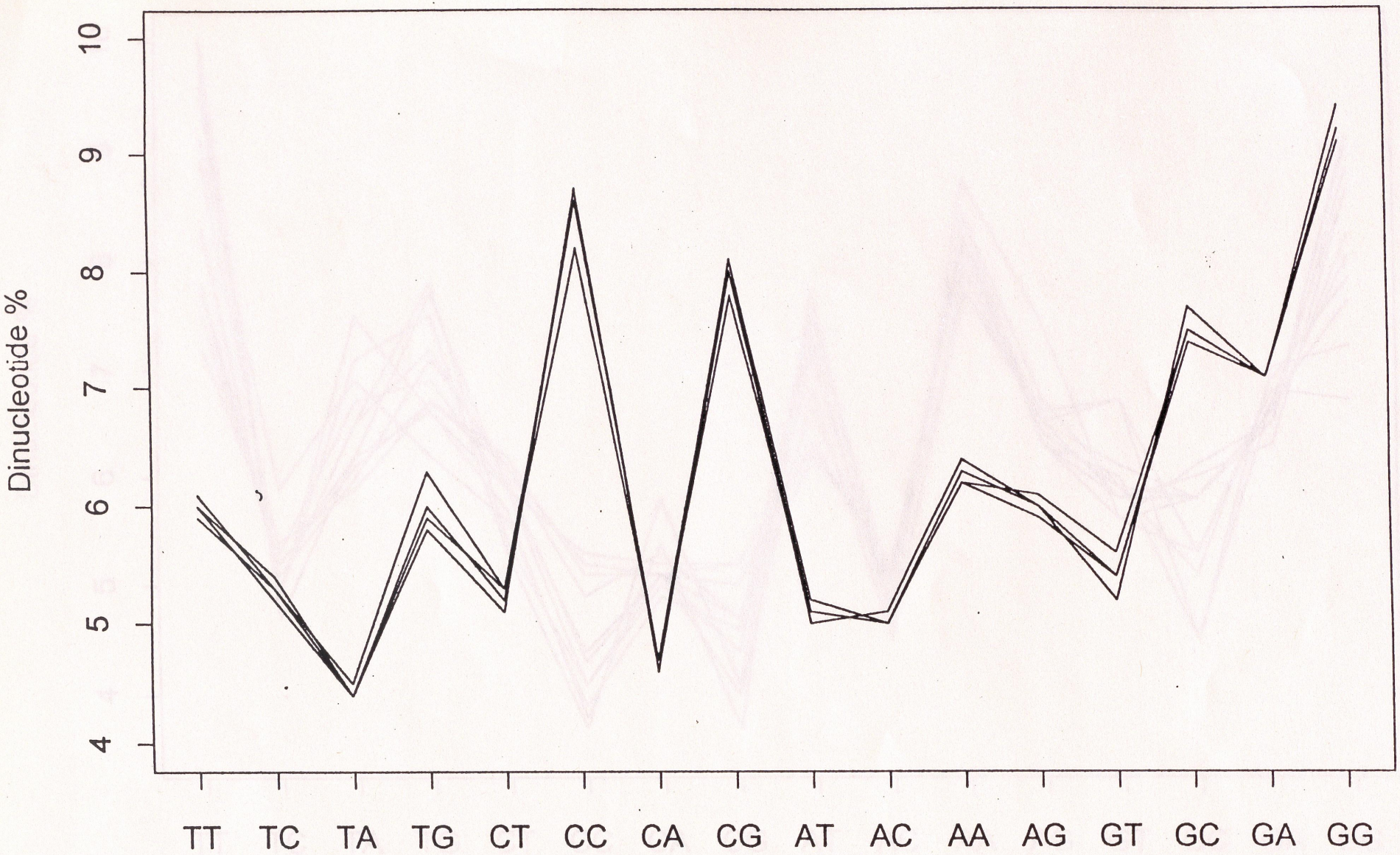
To come to our own work we are playing with the dinucleotide frequencies of several genes and genomes in search of simpler methods for studying the phylogenetic relationship. This is also being done by two other laboratories, one Russian (Nussinov) and another American (Karlin). This approach will be elaborated. The methods of building dendograms and phylogenetic trees based on changes in base composition coupled with parsimony and statistical methods (like bootstrapping) will also be discussed.

SUMMARY OF COMMENTS AND SUGGESTIONS FROM STATISTICIANS AND PROBABILISTS

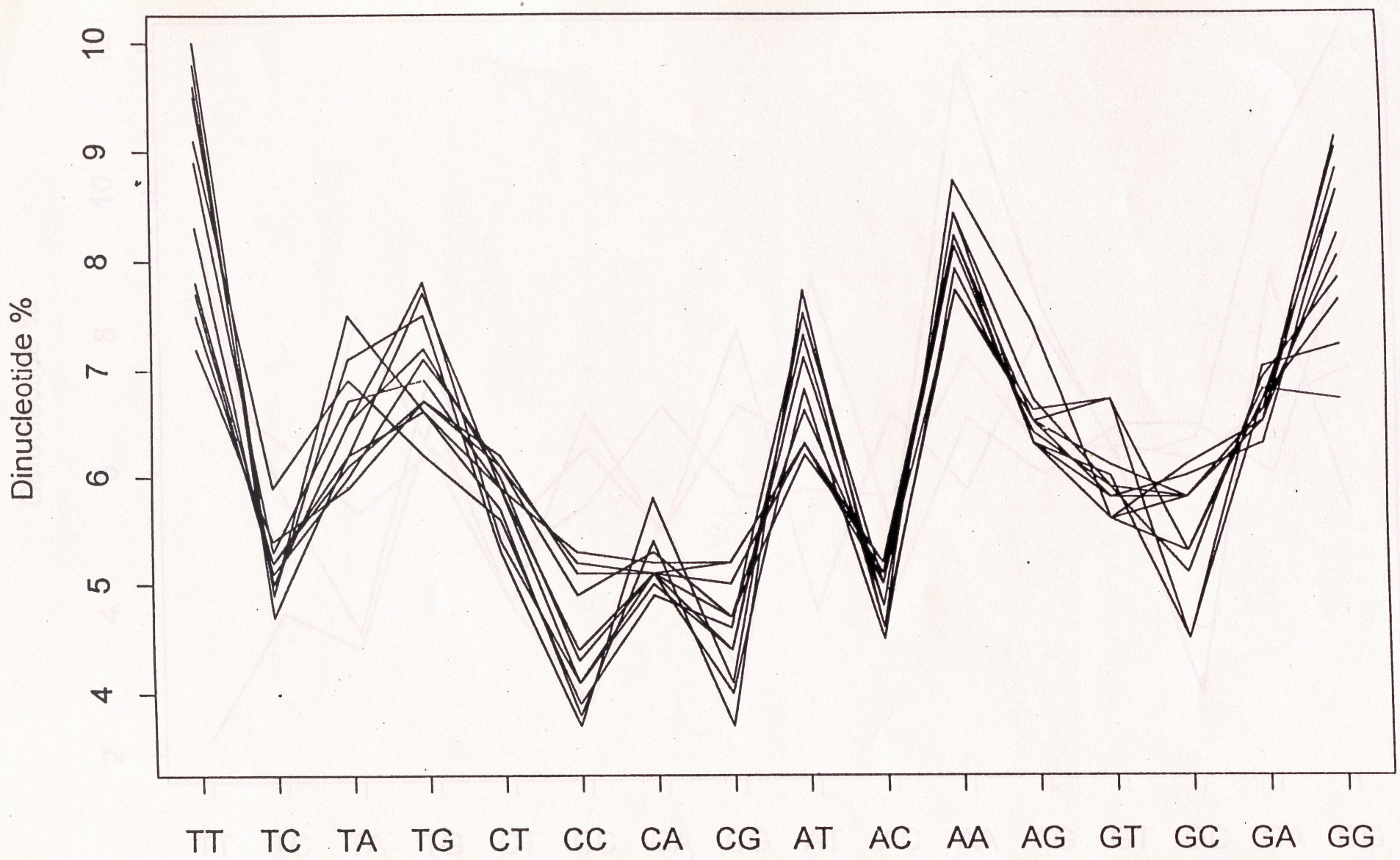
Dr. Burma is at present involved in collaborative research with statisticians and probabilists in Indian Statistical Institute, Calcutta and Calcutta University. Very recently they have jointly started looking at various statistical analysis of DNA data summarized through di- and tri-nucleotide frequencies. At this stage data on ribosomal DNA sequences and heat shock protein DNA sequences for a large number of organisms and data on DNA sequences for bacteriophages have been considered. The prime objective at this moment is to develop an understanding of the information content of di- and tri-nucleotide frequencies as far as phylogenetic relationships that exist among different organisms are concerned.

The preliminary analysis carried out so far are based on (i) linear plots of di-nucleotide frequencies, (ii) different types of cluster analysis (e.g. single and average linkage hierarchical clusters) and construction of dendrograms, and (iii) construction of minimal spanning trees. Some of the results are enclosed herewith. It seems that the results convey mixed messages. In some cases there are strong indications of possible phylogenetic relationships in the frequency data that are in conformity with known biological facts. Biologically close organisms are observed to form tight clusters when their di-nucleotide frequencies are compared. On the other hand, there are some abnormal (biologically) links noticed in the constructed trees, and sometimes frequency data tends to exhibit very wide variation as reflected in some of the linear plots.

