(261)

APPLICATION OF INFORMATION THEORY TO EVOLUTION

Dissertation Submitted in Partial Fulfilment of the
Requirements for the Degree of
MASTER OF PHILOSOPHY
by

ZEHRA HAMID

SCHOOL OF ENVIRONMENTAL SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI-110067

1978

CERTIFICATE

The research work presented in this dissertation has been carried out in the School of Environmental Sciences, Jawaharlal Nehru University, New Delhi-110067. The work is original and has not been submitted in part or full for any other degree or diploma of any other University.

Zehra Hamid

ZEHRA HAMID

Subba Rao

Dr. J. SUBBA RAO
Supervisor

B. Bhatia

Dr. B. BHATIA
Dean

Date: August 4, 1978.

SCHOOL OF ENVIRONMENTAL SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI-110067.

ACKNOWLEDGEMENT

I take this opportunity to express my deep gratitude to Dr. J. Subba Rao for suggesting me this problem and guiding me through this research work.

It was a privilege to work under the guidance of Dr. J. Subba Rao and I am grateful for his valuable suggestions. I am also indebted to Dr. (Mrs.) G. Subba Rao for her wise counsel and encouragement.

I extend my sincere thanks to the Dean, Professor N. Bhatia for the facilities extended to me during the course of this work.

I gratefully acknowledge my indebtedness to Miss Anuradha Sinha and Mr. O. Manickam for their invaluable comradeship in and out of the academic sphere.

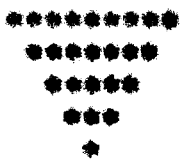
Thanks are also due to other friends of the School of Environmental Sciences for their kind co-operation.

I am very grateful to Mr. Pahwa who has taken enough pain in typing out this manuscript.

Zehra Hamid

August 4, 1978. ZEHRA HAMID

CONTENTS



CHAPTER I

INTRODUCTION

Darwinism is so well established that it is difficult to think of evolution except in terms of selection for desirable characteristics and advantageous genes. New technical developments such as the sequential analysis of proteins and the deciphering of the Genetic Code have made a much closer examination of evolutionary process possible. Patterns of evolutionary change at the phenotypic level do not necessarily apply at the genotypic and molecular level. And so new rules have to be formulated in order to understand the patterns and dynamics of molecular evolution.

By evolution we mean the divergence in the gene pool of population of organisms. This results in differential selection, by differential reproduction of genetic variations in the population concerned and also results in random genetic drift.

Molecular Evolution, on the other hand is the study of the changes which can be identified at the molecular integrative level at different points of the phylogenetic sequences of organisms (bio-chemical evolution).

Processes of Evolution

Evolution thus implies change with respect to time, but a formulation (Hardy-Weinberg - 1908 - App. 1) states that

the population remains constant. This paradox can however be resolved by examining the factors or forces that might upset the genetic equilibrium. The four processes in evolution are mutation, recombination, natural selection and isolation.

Mutation

The first important factor of evolution is mutation. It was pointed out (Charles Darwin - 1859) that any attempt to understand the mechanism of evolution must start with an investigation of the sources of hereditary variation.

Any population of organisms in nature is adapting to the given environmental situation in which it finds itself. A given species of organisms or a given population occupies an ecological niche, which is defined as the sum total of all the environmental inputs on the organism. The tendency for a population to become homozygous can be spoken as "Adaptation". Adaptation serves the well-being of the species only as long as the environment stays constant. The study of the geological record reveals that the environment does not remain constant. But each population has the tendency of becoming 'heterozygous' which is also defined as 'adaptability'. We are thus led to the problem: what is the source of variability, if a population has to maintain heterozygosity? The geneticist's answer is mutation - both genetic and chromosomal. An elaborate study of earlier works (Dobzhansky - 1970) reveals that mutation provides the raw material for evolution. It supplies the

genetic variation to a population that makes possible the new genotypes that may be necessary for the evolutionary success of the population.

Recombination

The argument about the necessity for variability as a compensation for environmental change, if correct, gives rise to another question. How significant mutation will be at the time of environmental change? An answer to this, as to the importance of mutation is that mutation alone is insufficient because spontaneous mutation rates are very low. (10^{-6} per locus) And such a low rate is not likely to be sufficient to provide the necessary variability at the time it is needed by the population. Mutation does not act alone, and the second factor important for evolution - "recombination" must also be considered. Mutations with time builds up a large storehouse of variation. Recombination of this genetic variation provides the many possible genotypes that might mean the difference between survival and extinction for a given population.

Natural Selection

Natural selection can be defined as the difference in gene pools from one generation to the next. Evolutionary changes at the different levels (morphological, functional and behavioural) results from the process of natural selection operating through adaptive changes in DNA. This at the same times does

not mean that all or most evolutionary changes in DNA is necessarily due to the action of Darwinian natural selection. The idea of selectively neutral change at the molecular level has not been readily accepted by many classical evolutionists, perhaps because of the pervasiveness of the Darwinian thoughts. Change in DNA and protein, when it is thought of, is limited to a response to activities at a higher level. (Simpson) But agreement with this statement will mean that DNA is a passive carrier of the evolutionary message. Evolutionary change is not imposed upon DNA from without, it arises from within. Natural selection is the editor rather than the composer of the genetic messages. One thing, the editor does not do is to remove changes which it is unable to perceive.

Isolation

Three of the four evolutionary processes has been described briefly. Before going into the fourth, let us first define a species. The most prominent neo-Darwinism defines it as - "An ancestral species is transformed into two or more derived species when an array of inter breeding Mendelian population becomes segregated into two or more reproductively isolated arrays. Species are, accordingly, systems of populations; the gene exchange between these systems is limited or prevented in nature by a reproductive isolating mechanism or perhaps by a combination of several mechanisms."
(Dobzhansky - 1970)

In short, a species is the most inclusive Mendelian population. Species are dynamic and not static units. From the very same definition, reproduced above, the concept of isolation and isolating mechanisms is clear. We need not go into further details.

After this, let us come back to our own work. At this stage, the problem being dealt in this work could be crudely stated as follows.

How are the species that exist today accounted for? How does the information contained in a particular DNA molecule viz. (a) in a particular kind of protein, and (b) in different species define the problem of evolutionary processes?

Let us now try to analyse these problems one by one.

Comparing Evolution in Protein and in DNA

Species divergence can be measured at the protein level through sequence analysis, and independently at the DNA level through in vitro hybridization. The measurement of DNA species divergence is complicated by the existence of repetitive DNA sequences of unknown function, discovery of these sequences has been the most important finding of the hybridization experiment.

The number of base changes (in m-RNA's derived by decoding the proteins) needed to give the same amino acid

sequence is used to construct phylogenetic trees. But amino acid substitution frequencies cannot be rationalized this way. The minimum base change method poorly reflects homology and also underestimates the total number of fixations inferred from phyletic data.