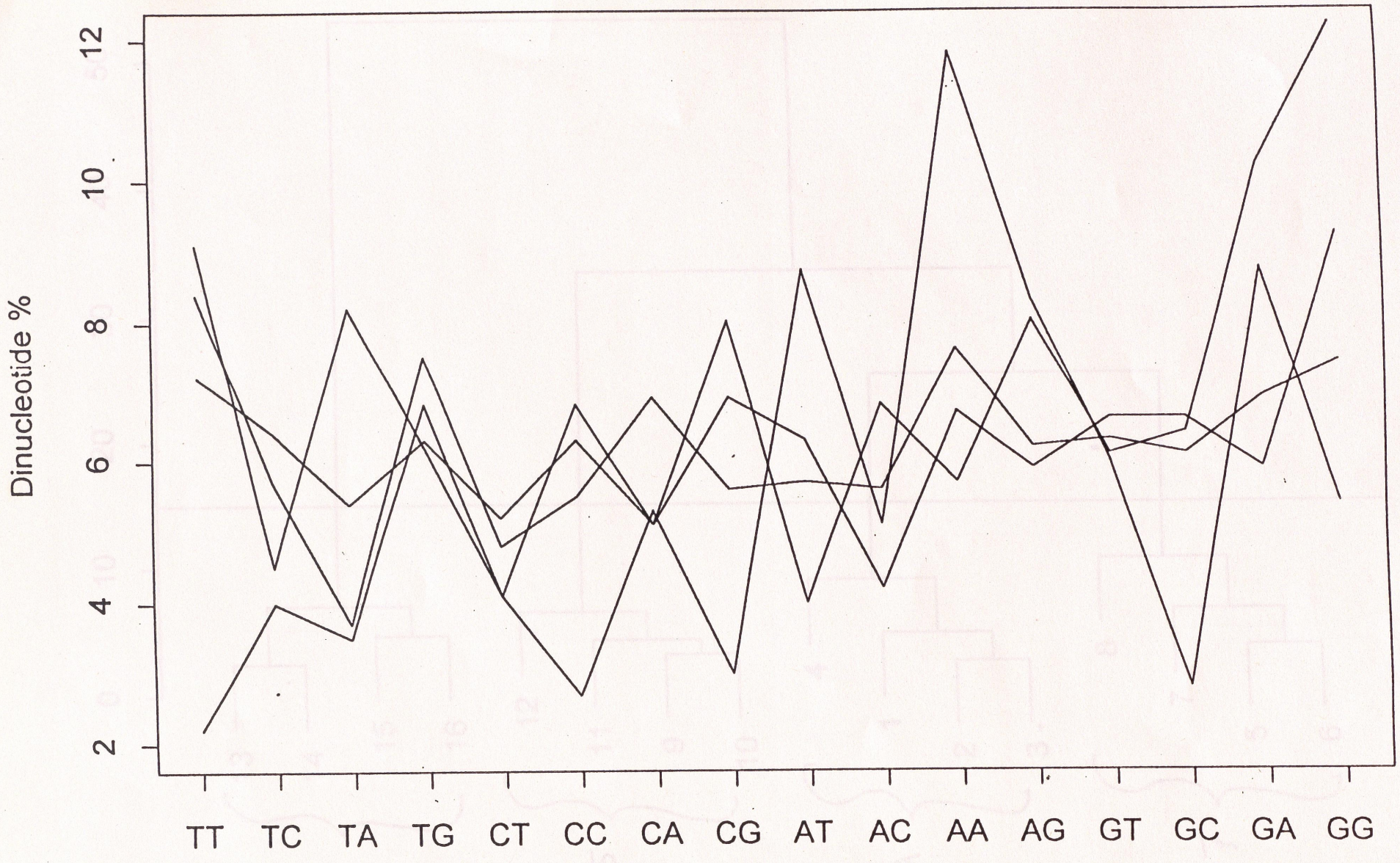
Cluster analysis as well as the construction of dendrograms and minimal spanning trees heavily depend on choice of the distance measure that quantifies the dissimilarity between the frequency vectors associated with a pair of DNA sequences. At this moment it is unclear what is the best possible method for weighing different frequencies for different nucleotide pairs and triples while forming the distance measure so that the results conform with well known biological facts. This is an important research problem which requires careful consideration. It seems appropriate to rank the frequency data so that relative abundance or scarcity of different nucleotide pairs or triples can be incorporated in the analysis. In order to derive biologically meaningful results, one should use some appropriate weights based on biochemical implications of changes in nucleic acid bases when a pair of DNA sequences are compared through frequency data.



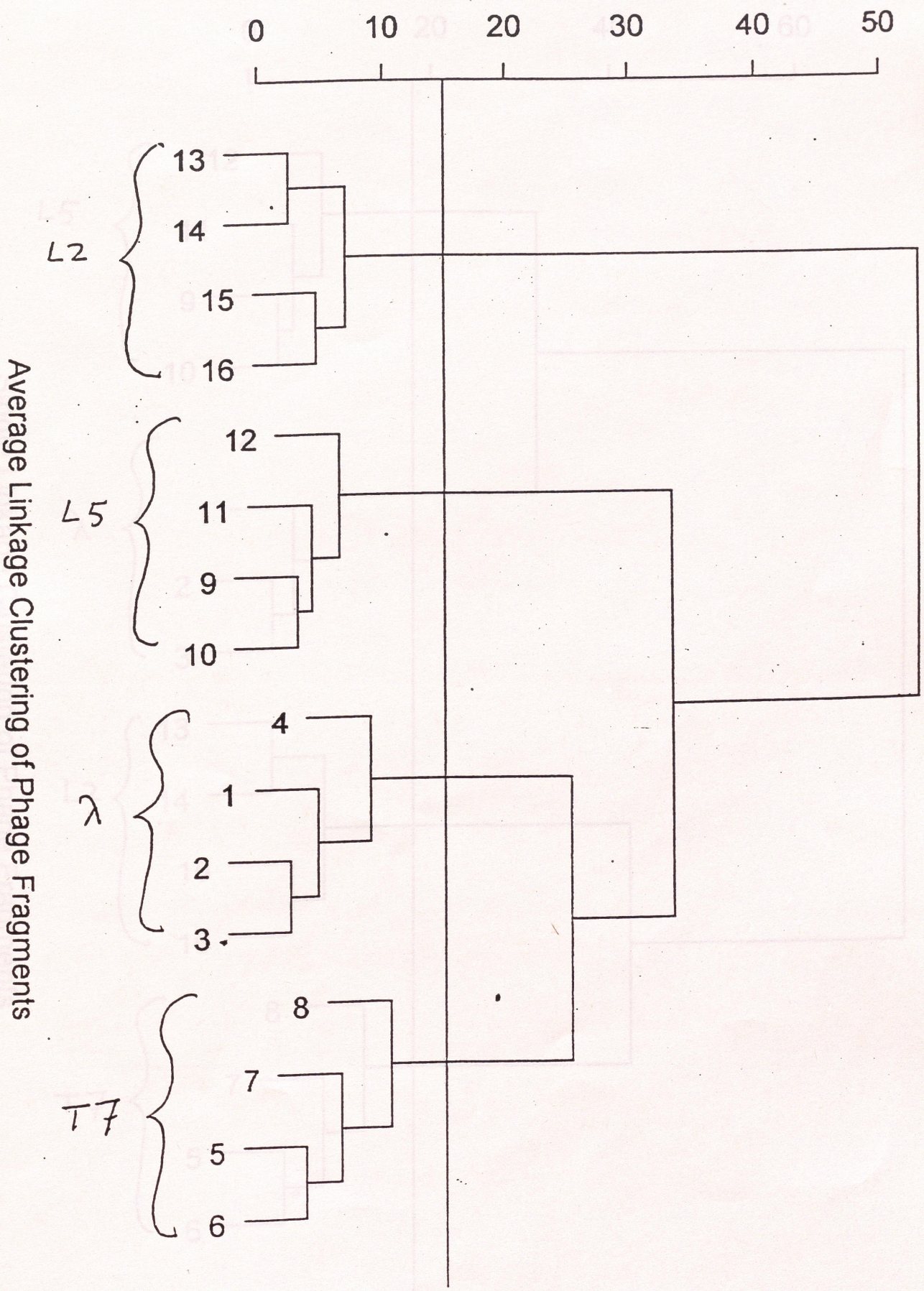
Dinucleotide Percentage Graph of Mammal rDna

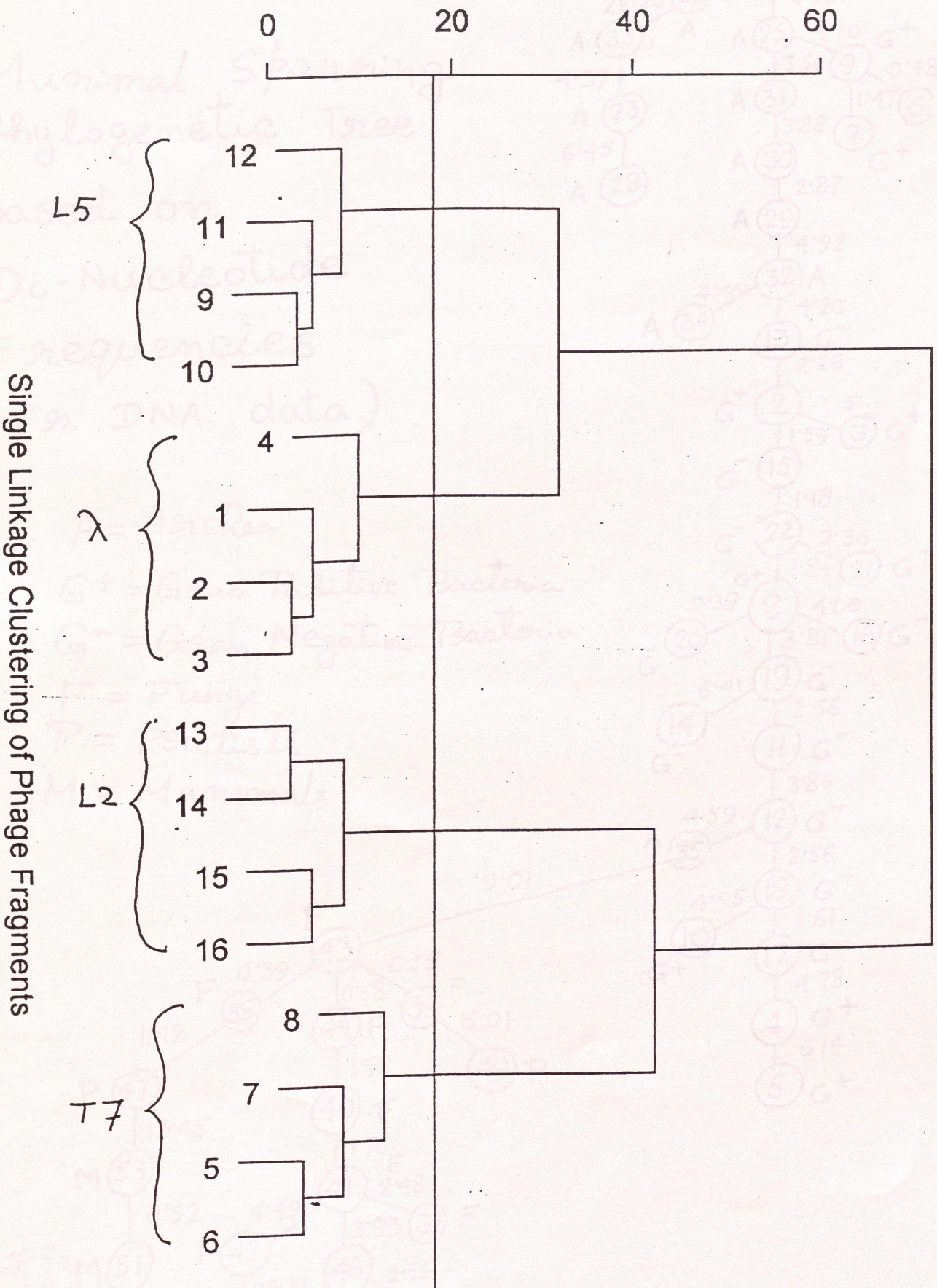


Dinucleotide Percentage Graph of Fungus rDna



Dinucleotide Percentage Graph of Protist rDna

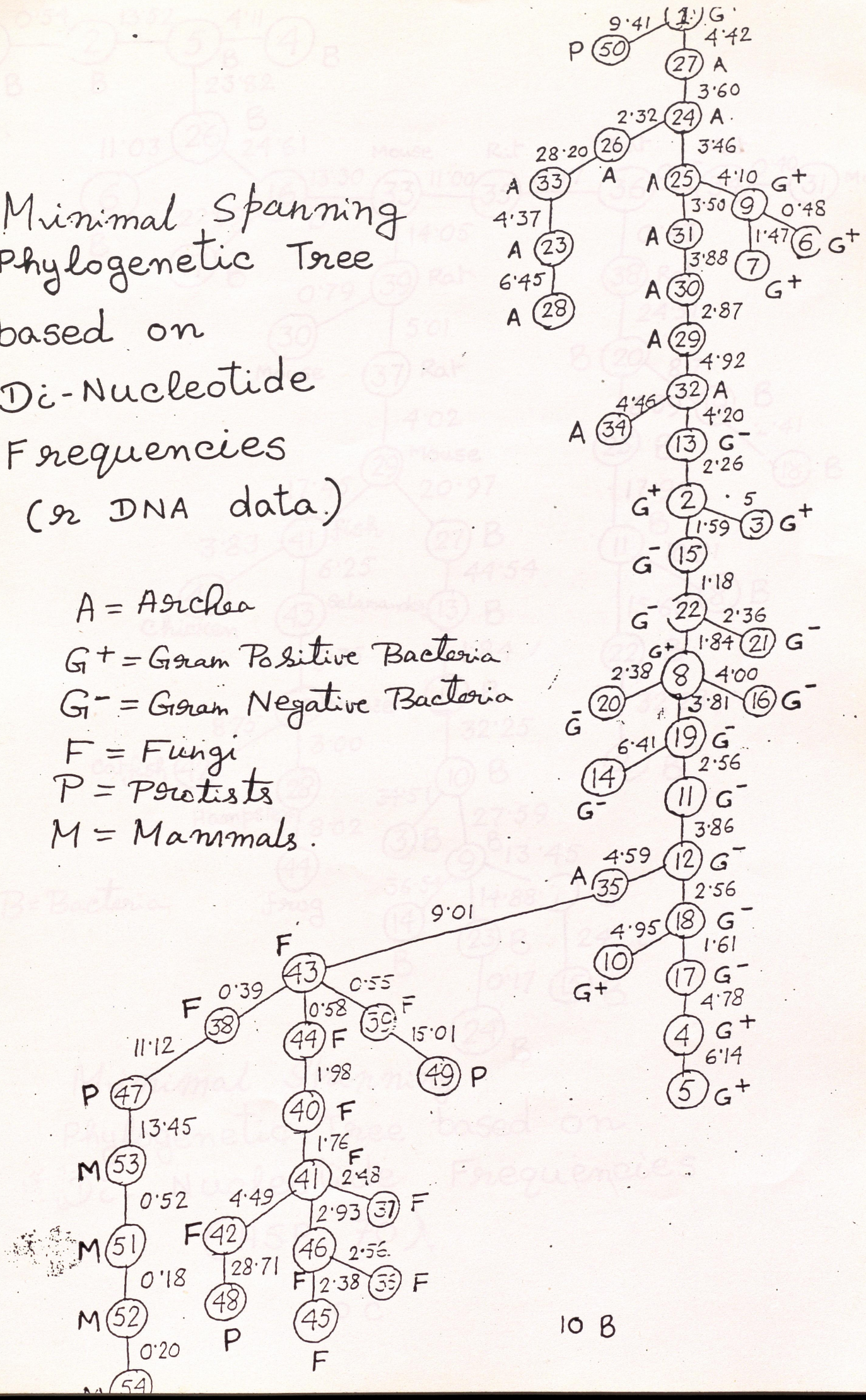

Average Linkage Clustering of Phage Fragments
0
10
20
30
40
50
13
14
15
16
L2
12
11
9
10
L5
4
1
2
3
λ
8
7
5
6
T7

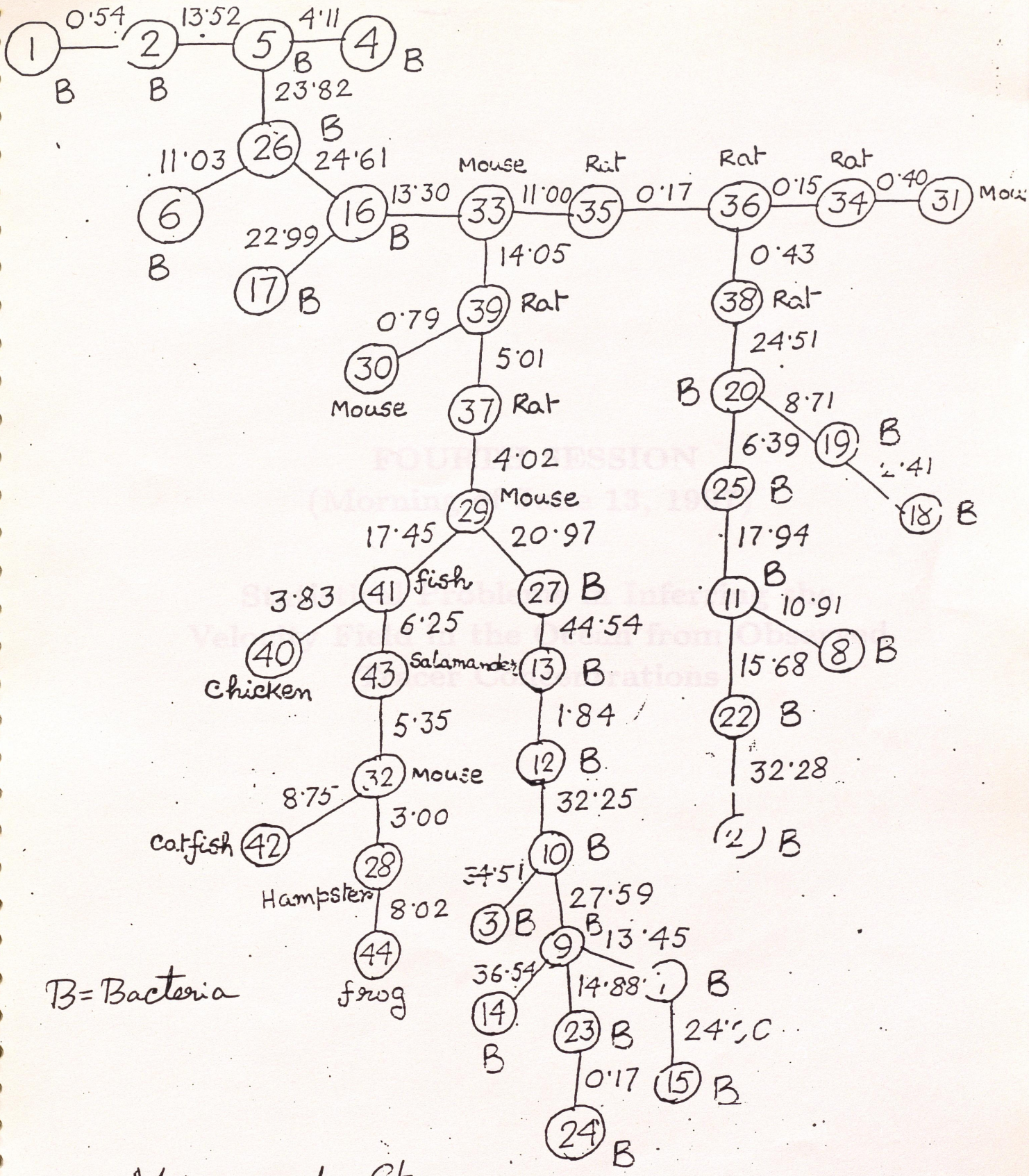


Single Linkage Clustering of Phage Fragments

Minimal Spanning Phylogenetic Tree based on Di-Nucleotide Frequencies (of DNA data.)

A = Archea
G^+ = Gram Positive Bacteria
G^- = Gram Negative Bacteria
F = Fungi
P = Protists
M = Mammals.





B = Bacteria

Minimal Spanning Phylogenetic Tree based on Di-Nucleotide Frequencies (HSP 70)

FOURTH SESSION
(Morning of June 13, 1997)

Statistical Problems in Inferring the Velocity Field in the Ocean from Observed Tracer Concentrations

DR. DEBASHIS SENGUPTA
IISC- BANGALORE

INFERRING THE VELOCITY FIELD IN THE OCEAN FROM OBSERVED TRACER CONCENTRATIONS

The Basic Idea

Traditionally, oceanographers have measured and mapped the temperature (T) and salinity (S) of seawater using research ships. They have also measured concentrations of other tracers such as dissolved oxygen and chemical species such as phosphate and nitrate ions. A present, three-dimensional maps of the distributions these quantities in the world ocean are available. Such maps are not available of the velocity fields in the ocean, because direct measurements of the flow of seawater are hard to make. We propose to obtain some information about the large spatial scale, time mean flow in the Indian ocean using the observed (historical records of) tracer distributions.

The idea is simple - the tracers are advected by the flow, therefore the observed tracer distributions are determined (partly) by the circulation of water in the ocean.[1] Since tracer distributions are influenced by the flow, we consider the inverse problem: Given the distribution of one or more tracers from observations, infer the velocity field that gave rise to this distribution.

About Ocean Circulation

The region of interest to us is the Indian ocean north of 30°S, including the Bay of Bengal but excluding the Arabian Sea. We are interested in that part of the large scale

[1] Apart from the circulation, sources or sinks also influence the tracer distributions. For example, if temperature is considered to be the tracer, its distribution is influenced by solar heating at the sea surface, by diffusion of heat and by advection due to the flow of seawater.

circulation called the meridional circulation. This is the flow in the meridional plane, i.e. the latitude-depth plane (see Fig. 2).