There is some agreement between codon relatedness and amino acid substitution rate. Protein evolution is more related to properties' difference between substituted and substituting amino acid than to a priori probability for minimising codon changes. Many base changes due to chromosomal and replication errors evidently occur and reoccur before any one is fixed in the population; so that mutations are not strongly selected by the codon transformations they require.

Evolutionary Changes in Certain Homologous Polypeptide Chains

The evolution of homologous polypeptide chains is studied by comparing amino acid sequences in the polypeptide chains of samples obtained from different species of living organisms. The sequences are aligned by placing identical residues in juxtaposition, and the number and types of identical and different residues compared. Amino acids can be compared in terms of the Genetic Code, so that the homology may be expressed between two polypeptide chains either as the number of amino acid differences or the minimum number of base changes in the Genetic Code necessary to convert one chain

into the other. This comparison emphasises evolutionary changes in DNA that are phenotypically expressed as changes in proteins.

The Phylogenetic Tree

Let us reconstruct the phylogenetic tree for a given class of proteins, having in mind that:

(i) it is interesting to compare a phylogenetic tree deduced from the study of functionally homologous molecular species (e.g. Cytochrome C etc.)

(ii) to compare two phylogenetic trees corresponding to two different class of proteins (Cytochrome C and Haemoglobin)

(iii) the reconstructed structures of ancestor sequences which are not amenable to the experiment, may exhibit certain regularities that are no longer apparent in protein of the present living species.

The uniqueness of the topology of a Genetic Tree deduced from a given set of experimentally established protein sequences has not yet been proved.

The second approach to the study of homologous proteins is a structural one. Forgetting the question of protein evolution, one can focus the attention on common factors in the structure of functionally homologous proteins. Attention can

be laid on the structure forgetting the genesis. But this is not the aim of our work. Let us come back to the central aim of our work viz. "The Information Theory".

Heredity and Information

The study of evolution and of heredity deals with the influence of species of past events on the present. There still exists a considerable uncertainty about the application of Information Theory in Biology. At present attention is directed to individual entities as proteins, chromosomes, genes etc. Through the use of Information Theory, we can arrive at some means of exact treatment of the system of organisation of which these units are but the parts.

Communication is the transmission of information from mind to mind (Gabor - 1960). A communication system consists of five parts. (Shannon - 1948)

1. An information source which produces a message or sequence of message to be transmitted to the receiving terminal.

2. A transmitter which operates on the message in some way to produce a signal suitable for the transmission over the channel.

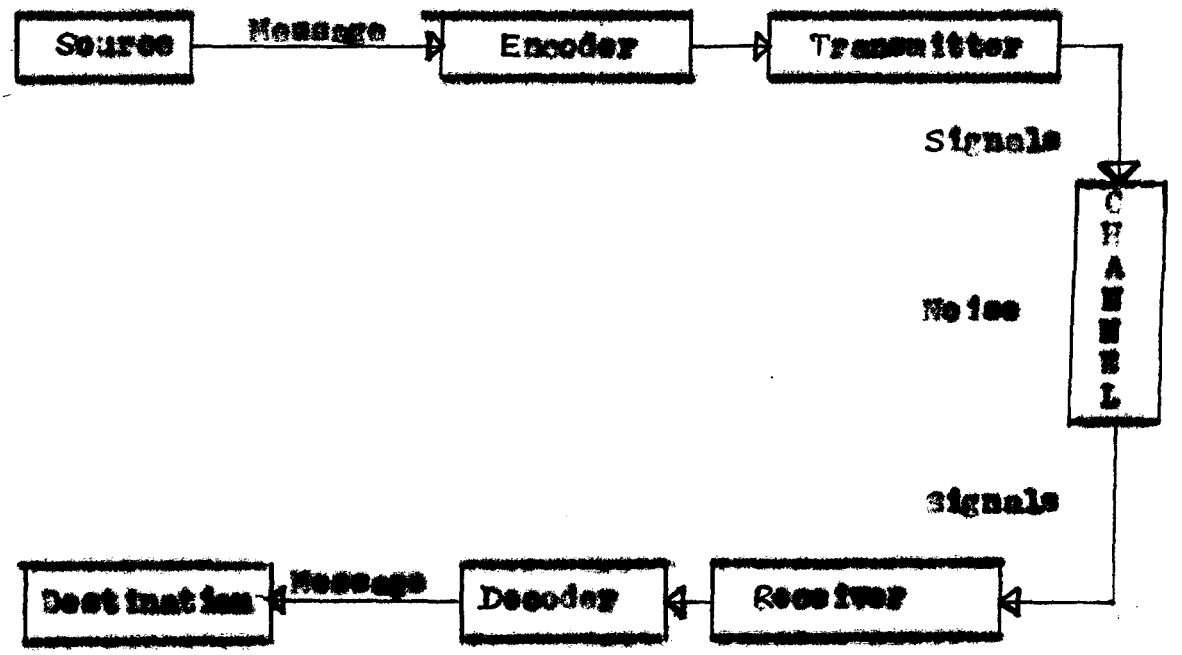
3. The channel used or the medium used to transmit the signal from transmitter to receiver.

4. The receiver ordinarily performs the inverse operation of that done by the transmitter, reconstructing the message from the signal.

5. The destination is the person or thing for whom the message is intended.

The problem for the biologist, is to use the method of communication theory to identify analogues of these five elements in living systems.

Diagrammatic Representation of a Communication System*

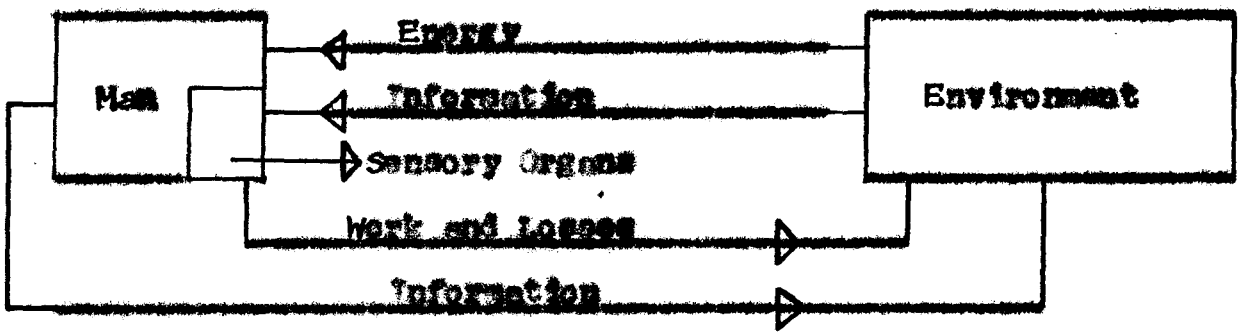


* Reproduced from Wolpert and Apter - 1963.