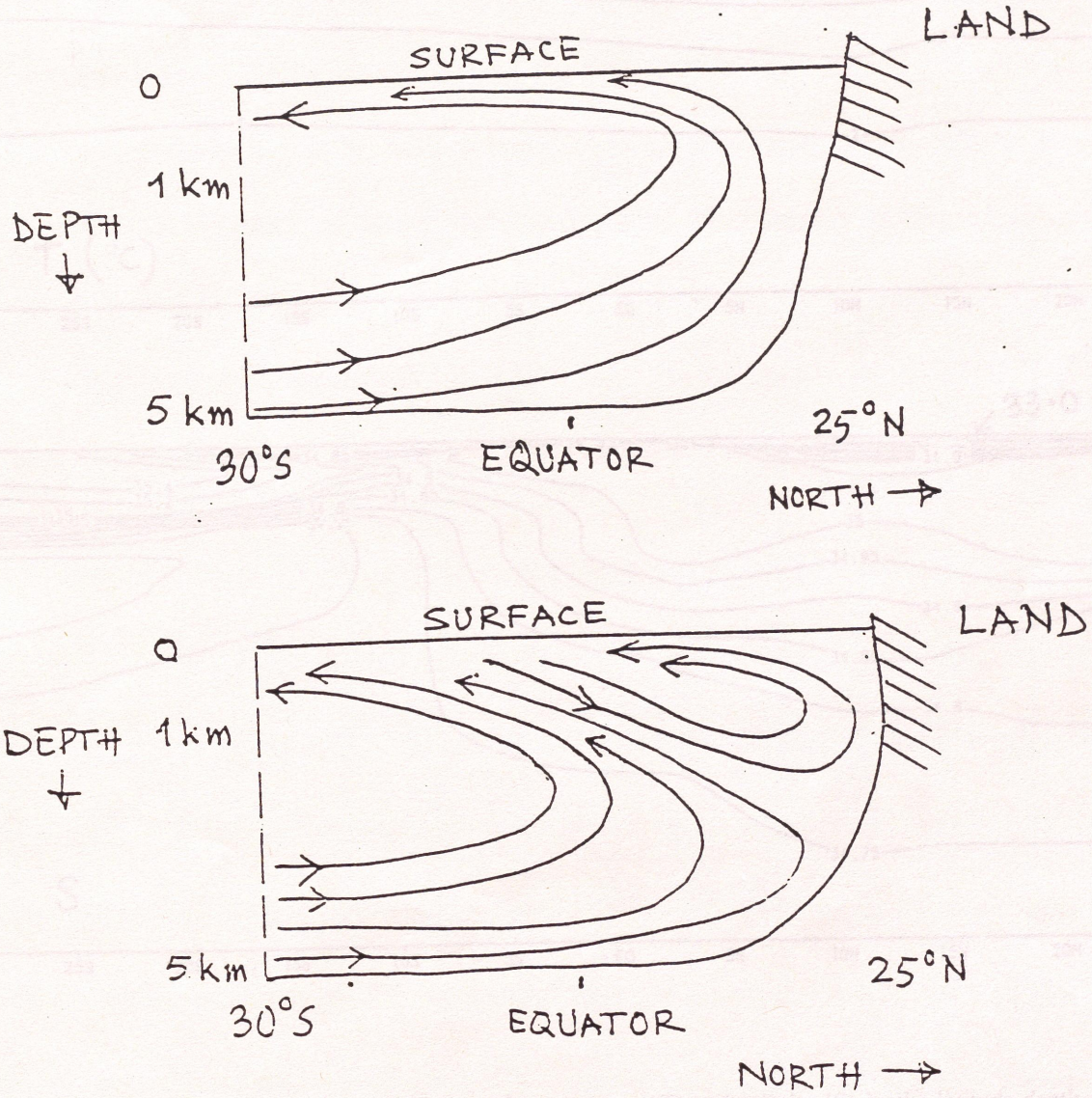


Figure 2: Vertical cross-section through the Indian ocean, showing two possible circulation patterns in the latitude-depth plane. The Bay of Bengal lies in the latitudes between 8^0 North and 23^0 North.

We propose to infer the two-dimensional meridional circulation using the observed distribution of tracers in the latitude-depth plane (see Fig. 3).

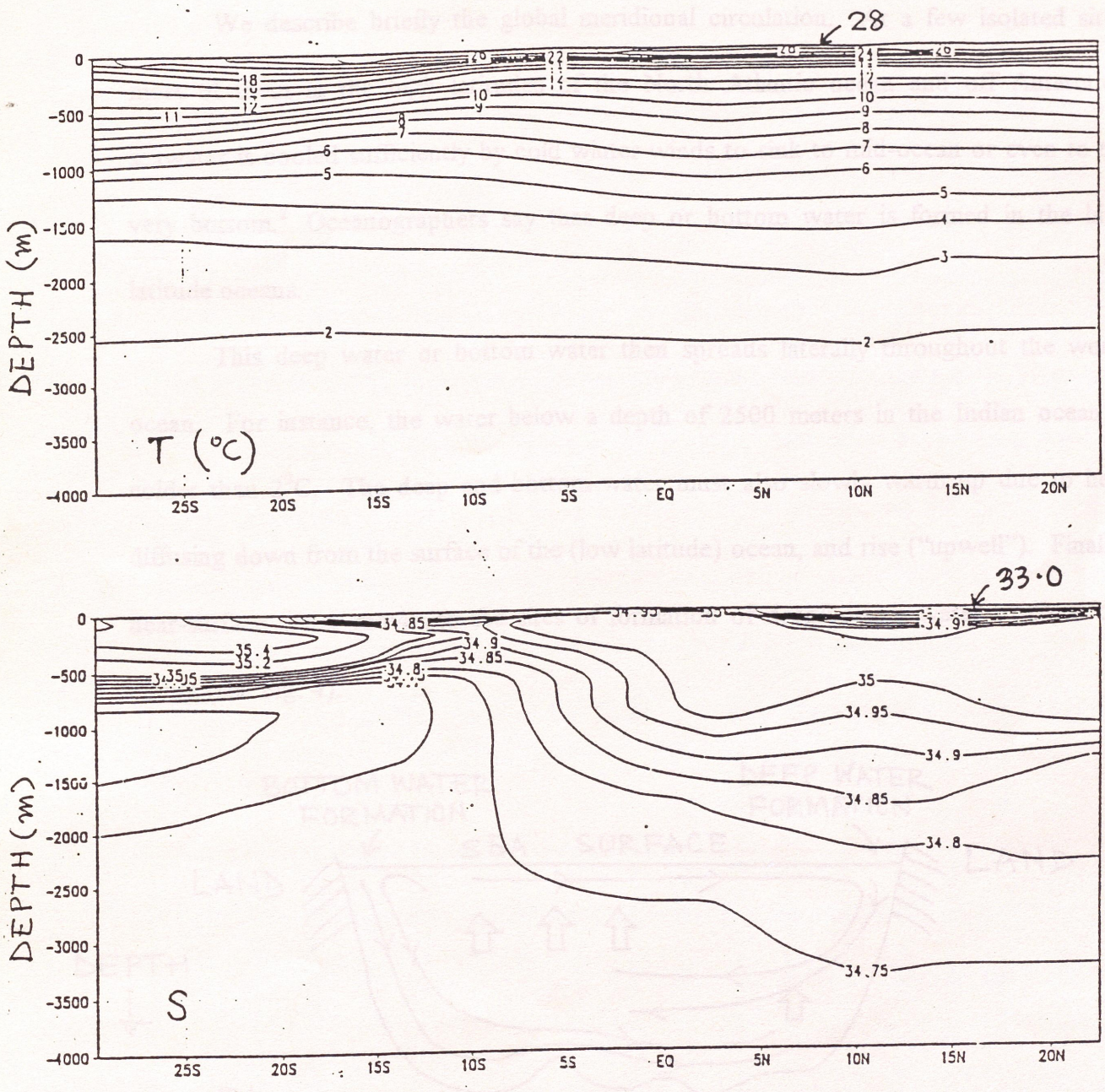


Figure 3: Observed distribution of temperature (T) and salinity (S) in the latitude-depth plane in the Indian ocean. The two dimensional distribution has been obtained by averaging in the east-west direction, i.e. averaging over longitude. Notice T decreases with depth; the deepest part of the ocean is filled with cold water whose origin is in the seas off Antarctica, and the high-latitude North Atlantic. Salinity is measured in dimensionless units of parts per thousand; e.g., a salinity of 35 means that a kilogram of seawater carries 3.5 grams of dissolved salts. The average salinity of seawater is about 35. Notice the low salinities near the surface in the Bay of Bengal - they are due to dilution of seawater by rain and river runoff.

We describe briefly the global meridional circulation. At a few isolated sites, most of them in the high latitudes of the North Atlantic ocean and off Antarctica, seawater is cooled sufficiently by cold winter winds to sink to mid-ocean or even to the very bottom.[2] Oceanographers say that deep or bottom water is formed in the high latitude oceans.

This deep water or bottom water then spreads laterally throughout the world ocean. For instance, the water below a depth of 2500 meters in the Indian ocean is colder than 2^{0}C. The deep and bottom water must also slowly warm up due to heat diffusing down from the surface of the (low latitude) ocean, and rise ("upwell"). Finally, near-surface water moves to the sites of formation of deep/bottom water to close the circuit (see Fig. 4).