Man and His Environment

Evolution is a continuing process. Life is a continuum of events. A system is a collection of things or events contained with some specified boundary. Man himself is such a system, or more properly a sub-system operating within a larger system - the environment. To man there are inputs (information or noise and energy). From man there are outputs (information, work or losses - noise etc.)

The Human Being as a Black Box in His Environment



The population of living organisms are complicated open systems that maintain a high level of organized activity. They resist the tendency to dissipate into their surroundings. They do it by virtue of the properties they have of receiving, storing and acting upon large amount of information. The difference between organism and environment becomes greater and greater, as the organisms enter more and more difficult environments (Young - 1968). The organisms remain intact by its

ability to follow changes in the environment, which may be regarded as the source of information.

Information in any system may be defined as that feature of it which remains invariant under re-coding. (Shannon - 1948)

Each individual can be considered as a machine that receives these changes, translates them into a coded version and transmits them through its system to effector agents. These decode the message and produce an appropriate action upon the environment, which is thus the destination. We therefore have to study the flux of events in each set of organisms, to discover how they are related to change in the environment and how they produce changes in the latter.

We may say, that every population of organisms operates at each moment under the influence of three sources of information, controlling the course of future behaviour:

1. There is the information provided by the events in the world around.
2. There is information received from ancestors.
3. There is the information that each organism has acquired and stored during life time.

Information received from the Environment

Any environmental change may produce change in the organism; but the information of heredity ensures that certain

especially reactive points are provided. For example plants have chlorophyll which is activated by light.

The organism thus selects the message to be transmitted. The organisms are influenced by changes in the environment. Their variations produces changes in the effector agents, which in turn act upon the environment.

Information Received by Heredity

Events occur around population of organisms and are received, and coded by their influence upon genetic mechanisms. Selection ensures that the genes of a population include those items that have produced survival under the set of conditions that the population has experienced. In this sense, the genes are a code, which stands for certain features of environment. We may say that the heredity information is decoded in each generation by the process that goes on in development and throughout the life of an individual.

Adaptation as Storage of Information

The physico-chemical reactions between environment and organism leave a more or less permanent impression on the activities of the latter. This constitutes a coded record or memory of the information received. As a result of this adaptability, every organism carries a memory of its immediate past, which serves as it were to forecast the future. For example, if a plant has received relatively

little water or an actual little oxygen, it will be provided with relatively long roots or more haemoglobin.

Molecular Information

The entropy of a DNA molecule (as in thermodynamics) is a function of configuration in which the atoms are ordered. It is expressed in the relation k log D, (k is Boltzmann constant and D is a measure of atomic disorder)

In statistical terms, entropy is defined as

$$S = - K \sum_{i=1}^{n} \log p_i$$

such that $\sum_{i}^{n} p_i = 1$

where p_i is the relative probability of the ith symbol generated by a source.

The Information Content I corresponds to a number

$$I \text{ (bits)} = \log_2 n^n$$

when dealt with several (n)

symbols with several (m) symbol types,

$$I \text{ (bit)} = \log_2 m^n = n \log_2 m =$$

$$3.32 \; n \log_{10} m^m \text{ (bits)}$$

An amino acid may be considered as having information content as that of a word and a protein as that of a prose paragraph.

Information and Entropy

The broad use of the term entropy is as a quantitative measure of the amount of disorder in a system or subsystem. Reproducing the view of Schrodinger (1943):

"Life seems to be ordered and lawful behaviour of matter, not based exclusively on its tendency to go over from order to disorder, but based partly on existing order that is kept up. The whole evolutionary process, both cosmic and organic is one. In the overall balance, cosmic and biological evolution result in an increase of entropy as predicted by the second law of thermodynamics."

In the beginning thus, it was not clear as to how to relate negentropy, which is the difference between the observed and the maximum uncertainty at a given GC composition, to the evolution of organisms.

Example - Let us consider a regular DNA base sequence -

```
.... T - C A G T C A G T C A G ....
       |   | | | | | | | | | |
.... A   G T C A G T C A G T C ....
```

This sequence can at the most code four kinds of amino acids. The entropy of the amino acid sequence coded

by this DNA is thus very low. We may conclude that this DNA base sequence belongs to much higher evolutionary rank than any existing organism (Smith - 1969).

Today the informational uncertainty or informational entropy H is regarded as a measure of evolution.

The Transfer of Information (The Question of Coding)

The original Watson-Crick scheme of replication is a semi-conservative one. Our major discussion has dealt with the process of synthesis of protein as directed by the information encoded in the DNA structure. One speaks of the Information Content of the DNA or the code for genetic information. This means that there is a specific set of directions available in the structure of the DNA molecules which specifies the precise primary structure of every protein synthesised by the particular cell involved.

The Genetic Code

A detailed account will follow in another chapter. The generalisation will be made that primary gene action consists of the transcription of the coded genetic information in DNA into the ribo polynucleotide sequence of mRNA and the subsequent translation of this information. This is contained in the four letter alphabet, A, U, G, and C of mRNA, into specific amino acid sequences.

The Genetic Code is degenerate, but the degeneracy is not evenly distributed. The natural Genetic Code with the base triplet mechanism approaches remarkably close to the optimum coding. This observation seems of significance when considering the evolution of the code by natural selection.

Bio-Chemical Homology

A degree of isology in primary structure greater than possible by probability is the evidence for common origin in the case of nucleic acids and proteins.

A protein chemist sees the replacement as being related solely to function. The external regions of the protein molecule are less restricted to change than are the internal regions, which must be occupied by hydrophobic side chains. Certain residues are invariant because they are essential to enzymatic function. It is the necessary properties of the protein that dictates its primary structure. This view tends to push DNA as the driving force in evolution into the background.

The second view is that the protein molecule is continually challenged by mutational changes resulting from base substitutions and other mutational events in DNA. Natural selection screens these changes. It expresses the random nature of point mutations primarily and only secondary of protein function.

Environmental Specialization

Molecules that occur in living matter has been classified into three categories, according to the degree to which the specific information contained in an organism is reflected in them (Zuckerandl and Pauling - 1966).

(a) Semantopheretic molecules or Semantides

The molecules carry the information of the genes or a transcript thereof. The genes themselves are the primary semantides (linear sense carrying units). mRNA are secondary ones and polypeptide molecules are tertiary semantides.

(b) Episemantic Molecules

These molecules are synthesized under the control of tertiary semantides. All molecules which are built by enzyme in the absence of a template are in this class. Though they do not express the information content of the semantides, they are a product of this information.

(c) Asemantic Molecules

These molecules are not produced by the organism and so do not express, either directly or indirectly, any of the information that this organism contains.

The evolution of the cells in the biosphere has involved a continuous simultaneous gain and loss of properties. In consequence there has been a great shuffling and

assortment to yield the enormous variety of cells in existence. Thus as the environment changed, the acquisition of new properties by certain organisms was essential for the survival of all other living forms. This itself must have initiated many inter-dependent relations.

With the existence of enclosed regions differing in chemical composition from their environment, new varieties of chemical reactions become possible. Some of the primitive aggregates promote the synthesis of their own constituents from other materials available in the surroundings. When these objects grow to a certain size, they become unstable and divided. Certain sorts of molecules and aggregates are able to survive the vicissitudes of the changing environment and grow at the expense of the others.

A part of the biological diversity stems from inheritance, part from interaction of the organism with its environment. In order to understand the mechanisms that provide a cell with its features or phenotype, we must question the nature of cell heredity and gene action.

Fortunately there is order in Biology, just as there is in Chemistry, Physics and Mathematics. Information Theory can be used to put the postulates of Molecular Biology in mathematical form. Information Theory contemplates the changes in accordance with the Central Dogma, the Sequence Hypothesis and the Genetic Code.

The Information Theory being discussed is based on fundamental mathematical theorems and on the first principles of Molecular Biology. It can serve to be an instrument of discovery and evaluation. The Central dogma and the sequence hypothesis cannot escape the consequences of error in protein synthesis.

We have used these very same concepts in a very natural way to calculate the Information Content of the biochemical function of different proteins. This knowledge may contribute in a better understanding of the origin of life and of evolution.

The taxonomist classifies various organisms into groups whose members have overlapping features. The cytologist labels classes of cell structures. The chemist groups molecules of similar structure and so on. Knowledge thus allows the rational assortment of information into orderly arrays. These continue to expose the secrets of nature. The chemical story of the cell is the synthesis of small molecules and their union to form specific classes of macro molecules.

****

References

Anfinsen, C.B. in The Molecular Basis of Evolution, Wiley, New York (1960).

Anfinsen, C.B., in Informational Macromolecules, H.J.Vogel, V. Bryson and J.O. Lampen (eds.), Academic Press, New York (1963).

Apter, M.J., and L. Wolpert, J. Theor. Biol., 8,244 (1965)

Dobzhansky, Th. in The Genetics of Evolutionary Process, Columbia University Press, New York (1970).

Ingraham, V. in The Bio-Synthesis of Macromolecules, W.A. Benjamin Inc., New York (1966).

Jukes, T.H. in Molecules and Evolution, Columbia University Press, New York (1966).

King, J.L. and T.H. Jukes, Science, 164,788 (1969).

Shannon, C. and W. Weaver in The Mathematical Theory of Communication, University of Illinois Press, Urbana, Ill. (1949).

Smith, T.F., Math. Bio Sciences, 4, 179 (1969).

Watson, J.D. in Molecular Biology of the Gene, W.A. Benjamin Inc., New York (1977).

Zuckerandl, E. and L. Pauling in Evolving Genes and Proteins, Academic Press, New York (1965).

The Molecular Basis of Life, Readings from Sc. Amer., W.H. Freeman and Co. San Francisco (1968).
The articles by Crick, Holley, Nirenberg, Yanofsky and Clark and Mark.

CHAPTER II

THE GENETIC CODE

Due to the stand point adopted, it is convenient (Woese - 1968) to depart from the usual presentation of the subject. A coding problem is specified by viewing the protein of an organism (effectively a cell) as a "topologically linear text" in a twenty letter alphabet - that is "the standard set of amino acids". The text is encoded in the sequence of nucleotides in the topologically linear hereditary molecules of DNA (information, genetic etc.). From this "Sequence Hypothesis" (Crick - 1970) follows the 'Coding Problem': How is the one text (the DNA nucleotide sequence) translated into the other? (the structure of the protein)

In view of the assumption that the protein is a topologically linear text, the structure referred to is its primary structure. This assumption is justified since the tertiary structure of proteins arises as the most favourable from an energetic stand point - 3 D configuration and requires no supplementary informational input. Thus proteins on being disordered (denatured) by an environmental perturbation (e.g. heating) will, on restoration of the original conditions spontaneously return to their original configuration. This return may involve the formation of covalent bonds as well as 'weak' bonds.

Equivalently, one can speak of the grammar which generates the protein text from the nucleotide alphabet. From a more speculative view point but justified by the scanty knowledge of the molecular basis of evolution and taxonomic classification, the set of all protein text is considered; structured into subsets by a series of integrations at the metabolic, physiological, organismic, taxonomic and evolutionary level. Then only the structure and nature of the code in relation to these subsets and groups of subsets can be questioned.

The "Sequence Hypothesis" is supplemented by the "Central Dogma" (Crick - 1970) namely that the protein text is stable. This implies the usual formulations. DNA is self replicating with very low transcription error frequency. Further the transfer of information from DNA to protein is one way. This means that a perturbation in the protein text cannot be transmitted to succeeding generations through a protein to DNA information transfer. This accords with the ideas of phenotype continuity and of evolution as a stochastic selection process.

The Nucleic Acids

There are of two main types -

(a) Deoxyribo nucleic acid (DNA)

This is the genetic material of all organisms except certain (RNA) viruses. It consists of two chains wound

around each other. Each chain consists of deoxy ribose (five carbon) sugars linked by 5'-3' phosphate diester bonds. Each sugar unit has a base (either a purine-~~Cytosine~~ Cytosine (C), or Thymine (T), or a pyrimidine-Adenine (A) or Guanine (G)) attached to it at the 1' carbon. Bases opposite to each other on the two chains are bound together with two (A=T) or three (G≡C) hydrogen bonds. This pairing is accurately followed to an accuracy of approximately 1 in 10^7 duplications. The self replicating mechanism available due to this selective pairing rule is obvious and involves the activity of only one enzyme. For example, DNA polymerase, needed to form the covalent bonds of the back bone between adjacent nucleotides (base-sugar-phosphate units) on the newly forming strand.

(b) Ribo nucleic acid (RNA)

The differences of RNA from DNA are -

(i) Thymine (T) is replaced by Uracil (U) which differs from T by the loss of a methyl group.

(ii) The backbone sugar is ribose which differs from deoxy-ribose by the addition of a hydroxyl atom (OH) in place of a hydrogen atom (H) (OH→H) at the 2' carbon.

(iii) The normal configuration is single stranded and has a short life time in vivo due to the presence of degradative enzymes.