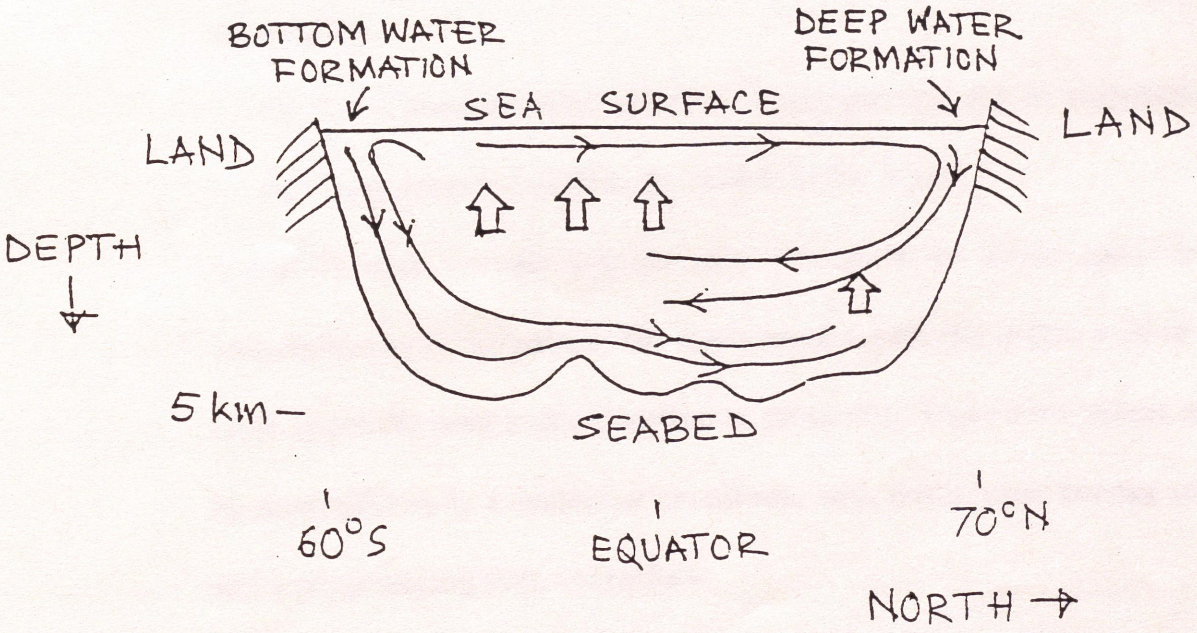


Figure 4: Schematic of meridional circulation in the ocean. Upwelling is shown as wide arrows.

[2] The density of seawater increases with decreasing temperature (and increasing salinity).

There must be a vigorous meridional circulation in the Indian ocean. There are no direct measurements of currents that reveal this circulation, but there are several lines of evidence for its existence.

- The bottom and deep water in the Indian ocean comes from the Antarctic/North Atlantic, entering the Indian ocean from the south. Since the Indian ocean is closed to the north by land, there must be a return flow at shallower depths. An indirect estimate of the flow across 30^0 South in the Indian ocean supports this inference. It shows water below 2000 metres depth moving northwards, and water above 2000 m moving south.
- The Indian ocean north of 15^0 South gains heat through the surface. That is, the quantity

  Q_N = (net incoming solar radiation - latent heat flux due to evaporation - net longwave radiation from the surface of the ocean)

  is positive when averaged over the year. Therefore, for steady annual mean temperature to be maintained, the ocean must export this excess heat to the south (since the basin is closed by land to the north). This export of heat must be accomplished by a meridional circulation, with warm water moving south and a compensating deep, cool inflow.

Direct measurements of this meridional circulation are impossible at present. In principle, however, the known tracer distributions can be used to reconstruct this flow.

OUR APPROACH

We propose to take the following approach to inferring the meridional circulation in the Indian ocean north of, say, 30^0S.

We shall first construct the tracer fields in the latitude-depth plane by averaging the known observed T and S values (at each latitude and depth) over the east-west extent of the Indian ocean basin.

The tracer conservation model is, in latitude-depth coordinates,

$$\frac{\partial T}{\partial t} + \frac{v}{r}\frac{\partial T}{\partial \phi} + w\frac{\partial T}{\partial z} = K\frac{\partial^2 T}{\partial z^2} \qquad (1)$$

$$\frac{\partial S}{\partial t} + \frac{v}{r}\frac{\partial S}{\partial \phi} + w\frac{\partial S}{\partial z} = K\frac{\partial^2 S}{\partial z^2} \qquad (2)$$

where φ is latitude, z is depth, v is the component of velocity of seawater in the direction of increasing φ, i.e. northward speed; w is vertical velocity (positive upwards) and K is a diffusion constant; r is radius of the earth.

T and S are specified at the surface and bottom from observations. T and S are also specified over the full depth at the southern boundary of our domain. That is, the boundary conditions are

$$T = T_{observed}, \; S = S_{observed} \quad \text{at } z = 0, \; z = -H \qquad (3)$$

and

$$T = T_{observed}\ ,\ S = S_{observed} \quad \text{at} \quad \phi = 30°S \tag{4}$$

We then take a guess velocity field

$$V = V_{guess}(\phi, z)\ ,\quad W = W_{guess}(\phi, z)$$

and integrate eqs. (1) and (2), subject to boundary conditions (3) and (4), forward in time. Initial conditions are uniform T and S within the domain $T(t = 0) = T_0$ and $S(t = 0) = S_0$. If the model is integrated long enough, the T and S fields will reach a steady state distribution. The model steady state $T_m(\varphi,z)$ and $S_m(\varphi, z)$ will differ from the observed tracer fields $T_{obs}(\varphi,z)$ and $S_{obs}(\varphi,z)$.

Our problem is then to vary the velocity field so that we minimise the difference between the model steady state T, S and the observed T, S distributions. Notice that with two tracer fields, we can in principle infer the velocity fields $v(\varphi,z)$ and $w(\varphi,z)$. In reality, however, these two fields are not independent - they are connected via the continuity equation.

The condition that sea level does not rise or fall in the mean is

$$W = 0 \quad \text{at} \quad z = 0 \tag{5}$$

Also, at the bottom boundary (assuming it is flat)

$$W = 0 \quad \text{at} \quad z = -H \tag{6}$$

Given the v field, the w field can be obtained from it using the conditions (5) and (6) and the continuity equation in φ and z.

$$r\cos\phi\,\frac{\partial w}{\partial z} + \frac{\partial}{\partial \phi}(v\cos\phi) = 0 \qquad (7)$$

Alternatively, we can introduce the stream function ψ such that

$$v\cos\phi = -\frac{\partial \psi}{\partial z}\,, \quad r w\cos\phi = \frac{\partial \psi}{\partial \phi} \qquad (8)$$

In principle, therefore, we can infer the entire velocity field (i.e. v and w) from the known distribution of a single tracer, either T or S.

The object is to develop the methodology to do this. There are several questions we can ask:

- If we define **e** to be the mismatch between model and observed tracer distribution, how can we minimise **e**? Can we find a unique minimum?
- Suppose the velocity field that minimises **e** for the temperature distribution is $\vec{u}_T$, and the velocity field that minimises **e** for the salinity distribution is $\vec{u}_S$. How different are $\vec{u}_T$ and $\vec{u}_S$?
- What is the range of velocities (velocity fields) that is compatible with the observed T or S distributions? Alternatively, given an uncertainty in the T or S values, what is the uncertainty in the optimal velocity field $\vec{u}_T$ or $\vec{u}_S$?

Forwarded message:
> From dsen Tue Jan 7 14:17:06 1997
> Date: Tue, 7 Jan 1997 12:02:23 +0530 (IST)
> From: Dr D Sengupta <dsen@caos.iisc.ernet.in>
> To: probal <probal@isical.ernet.in>
> Cc: Mohan Delampady <mohan@isibang.ernet.in>
> Subject: Pune meeting write up
> Message-Id: <Pine.LNX.3.91.970107113353.841A-100000@caos.iisc.ernet.in>
> Mime-Version: 1.0
> Content-Type: TEXT/PLAIN; charset=US-ASCII
>
> Dear Probal ,
> I saw your mail from Carolina . My problem is about inferring the
> optimal flow field (the flow is unknown) from tracer distributions (the
> tracer distributions are known). The problem with tracer distributions
> in the ocean are that they are based on very few data - therefore the
> atlases contain a lot of interpolation , both in time and space . So I
> have taken an approach that does not involve estimating gradients of the
> tracer concentrations from the observations .
> Of course , as you have noticed , there are terms involving tracer
> gradients in the simple model ; but the model is meant to compute
> the tracer concentrations and their gradients INTERNALLY , without any
> observed tracer information being input (except at the boundaries) .
> The gradients all pertain to MODEL tracer fields . The problem you mention ,
> relating to the estimation of gradients from the observed tracer fields ,
> is of course an important open problem in oceanography . But it is
> distinct from the problem I posed .
> June is fine for the meeting . Please let me know if there are any
> questions . With regards ,
>
> - Debasis
>

Dear Debashis and Mohan :

I have given some more thoughts to the problem that Debashis sent to us. First of all, I need answers to the following questions in order to get things clarified. It should be easy for Debashis to answer them.

(A) Are there known analytic solutions to the system of partial differential equations (1) & (2) connecting the flow fields with the tracer fields ? If the answer is yes, could you cast some light on how to get those analytic solutions for given flow fields. If analytic solutions are not possible to get, are there known ways of solving the equations numerically to determine the "model tracer fields" for given flow fields ? How are such numerical solutions computed ?

(B) Some of us are a bit confused with the statement "if the model is integrated long enough, the T and S fields will reach a steady state distribution". Specifically, we would like to know what is the implication of that on the partial derivatives of T and S w.r.t. t that appear in equations (1) and (2) - are these partial derivatives assumed to be zero or a constant in the steady state ?

I understand Debashis' concern about estimating the derivatives of the tracer fields from the data in view of the fact that tracer distributions are based on very little data and atlases contain an excessive amount of interpolation. Now, if we look at equations (1) and (2) they can immediately lead to neat analytic expressions for V and W in terms of various partial derivatives of T and S. So it is tempting for people to try to estimate those derivatives from whatever data that is available and that will lead to some estimates for the flow fields. The interesting point is that such an estimate (which Debashis does not like and he has very good reasons for that) will satisfy equations (1) and (2) 'perfectly'. The advantage in Debashis' approach is that it avoids the derivatives of the tracer fields being estimated from the data but one has to settle question (A) above (how easy is that ?). Also, the "mismatch between the model and the observed tracer fields" discussed in the last page of the write up does not seem to be a simple object to be minimized - especially, if we need to solve the minimization problem over a very large family of candidate flow fields. This is again tied to question (A) above.

I will stop here and let you two speak. I am sending copies of our correspondence to statisticians in Calcutta with whom I discussed the problem before coming to Chapel Hill, North Carolina.

Best regards----------PROBAL

PS. Please send e-mail to probal@stat.unc.edu - mails sent to Calcutta address do get forwarded here but it causes unnecessary delay. Thanks.

From dsen@caos.iisc.ernet.in ...
Date: Thu, 20 Feb 1997 12:03:15 +0530 (IST)
From: Dr D Sengupta <dsen@caos.iisc.ernet.in>
To: probal@stat.unc.edu
cc: Mohan Delampady <mohan@isibang.ernet.in>
Subject: probal's msg.