Proteins

The problem is the encoding of polypeptide chain sequences since many proteins are assembled from two or more separately synthesized polypeptide chains. A polypeptide chain is a sequence of amino acids $NH_2 - \overset{H}{\underset{R}{C}} - C\overset{O}{\underset{OH}{}}$ where R is a side group different for each acid. It may be an alkyl, phenyl, cyclic group or a combination of these. Thus the letters of the protein text are amino acids.

These amino acids are divided into a standard set of twenty amino acids and a super-numerary set of the other amino acids found in natural polypeptides. The criteria for this division were:

(a) The amino acids of the standard set occur in all polypeptides except for statistical fluctuations, and

(b) The presence of amino acids of the supernumerary set in proteins is due to post synthetic enzymatic modifications of the standard amino acids which are the only ones that are incorporated during synthesis itself. Thus Proline and Lysine are much better precursors for collagen synthesis than their supernumerary set counterparts i.e. Hydroxyl Proline and Hydroxyl Lysine, which are the actual occurring components in collagen.

This division into standard set and supernumerary set is evidently important, if the code is to have a simple and universal structure. This also provides a clearly defined alphabet for the text to be encoded.

Code Reading in Cells

As the purpose of this chapter is to present the code and language of the gene, so the general mechanism of translating sequence information in one polymer into the sequence of another is being described very briefly.

(a) Replication

This is the process of DNA duplication. The strands separate and a new strand is assembled on each parent strand by complementary base pairing.

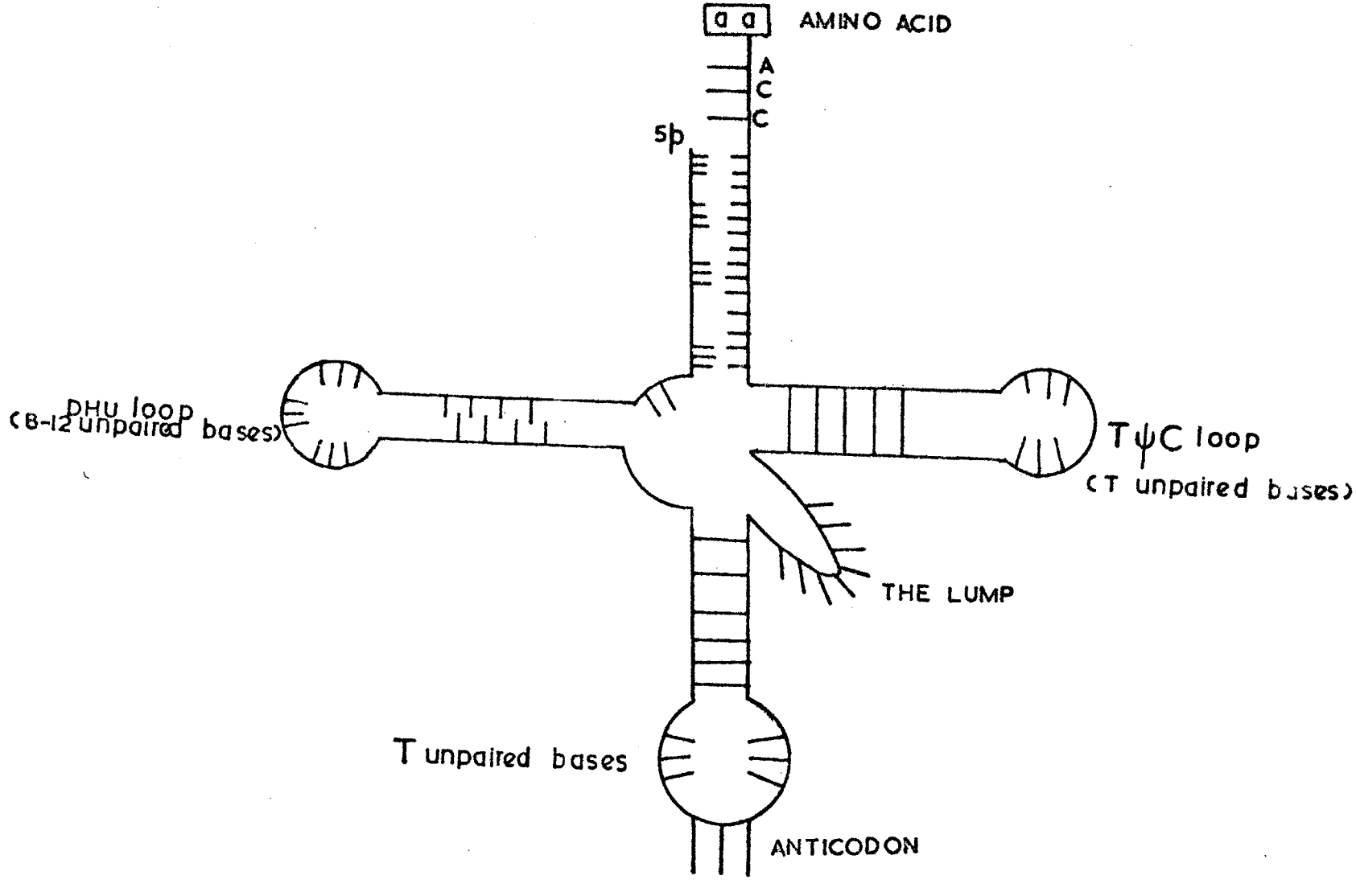
(b) Transcription

This is the production of mRNA by the formation of a polyribo nucleotide (with T→U) on one of the DNA chains. Here 'm' stands for messenger, since it is this molecule which carries information from the genetic memory (DNA) to the translation site (ribosome). The polypeptide is sequentially constructed at the ribosome. The mechanism is again complementary base pairing (A-U, T-A, and G-C). The enzyme used is RNA polymerase.

(c) Translation

This is the process whereby the information in the mRNA base sequence is translated into the amino acid sequence of the polypeptide chain. As the mRNA bases are in general

Fig 1. Schematic diagram of an amino-acyl-tRNA showing clover leaf shape.



incapable of forming complexes with amino acids, so an intermediate adaptor molecule is required. This is another kind of RNA, the transfer RNA (t-RNA). Each t-RNA molecule has a sequence of three unpaired bases (anticodon) which due to the stereochemical arrangement of the molecule are uniquely adapted out of all the bases in the t-RNA to fit to a specific sequence of nucleotides (codon) on the mRNA by the usual base pairing mechanism. On the other hand there exists enzymes (amino-acyl synthesase) which are capable of recognising specific amino acids and t-RNA's and binding the two together. The DHU loop (Fig. 1) is thought to be involved in this recognition.

Amino Acid (AA) + Adenosine triphosphate (ATP)

$$\xrightarrow[\text{Amino Acyl}]{}\quad \underset{\text{Synthesase (AAS)}}{\longleftarrow} \quad AA \sim AMP + P \sim P$$

where AMP is the adenylic acid group

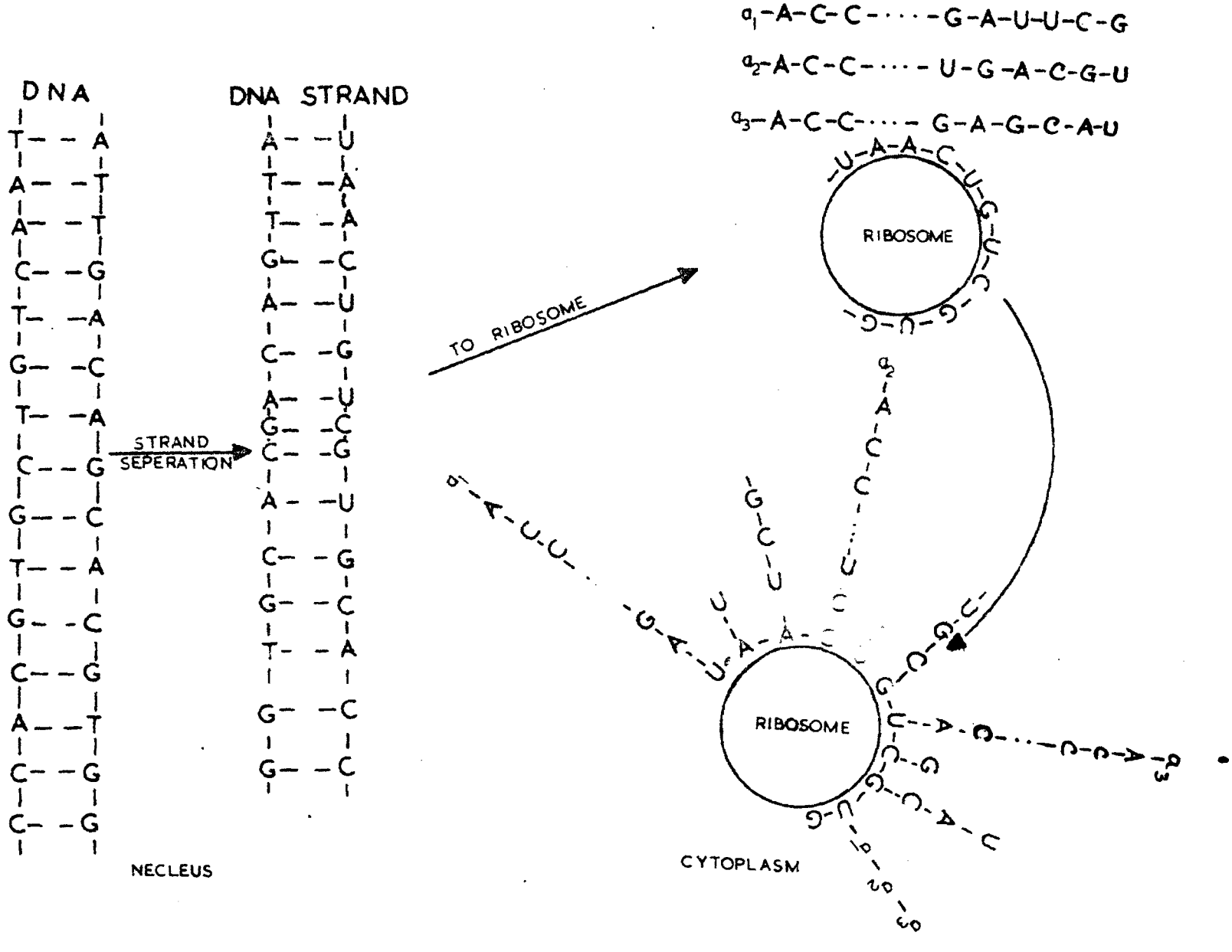
$$AA \sim AMP + t\ RNA \underset{}{\overset{AAS}{\rightleftharpoons}} AA \sim t\ RNA + AMP.$$

The generalised scheme for the interaction of these various molecular units is shown diagrammatically in Fig. 2.

Solution of the Problem - Cracking the Code

An account of the methods used to arrive at an understanding of the reading mechanism would be extremely lengthy. The details of the process are also not central to basic

Fig 2. Formation of RNA.



information transfer processes which basically involve only base pairing (DNA - mRNA, mRNA - t RNA and amino acid - t RNA association), it is being neglected. Instead, the attention is being confined to the methods used to arrive at the results supporting the characteristics of the code as given above.

Basically the methods of code cracking fall into two categories: in vivo and in vitro studies. The former provides information on the nature of the code while the latter is more useful in investigating the codon-amino acid assignments. The two approaches thus complement each other and most of the code's properties are deducible from data obtained by either. The few contradictions remaining are being steadily reduced as the understanding of the parameters controlling the functioning of cell free (in vitro) studies increases.

<u>In Vivo Studies</u>

These rely on two techniques.

(i) Genetic mapping using the property of "crossing over" (that is the breakage and intercombination of similar chromosomes). This leads to gene assortment. As the breaking frequency between two genetic loci obviously depends on the distance between them - this allows the construction of a genetic map of loci controlling various distinct functions.

(ii) In tandem, studies of protein structure modification by mutations at various genetic loci.

One or two clear cut experiments will be recounted in each case. Such techniques permit the answering of the questions of the colinearity, overlapping, coding ratio, inter-codon punctuation and degeneracy of the code. Each experiment is well confirmed by other data.

Colinearity

Yanofsky et al (1967) determined the locations of amino acid replacements in a seventy five amino acid segments of A protein of E. coli Tryptophan synthesase, due to mutations in the corresponding gene whose locations were determined by usual genetic mapping techniques. In every case the mutations were arranged in the same sequence on the genetic map as the amino acid replacements in the protein sequence. The distances on the two maps do not correspond due to differential probabilities of cross over at various loci. Later the sequence of all 267 amino acids in the protein was determined and amino acid replacements again correlated sequentially with the corresponding mutations (Fig. 3).