Dear Probal , Mohan ,

I have been thinking about our correspondence so far . It seems that we might really have two problems in hand . One is what I suggested , with all derivatives computed from the model rather than the data . The second problem is the one Probal describes so clearly in the last mail - express v and w in terms of the temperature and salinity fields using the tracer conservation equations , and estimate them (i.e. the velocities) from whatever data we have . The second problem might be worth doing as well, because one cannot say that the data is definitely too meagre to attempt this approach . Also , since we are interested in the time-mean flow , the time derivative terms may be dropped from the tracer conservation equations in this approach .

Now to take up Probal's questions A and B .

(A) There are no analytic solutions to the equations that I know of . For even the simplest flow fields (v,w), no simple approach will allow analytic solutions to be got . There are , however , standard numerical methods to solve the advection-diffusion equations for the tracer fields, for a given velocity field of general form . These should be very easy to implement , and will not involve heavy computation .(The temperature and salinity fields in the ocean have correlation scales of the order of 100 kilometres in the horizontal , and 50 metres or more in the vertical. So our grid will have about (60x40) points only . Moreover , the problem is linear .)

(B) We expect that as the evolution equations for the tracer fields are integrated in time , starting from spatially uniform initial conditions, eventually a balance will be approached . This is a balance between advection, and diffusion from the boundaries . As the balance is approached more and more closely , the rate of change of the tracer fields will become slower and slower . So we could define a steady state by examining the size of the time derivative term - if it falls below a certain value everywhere in the domain , we could declare steady state .

To answer Dr. Diwanji's point , I think there is enough data to meaningfully compare model and observations . However , there may not be enough data to obtain meaningful estimates of spatial derivatives of the tracer concentrations (although this needs to be examined - see the first paragraph above) . Since the spatial gradients of the tracer fields are small in the deeper parts of the ocean, it is possible that a range of velocities (v or w) might be compatible with the tracer data .

One of the goals should be to ascertain this range.

My feeling is that we might not have to examine a very large number of different PATTERNS of flow - the one-cell or two-cell pattern I sketched in my write up might do (or maybe there is a three-cell pattern).

We know definitely that the flow in the uppermost layer is southward because it is directly driven by the winds ; we also know that the slow flow in the deepest layer is northward because the seawater there bears the characteristics of Antarctic water.I am hoping that the error-minimising algorithm will help to discriminate between the patterns, and to deduce the strength of the cells.

Regards ,

_ Debasis

I had a small chat with Ramasubramanian just now. What follows is what I understood there.

Consider (1) of DS. Changing the notation slightly and calling ϕ as z_1 and z as z_2, rewrite it as follows:

$$\frac{\partial T}{\partial t}(z_1, z_2) = \frac{K^2}{2}\frac{\partial^2 T}{\partial z_2^2}(z_1, z_2) - v(z_1, z_2)\frac{\partial T}{\partial z_1}(z_1, z_2) - w(z_1, z_2)\frac{\partial T}{\partial z_2}(z_1, z_2). \quad (1)$$

This is the forward Kolmogorov equation corresponding to a *degenerate* (because $\frac{\partial^2 T}{\partial z_1^2}$ doesn't appear) diffusion

$$\begin{aligned} X_1(t) &= \int_0^t v(X_1(s), X_2(s))\, ds + X_1(0), \\ X_2(t) &= KB_2(t) + \int_0^t w(X_1(s), X_2(s))\, ds + X_2(0). \end{aligned} \quad (2)$$

This is a 2-dimensional diffusion with infinitesimal dispersion $\begin{pmatrix} 0 & 0 \\ 0 & K \end{pmatrix}$ and infinitesimal drift $\begin{pmatrix} v(x_1, x_2) \\ w(x_1, x_2) \end{pmatrix}$.

Problem is to estimate the drift. The top and bottom boundaries of the diffusion are known. It may be better to eliminate families of functions from the solution set rather than trying to find the solution directly. Make some reasonable guess at the two vertical boundaries also (?). Then try to find solutions to (2). From (5), (6) and (7) of DS $w(.,.)$ seems to be a periodic function in the second variable with period H. So, start with a family like $\begin{pmatrix} v(x_1, x_2) \\ w(x_1, x_2) \end{pmatrix} = \theta \begin{pmatrix} \hat{v}(x_2) \\ \tan(x_1) \int_0^{x_2} \hat{v}(u)\, du \end{pmatrix}$, where $\hat{v}$ is a known function such that $\int_0^{x_2} \hat{v}(u)\, du$ is periodic (with period H). More generally one could consider $\begin{pmatrix} v(x_1, x_2) \\ w(x_1, x_2) \end{pmatrix} = \theta A(x_1, x_2)$ where A is a known pair of functions with the second being periodic (with period H). Instead of a scalar θ one could have more parameters also.

Now use (4) of DS as data to obtain an MLE or some such estimate of θ.

As you can readily see, I haven't worked much on this approach. Try getting some help from Rajeev or BV.

Date: Thu, 24 April 1997

Debashis gave an excellent lecture this afternoon. The physics of oceans became almost entirely clear to all of us. This is not to say that we got any nearer to solving his problem! Some of the points that were discussed are the following:

(1) What about writing (v, w) as a smooth function f of (T, S, x) and based of some more (?) data on (v, w) can we estimate f?

This was shot down immediately by DS with the argument that x above will include pressure gradients which are not available.

(2) What about dropping $\frac{\partial S}{\partial t}$ and $\frac{\partial S}{\partial t}$ from the conservation equations and expressing (v, w) in terms of the other partials, then estimate these partials from observed data?

You must have noted that this was discussed in our correspondence earlier. DS now thinks that this may be ok, but one has to see what pattern emerges and then only its merits, if any, can be judged.

(3) Ramasubramanian thought that if $\frac{\partial S}{\partial t}$ and $\frac{\partial S}{\partial t}$ are dropped then the problem is perhaps estimating the drift term of a diffusion process and on this there is much literature. I am yet to understand this and will give it some thought in the coming days.

(4) Do the conservation equations (along with the boundary conditions) definitely indicate whether the flow patterns are single cell or multiple cell?

DS didn't know. I feel that this information may not be available in those equations in which case the velocity field will be greatly influenced by what pattern is assumed.

(5) Numerical solution of the conservation equations subject to boundary conditions is still what DS thinks is the most reasonable approach. Then he says minimizing the 'error' between the stable solution and the observed pattern is where he wants the statisticians to help him.

ments, both of S and T tend to have lots of smoothness and regularity as functions of Φ and z. So, estimation of their derivatives from the data using standard nonparametric smoothers might not lead to very bad results. Also, measurements are comparatively less sparse in the upper layers of the ocean, and it is the velocity field at the upper layers that has significant influence on the climate.

Equations (1) and (2) in the writeup of DS are degenerate Fokker-Planck equations. Also the boundary conditions (of Dirichlet type) are prescribed on $z = 0$, $z = -H$ and $\Phi = 30^0 S$ but not on $\Phi = 25^0 N$ ($25^0 N$ corresponds roughly to land at the west of Bay of Bengal). Note that one can associate a degenerate diffusion process with the Fokker-Planck equation. So the problem is one of estimating the dirft of the diffusion with the dispersion being unknown. To make the problem more tractable one can impose a reflecting boundary condition (i.e. Neumann type condition) at the boundary at $\Phi = 25^0 N$. This means that there is no flux across the land. Also, such a condition seems reasonable in view of Fig. 3 given in DS's writeup. To make the problem even more tractable, we can add a diffusion term in the Φ direction (i.e. a second order term in Φ-variable). This would make the equation and the diffusion process non-degenerate. Thus we may try to model the basic dynamics of the problem by a non-degenerate diffusion process with normal reflecting boundary in a compact domain containing $[z = 0, z = -H] \times [\Phi = 30^0 S, \Phi = 25^0 N]$, whose boundary includes $[0, -H] \times \{25^0 N\}$. Also the continuity condition (7) implies that the Lebesgue measure is the stationary distribution. So one can perhaps meaningfully consider stationary or steady state analogues of (1) and (2). This would also perhaps justify to some extent estimation of the gradient or second order derivatives of T and S w.r.t Φ and z from the data available. Making the problem non-degenerate and imposing Neumann condition at $\Phi = 25^0 N$ will enable us to obtain (at least in principle) unique solutions to boundary value problems with given V, W, K and boundary data. One can give a stochastic representation for the solution.

From the available data (or part of it) one can have preliminary estimates of V and W. On the basis of these estimates, one can try to guess a parametric family of V and W. Corresponding to any V and W belonging to this parametric family, one may solve the non-degenerate steady-state boundary value problem by numerical or probabilistic methods. Matching against the available data for T and S, one can then hope to get hold of optimal value(s) of the parameter(s) involved.

NOTE : The following points were noted during the meeting. While the data is very sparse (especially in the deep sea) for S and T measure-

FIFTH SESSION
(Morning of June 13, 1997)

Statistical Problems in Chemical Engineering Kinetics : Nonlinear Experiments, Model Discrimination, Parameter Estimation and Inference.

DR. R. JAGANATHAN
CHEMICAL ENGINEERING DIVISION, NCL – PUNE

CHEMICAL KINETICS:

* DEALS WITH RATE OF CHEMICAL REACTIONS

* THE RATE OF A CHEMICAL REACTION IS MEASURED AS THE CHANGE IN CONCENTRATION PER UNIT TIME

* KINETIC STUDY INVOLVES DETERMINING THE DEPENDENCE OF THE RATE AS A FUNCTION OF VARIOUS OPERATING PARAMETERS

$$R = dC/dt = F(Ci, T, W)$$

* CHEMICAL REACTIONS CAN BE SIMPLE OR COMPLEX

1) $A \text{ ----------> } B$

$R = k C_A$

2) $A + B \text{ -------> } C$

$R = k C_A C_B$

3) NET WORK: $A + B \text{ -----> } C$

$A + C \text{ -------> } D$

$R = k C_A C_B / (1 + K_B B + K_C C)$

MAIN OBJECTIVE IS TO DETERMINE THE PARAMETERS k_i K_A , ETC TO ARRIVE AT THE CORRECT FORM OF RATE EQUATION

EXPERIMENTAL PROCEDURES:

REACTORS: BATCH AND CONTINUOUS REACTORS

METHOD OF ANALYSIS: DIFFERENTIAL AND INTEGRAL

Table 2

RATE EQUATIONS USED AND PARAMETERS OBTAINED

Sr. No.	Rate Equation	Temp - erature. K	k_1	K_A (m^3/kmole)	K_B x 10^{-3} (m^3/kmole)	ϕ_{min} x 10^{14} (kmole/m^3/s)^2
I	$\frac{wk_1 A^* B_l}{(1+K_A A^*+K_B B_l)^2}$	313	4.05×10^5	1.18×10^5	7.01	0.97
		333	1.60×10^5	1.43×10^5	2.79	5.57
II	$\frac{wk_1 A^* B_l}{(1+K_B B_l)^2}$	313	1.98×10^5	–	9.64	9.25
		333	1.38×10^5	–	2.80	25.50
III	$\frac{wk_1 A^* B_l}{(1+K_A A^*+K_B B_l)^2}$	313	2.04×10^5	1.66×10^5	2.92	1.09
		333	1.71×10^5	1.97×10^4	2.33	5.26
IV	$\frac{wk_1 \sqrt{A^*} B_l}{(1+K_A \sqrt{A^*}+K_B B_l)^2}$	313	1.79×10^4	1.89×10^3	21.10	1.17
		333	7.55×10^4	1.10×10^3	29.90	20.73
V	$\frac{wk_1 \sqrt{A^*} B_l}{(1+K_B B_l)^2}$	313	1.60×10^2	–	3.75	1.89
		323	4.00×10^2	–	2.70	4.02
		333	6.15×10^2	–	2.29	6.03
VI	$\frac{wk_1 \sqrt{A^*} B_l}{(1+\sqrt{K_A A^*})+(K_B B_l)^2}$	313	1.74×10^5	1.78×10^5	2.91	8.23
		333	1.22×10^5	3.92×10^5	2.64	33.22

Note : Units of k_1 for Eqs. I, II and III are (m^3 / kg) (m^3 / kmol) (1 / s) and for Eqs. IV, V and VI are (m^3 / kg) (m^3 / kmol)$^{0.5}$ (1 / s)

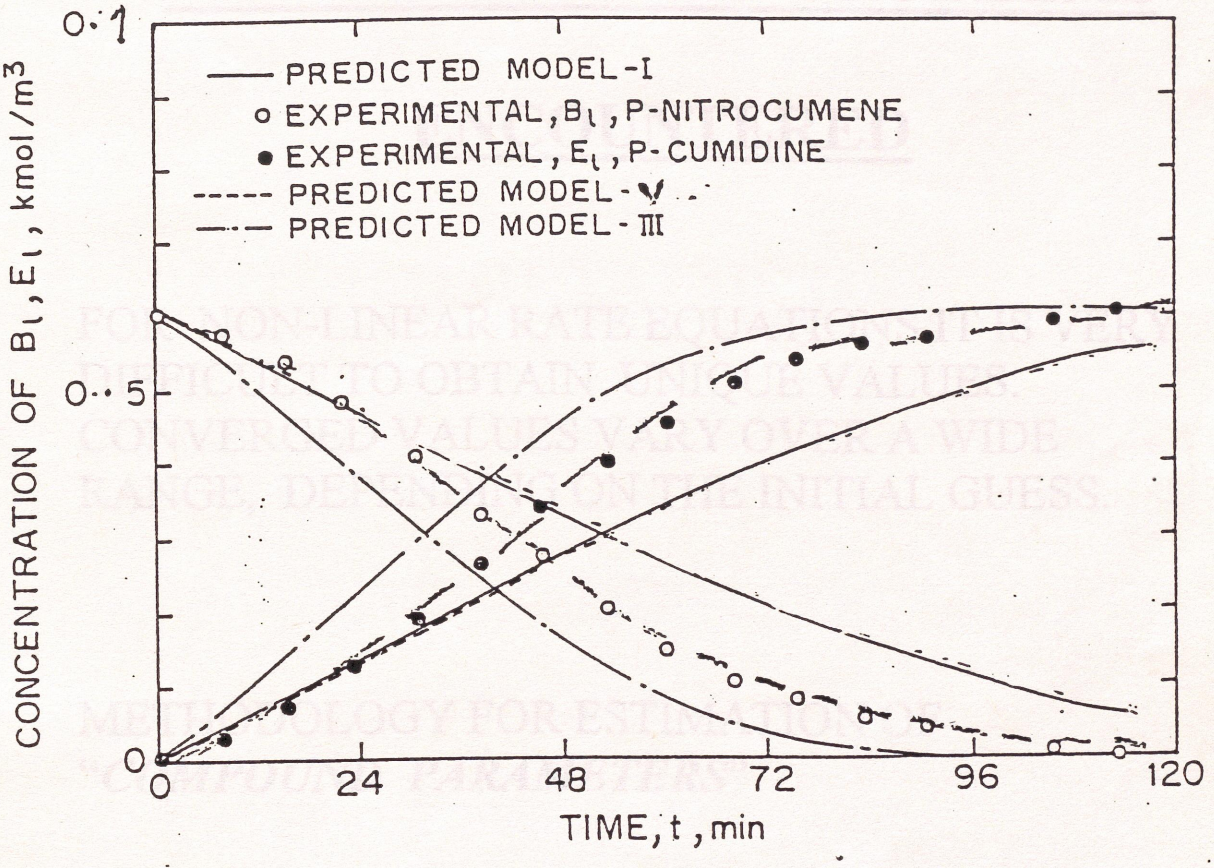


Figure 6 : LIQUID REACTANT / PRODUCT CONCENTRATION PROFILE AT 333, K FOR MODEL I, III AND V
CATALYST LOADING, W = 1·66, kg/m³
H_2 PRESSURE, PH_2 = 790, kPa

DIFFICULTIES AND PROBLEMS

ENCOUNTERED

FOR NON-LINEAR RATE EQUATIONS IT IS VERY DIFFICULT TO OBTAIN UNIQUE VALUES. CONVERGED VALUES VARY OVER A WIDE RANGE, DEPENDING ON THE INITIAL GUESS.

METHODOLOGY FOR ESTIMATION OF "***COMPOUND PARAMETERS***"

FOR INSTANCE:

$$R = \frac{k1 * k2\ (C1*C2)}{[(1 + (K_A + K_B)C2)]}$$

SIGNIFICANCE (PHYSICS) AND

IMPORTANCE OF :

* VARIANCE
* STANDARD DEVIATION
* CONFIDENCE LIMIT
* AVERAGE ERROR
* t - TEST
* $\in$ - TEST

AS APPLIED TO MULTIVARIANT / NON LINEAR PROBLEMS

STATISTICAL DESIGN OF EXPERIMENTS

IS IT POSSIBLE TO DECIDE ON THE NUMBER AND TYPE OF EXPERIMENTS *APRIORI* FROM THE TYPE OF THE RATE MODEL ?

SUMMARY OF COMMENTS AND SUGGESTIONS FROM STATISTICIANS AND PROBABILISTS

The problem of "compound parameters" is a "statistical nonidentifiability problem". Unless some additional information is available it is not possible to estimate statistically each component of a compound parameter from the data. This problem cannot be handled statistically – it needs to be handled in a different way.

The problems of multiple minima and lack of convergence of the iterative procedures as well as the dependence of convergence on the choice of initial values are well known to anyone working in non-linear least squares. It is difficult (virtually impossible) to provide a general solution to the problem. Standard statistical software packages, which can perform nonlinear least squares, too suffer from this problem for general models. However, since the nonlinear models that arise in chemical kinetics are of certain specific types one may try to develop some specialized versions of nonlinear least squares algorithm that will be suitable for those specific problems and will have good convergence properties. Also, using the specific nature of the nonlinear model, one may try to choose the initial values for starting the iteration. This appears to be a potential research problem.

The problem of efficient designing of an experiment for efficient estimation of parameters in a well specified nonlinear parametric model has received a fair amount of attention in statistical literature and some adaptive sequential designs are known and well studied in this context. On the other hand, the problem of efficient designing of experiments for model discrimination has not received that much attention in the statistics literature. This is a challenging problem that calls for careful research. One needs to deal with the cases for nested family of models and non-nested family of models separately and carefully. When it comes to choosing a model by comparing residual sum of squares one needs to be careful when the competing models are nested. Another important fact is that the experiment that emphasizes on discriminating among several models will typically be concentrated at points of maximum separation of the competing model functions. As a result, such a design will not be appropriate for efficient estimation of parameters in any of the competing models. One therefore needs to adopt some form of batch sequential strategy to switch back and forth between the optimal design for model discrimination and optimal design for parameter estimation in a given model. The entire problem needs to be considered carefully with a great deal of analytic investigation as well as numerical studies. It so seems that there are several promising research problems that can be pursued here.

SIXTH SESSION
(Afternoon of June 13, 1997)

During this session three speakers spoke, and each talk was followed by discussion from the audience. The first speaker was a statistician actively involved in industrial consulting, the second speaker was an ecologist and the third speaker was a biologist.

INDUSTRIAL APPLICATIONS OF STATISTICS

BY
A. D. DHARMADHIKARI

ABSTRACT

I am going to discuss two problems :

1. DEPARTURE OF LUBRICATION OIL PRESSURE FROM THE TARGET VALUE IN AN OIL ENGINE :

An oil engine producing company in Pune has observed that the lubrication oil pressure (LOP) is consistantly low. This can happen due to various reasons, viz., porosity within and outside the engine, departures from target values of components in the LOP system, and above all wrong engineering design.

Major components in lubrication oil passage are sump, suction- tub, delivery body, pump, relief valve, filter, filter head, cooler, cooler pipes, oil holes in gear casing and crank case, crank shaft, cam shaft, bearings, and connecting rod.

Metallergical and geometrical specifications of parts are not always based on rigorus mathematical derivations. Further, the parameters like hardness, strength, porosity, blow holes do affect machining of the components. Hence by considering only the geometrical dimentions LOP cannot be achieved.

We handled this problem using following statistical methods.

1. Design failure mode and evaluation analysis (FMEA).
2. Process FMEA.
3. Sampling plans for bought out components and
4. Statistical design of experiment.
   (Fractional factorial experiment with specified main effects and first order interactions).

2. ANALYSIS OF FAILURE DATA OF JIB CRANES TO PROPOSE THE PREVENTIVE MAINTENANCE SCHEDULE :

The problem dealt here is about the jib cranes used in an antomobile unit. To begin with we furnish some information of jib cranes.

Jib cranes are used in machine shop for circular moments of loads around the machines.

A jib crane can be divided into two subassembles, a mechanical and an electrical.

The mechanical subassembly, has a column and an arm resting on the top of the column. The moment of the arm depends upon the ball-bearings. The other end is free. Rating of the arm, is, the maximum load that can be picked up by the crane. This value depends upon the deflection of the arm. The deflection is measured periodically which is used to revise the rating of the crane. Major mechanical failure is wear and tear at the ball bearings due to friction.