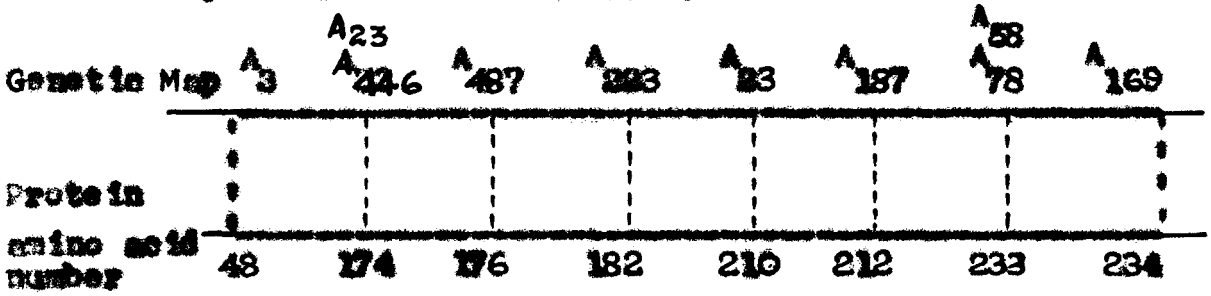
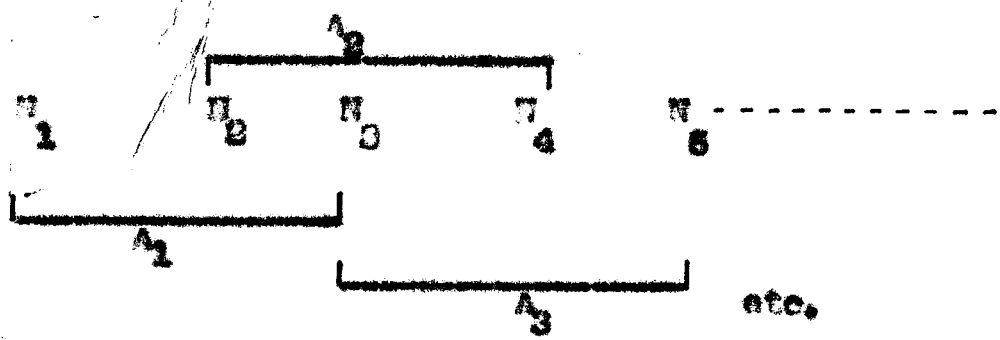


Fig. 3: Yanofsky (1967), reproduced from Wees (1969)

A number of similar experiments support the same conclusion.

Overlapping

An overlapping code restricts the possible amino acid sequences in proteins. The restrictions would depend on the degree of overlap. For example in a completely overlapping triplet code



A_1 can be followed only by an amino acid with N_2 N_3 as the first two letters of its codon.

The detailed statistical analysis have revealed no non-random behaviour among the observed succession frequencies. More so, of the four hundred possible dipeptides all but twenty three have been found to occur naturally (which involve the rarer amino acids e.g. Tryptophan or Methionine).

Further, mutations usually produce only one amino acid replacement rather than the several which would be expected from the altering of even one mutated base. Amino acid replacements in Tobacco Mosaic Virus (TMV) were observed by Wittman. Forty one were induced using Nitrous acid as a mutagen, eleven were

spontaneous. Almost all led to single mutations and none of the double mutations were adjacent. The Tryptophan synthesase mutation experiments all give only a single replacement.

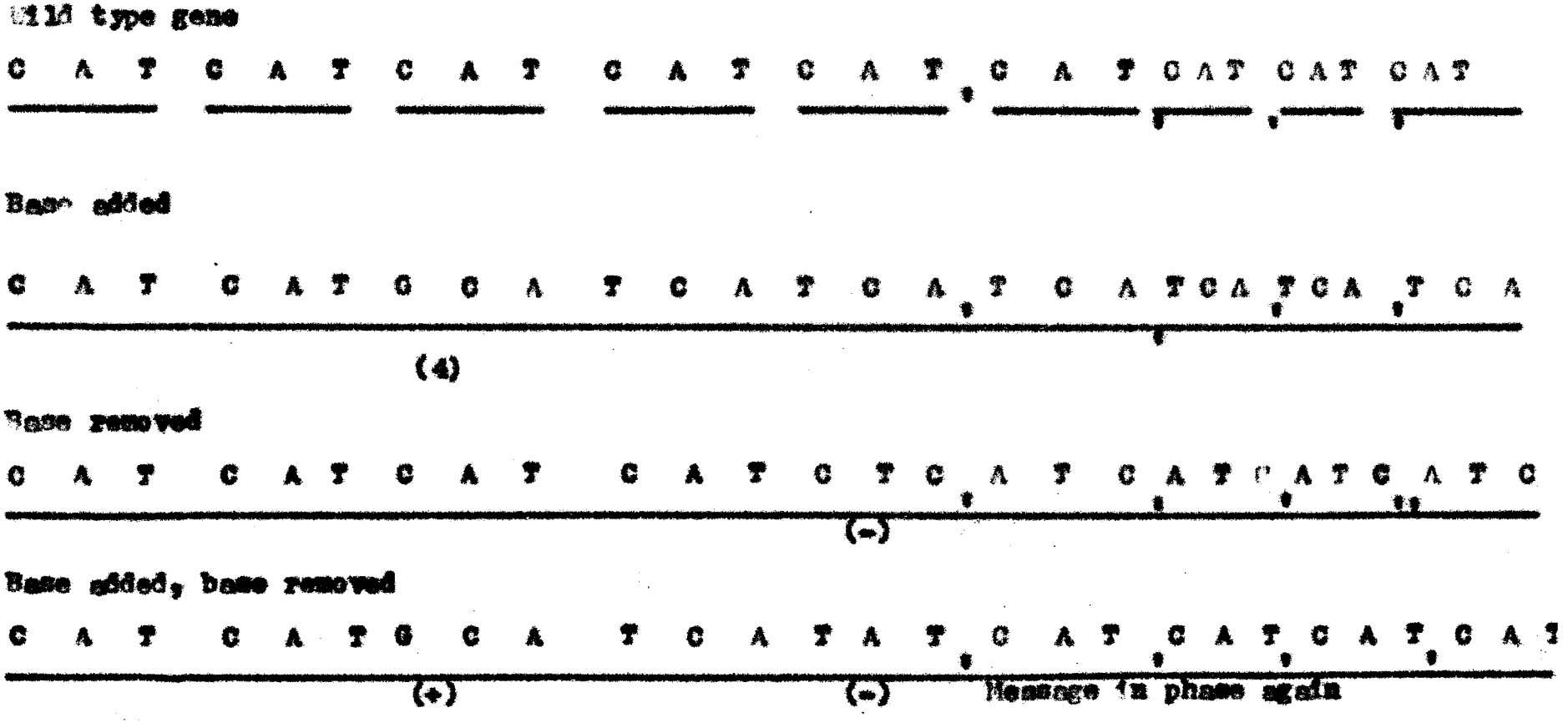
Coding Ratio : (Coding length for non-Overlapping Codes)

Apriori arguments favour a minimum codon length of 3. A four letter code will generate 4, 16, 64, ... for singlet, doublet, triplet etc. sequence codes. Since 20 amino acids must be coded, so the minimum codon length is 3.

Various attempts to determine the coding ratio by counting the number of amino acid replacements induced by mutagens, indirect comparisons of mRNA length and protein chain length etc. have been made but yield only approximate and bounding results.