The electrical subassembly has two control panels, one on the column, other, having free moments, called operator's panel and the motor, stationed in the hoist - on the arm. Cables are from main switch to operator panel via control panel on the column and hoist. There is a mild steel chain from the hoist to operator panel. The length of chain must be less than the length of the cable (wire) connected hoist and operator panel.

The hoist contains motor, gear box and the brake unit. Further, breake unit has spring and brake paddles. When the operator switchs on the motor, first the brakes are pused back and then motor starts. It is expected that when the operator switches off the motor, first the brakes will operate, which will stand still the verticle motion of the load and then motor stops.

It is expected that 80% of the failures will be either due to non-functioning of the brake-unit and electrical faults in the operator panel.

Hence, in following study, the failure modes are limited to two-mechanical (0) or electrical (1) only.

Hunch by the maintenance engineer was that chances of failure of an electrical failure followed by an electrical failure are more.

We have data on two cranes, referred to as A and B, hencefurther. Both of these are used for loading-unloading the crank-cases. Crane A is used for 3/4 - cylinder-crank-cases, while B is used for 6/8 - cylinder - crank-cases.

To do the analysis, we started with the question, whether the data from cranes A and B be pooled ?

Using tests from proportion it was found that the value 3.17, of the test statistics was significant at 5% level of significance. Hence the data of two cranes were analysed separately.

The individual data sets were, then, subjected to welknown tests for trend and dependence. The test statistics, and results are summarised in section 2. These tests indicated that data are not from iid set up.

Hence, we decided to recognize the bivariate nature of the data, viz., type of failure and the time gap between successive failures.

The theory of semi Markov process, (SMP) states that, if type of failures are homogeneous first order Markov-chain and given the types of failures, successive time gaps are independent the realization may be from SMP. Hence, it was decided to concentrate firstly on the stochastic processes induced by the types of failures say $\{X_n^i,\ n \geq 1\}$, $i = A,B$.

We came to the conclusion, that it was not possible to accept the first order Markov model.

However, 2nd order Markov chain model was accepted with in the class of 3rd order Markov chain in 5% level of significance.

Consequently, the model proposed to predict the failure pattern is more general than semi Markov process. Details of these are in rest of the paper.

How to estimate the ammount of a 'suicidal' substance in the parent cell, given the lifespans of the two daughter cells.?

DR. N. V. JOSHI (IISc, Bangalore)

Many of the cells in living organisms have specific life spans. In fact, a process called 'apoptosis' (programmed cell death) initiated by the cell itself, at an appropriate time, leads to the death of the cell. Apoptosis plays a very important role in the orderly development of the organism.

Prof. A. Antony of the Department of Microbiology and Cell Biology, Indian Institute of Science, Bangalore , and his student Dr. Mahavir Bhojani, have conducted a set of experiments, where they have monitored the fates of two daughter cells, which originate by the division of a single (mother) cell. Both the daughter cells undergo apoptosis, i.e., they self-destruct after a certain ammount of time, of the of the order of several hours.

What are the factors that determine the lifespans of the two cells? One of the simplest hypothesis is that the ammount of a certain substance inherited from the mother cell governs the life span; e.g. higher the ammount, shorter the life span. If the substance (ammount S in arbitrary units) were to be uniformly distributed throughout the mother cell (or distributed in a spherically symmetrical manner), then each of the two daughter cells (formed by the division of the mother cell) would inherit an equal ammount, and the expected lifespans of the two should be equal. If the substance were to be asymmetrically distributed (say uniformly, but restricted to a hemisphere), the difference between the ammount inherited by the two cells would vary between 0 and S, the total ammount of the substance.

The statistical question of interest is whether one can infer the value of 'S', (expressed as 'f', the fraction of the cell-volume, or cell-surface), given the data on lifespans.

The observable data consists of 2n lifetimes 'T', denoted by ,

T1i and T2i (i=1, 2, ... n)

obtained from the 2n pairs of daughter cells, starting with n mother cells, and the derived difference 'D' between the pair of life spans

Di (i=1, 2, ...n) where Di = abs (T1i - T2i)

Let us assume that the cell is a sphere, and that the the substance is uniformly distributed in some restricted portion of the cell. For example, it could form a single, regular patch on the cell with a circular boundary. Assume that it occupies a fraction 'f' of the surface. Assume that the two daughter cells are obtained by a bisecting the sphere by a 'random' plane passing through the center of the sphere, and that the ammount of the substance inherited by the

two cells is f1 and f2 (where f1+f2=f).

For a known value of 'f', it should be possible (for probability theorists and/or statisticians) to obtain the distribution of f1 and f2, and in turn, the distribution of the differences between the lifespans of the two daughter cells.

The more important question is of estimating 'f' from the values of T1i, T2i, i=1, ...n, and that too should be amenable to analytical solution by probability theorists/ statisticians. A simulation based brute force approach would definitely lead to the solution of the problem, and can easily be employed as a last resort.

{ There may be more and better questions that can be asked of this data. However, this can be done most effectively by examining the data itself. Unfortunately, my e-mail to the two investigators has not as yet led to any response, so at the moment the data is unavailable. }

ELSEVIER

Biochimica et Biophysica Acta 1281 (1996) 189–204

BBA Biochimica et Biophysica Acta

Assessment of molecular sieving across bacterial outer membrane of *Pseudomonas*

Satish B. Kulkarni, V. Somlata, V. Sitaramam *

Department of Biotechnology, University of Pune, 411 007 Pune, India

Received 22 November 1995; revised 15 January 1996; accepted 15 January 1996

Abstract

The role of the permeability barrier of the outer membrane of *Pseudomonas* was re-evaluated based on the physical theory of molecular sieving in view of its intrinsic antibiotic resistance. We developed a set of analytical procedures based on parametric and non-parametric statistical tests to evaluate, validate and adopt the better among a set of competing non-linear models of diffusion. The molecular mass dependence of uptake of non-electrolytes in bacteria yielded a quantitative measure to distinguish between sieving mechanisms and specific uptake/efflux mechanisms. The experimental data, supported by the physical model of DEAE-Sephadex and various analytical models and extensive simulation of the errors, both in measurement and models, yielded evidence consistent with the relaxation of the outer membrane matrix barrier in *Pseudomonas*.

Keywords: Renkin's equation; Pore size; Error simulation; Outer membrane; Nonelectrolyte; Diffusion; (*P. aeruginosa*)

SEVENTH SESSION
(Afternoon of June 13, 1997)

Concluding Remarks and Discussion

The following points were noted during the concluding session :

Such meetings can be used to identify specific thrust areas of multidisciplinary research which will involve scientists and engineers from various fields on one hand and statisticians and probabilists on the other. It was suggested that problems that arise in Electrical Engineering, Physics and Meteorology frequently involve statistical data analysis and probabilistic modeling. In future, attempts should be made to identify potential research problems in those areas with definite statistical and probabilistic components. It will be appropriate to make some efforts to contact some people from those disciplines who might act as resource persons for this purpose.

The need for making serious attempts to follow up on the problems that were discussed during different sessions was appreciated by all participants. All of the major problems discussed generated stimulating discussion and exchange of ideas. Many of the participants expressed their interest in carrying out further work on various research problems discussed. Small groups were formed among the participants with the objective of looking at specific problems and pursuing collaborative research. It is now hoped that this meeting will give rise to some successful research projects.

BRAINSTORMING SESSION
ON
SAMPLING METHODOLOGY FOR MONITORING THE BIODIVERSITY OF EASTERN AND WESTERN GHATS

Venue : Conference Hall
Jawaharlal Nehru Centre for Advanced Scientific Research, Indian Institute of Science Campus

September 16 , 1997

10:00 - 10.30 a.m.	Registration	
10.30 - 11.15 a.m.	Introduction	Madhav Gadgil
11.15 - 11.30 a.m.	*Coffee*	
11.30 - 12.00	Strata for locatiing study sites	Utkarsh Ghate
12.00 - 12.30 p.m.	Focal taxa	P. Pramod
12.30 - 1.00 p.m.	Earthscape elements	Harini Nagendra
1.00 - 2.00 p.m.	*Lunch*	
2.00 - 5.30 p.m.	Parallel sessions of the Working Groups on Strata, Focal Taxa and Earthscape Elements	

September 17 1997

Presentations of Recommendations of the Working Groups

9.30 - 10.15 a.m.	Strata
10.15 - 11.00 a.m.	Focal taxa
11.00 - 11.15 a.m.	*Coffee*
11.15 - 12.00 noon	Earthscape elements
12.00 - 1.00 p.m.	Study methodology
1.00 - 2.00 p.m.	*Lunch*
2.00 - 3.00 p.m.	Networking
3.00 - 3.15 p.m.	*Coffee*
3.15 - 4.30 p.m.	Open discussion

DISCUSSION MEETING ON MONITORING BIODIVERSITY OF EASTERN AND WESTERN GHATS

Programme- 18.6.1997

09.30 a.m.	-	10.30 a.m.	Madhav Gadgil	An overview of the Proposal
10.30 a.m.	-	11.00 a.m.	P. S. Roy	Remote sensing of biodiversity
11.00 a.m.	-	11.15 a.m.	**Tea**	
11.15 a.m.	-	11.35 a.m.	N. V. Joshi	Sampling for maximising the number of species encountered
11.35 a.m.	-	11.45 a.m.	A. P. Gore	Effort needed for Measurement of Biodiversity
11.45 a.m.	-	11.55 a.m.	S. D. Gore	Organizing and managing the database
11.55 a.m.	-	12.10 p.m.	V. J. .Nair	Role of Botanical Surveys
12.10 p.m.	-	12.25 p.m.	P. T. Cherian	Role of Zoological Surveys
12.25 p.m.	-	12.40 p.m.	Madhav Gogte	Role of Forest Departments
12.40 p.m.	-	12.55 p.m.	M. R. Alemeida	Role of Industries
12.55 p.m.	-	1.45 p.m.	**Lunch**	
1.45 p.m.	-	2.00 p.m.	C. A. Viraktamath	Role of Agricultural Universities
2.00 p.m.	-	2.15 p.m.	R. Palat	Involving Colleges
2.15 p.m.	-		**Open Discussion**	

(A) DISTINCTNESS

ANOVA

	$\bar{x}$	σ		pooled	
HABITATS	5%	100%	35 – 70 sp.	13%	18 Jne 97
SITES	7%			13%	

DIST.. around 7%

(B) SAMPLING