The best direct method determination is that of Reichmann (1964) using the Satellite Tobacco Necrosis Virus. Satellite viruses can multiply only in the presence of other viruses. The implication being, that since their observed RNA content is low - they have retained only their coat protein cistron (one cistron codes for one polypeptide chain) and require the cistrons of other viruses to accomplish other functions. Reichmann found that the coat protein contained 400 amino acids while the RNA length was of the order of 1200 nucleotides. Clark (1966) confirmed that this RNA coded principally for coat protein in vitro studies. This gives

Figure 4



Note: Effect of mutations that add or remove a base is to shift the reading of the genetic message, assuming that the reading begins at the left-hand end of the gene. The hypothetical message in the wild type gene is CAT, CAT.... Adding a base shifts the reading to TCA, TCA .... Removing a base makes it ATC, ATC .... Addition and removal of a base puts the message in phase again.

a coding ratio of exactly three. Crick et al (1961) gave an ingenious demonstration of several of the currently accepted features. A brief account is being given. They assumed a triplet or multi-triplet codon, a non-overlapping degenerate code read from a fixed starting point with no "commas". This implies that a shift in starting point would change the base sequence. Such a shift would be effectively produced by the insertion (+) or deletion (-) of a code (Fig. 4).

The reading frame would be restored by the insertion (+) of any base in a position succeeding the first deletion restoring the original codon sequence read except between the two mutations.

Further it is clear that if there are 3 n bases (n = 1, 2, ...) in a codon, then if 3 n mutations of the same sign were induced then the sequence would remain unaltered except between the first and last mutations while otherwise the whole sequence would remain scrambled. All these conclusions depend strongly on the assumptions made.

They then performed experiments using Acridine yellow and Proflavine generally accepted as deletion-insertion mutagens to study mutations in the B cistron of the rII region of the E. coli Bacteria phage T_4. Combinations of deletion-insertion (d-i) spontaneous mutations were studied by genetic mapping techniques. Combinations of 3 + 's, 3 - 's, 6 + 's etc. were shown to reproduce the original phenotype (as distinguished by shape etc.). The other combinations also

gave results consistent with this assumption. These results are not being reproduced here, but the conclusion of a triplet, commaless code with start-determination of reading frame may be taken to be established. This is supported by all other evidence. They gathered evidence supporting the view that there were two types of frame shift mutation and that pairs +-, -+ of these resulted in the original phenotype (with certain restrictions) and that similar pairs (++, --) never resulted in a revertant phenotype.

Degeneracy

The (G) + (C) content of DNA from various organisms varies as a fraction of total bases - from 0.25 to 0.75. The average composition of proteins however varies very little between widely different organisms. This already suggests a high degree of degeneracy. Upon consideration of various other possible causes such as letter codes with five letter codons, variation in transcription rules, non-universality of the code, non-genetic DNA basing etc. may be eliminated leaving degeneracy as the most plausible alternative. This is supported by other data and thus implies that degeneracy is to a considerable extent involving different codon letter composition rather than just sequence permutation.

Amino Acid Replacement

Most replacements are of single amino acids and should therefore result from the replacement of one base (as a deletion

or insertion would scramble the reading frame and since probability of adjacent base mutations is much lower than that of single base replacements). Firstly observed replacements are seen to be much less frequent than single base mutations indicating degeneracy. Statistical estimates of degeneracy have been made by classifying the replacements into permitted (achievable by single base substitution) and forbidden replacements and comparing the known replacement frequency list with the theoretical one generated by assuming a permitted transition number. He obtained 75 permitted translations which compares very well with the 76 predicted by the dictionary.

Thus in vivo, methods permit the determination of all the general features of the code. It remains to determine the actual codon assignments. This is best and most easily done by in vitro methods and then checking with in vivo systems.

Codon Assignment by 'in Vitro' Methods

These rest on experiments equivalent to the direct in vivo method: to determine the base sequence of a cistron and the corresponding amino acid sequence in a polypeptide chain. The in vivo method is quite unfeasible due to technical difficulties. However, the discovery (Nirenberg and Matthaei - 1961) that the addition of m-RNA (synthetic or natural) stimulated the incorporation of amino acid precursors (the incorporation is detectable only by using

traces of labelled amino acids as in vitro systems are still very inefficient compared with the corresponding 'in vivo ones). This immediately opened up the possibility of using synthetic RNA of known base sequence to produce polypeptide and this is evidently equivalent to the direct method and technically simple. Two technical achievements were the pre-conditions of this achievement.

(a) The Synthesis of RNA of A Known Base Sequence

Random ribo nucleotide polymers had been readily available since Grunberg, Manago and Ochoa's discovery in 1955 of the enzyme-polynucleotide phosphorylase. This enzyme can catalyse the formation of polymers of nucleosides from nucleoside triphosphates without needing a template. The base composition of the polymer being dependent on the concentration of precursors available and the sequence being random (within fluctuations). Further developments in poly-nucleotide synthesis and ordering will be described with their own labor.

(b) The Development of Stable Cell Free Systems

Previous investigators while observing the incor-poration of labelled amino acids in a substance precipitated by trichlore acetic acid (a property of proteins) had been unable to pursue investigations due to system instability arising from messenger degradation and enzyme inactivation.

These difficulties were overcome by Nirenberg et al. (1961). They prepared a workable E. coli cell free system as follows:

(i) Disruption of cells by grinding with alumina.

(ii) Extraction with buffer.

(iii) Centrifugation to remove debris and intact cells.

(iv) Ultracentrifugation to separate ribosome and supernatant (t-RNA and enzymes).

(v) Dialysis against buffer using members of various cell sizes to separate out t-RNAs and amino-acyl enzymes. Mercapto ethanol added to buffer stabilized components and made low temperature storage possible.

(vi) A mixture of washed ribosomes (this separates out the attached m-RNA etc.) and t-RNA was backed by an ATP generation system and 20 labelled amino acids.

This system was able to incorporate labelled precursors into polypeptide chains. Addition of DNA degrading enzyme deoxyribonuclease stopped incorporation implying that m-RNA was being constantly degraded and needed renewal by synthesis. Finally, addition of external m-RNA from various sources, natural and synthetic, produced a several hundred fold incorporation rate increase. It was subsequently confirmed that no anomalous process was taking place by demonstrating that virus coat protein chains were synthesised upon addition

of viral m-RNA fractions to the E. coli cell free system (also supporting the hypothesis of code universality).

The methods of decoding, using synthetic polymers, di, tri and oligo (few) nucleotides will now be discussed briefly. These fall into two sets of two classes according to whether the t RNA- ribosome binding or polypeptide chain composition is studied and whether a random or defined sequence nucleotides are used. A discussion of materials and methods (Electrophoresis, Paper Chromatography, UV spectroscopy, Base preparation etc.) will also be omitted.

Random Polynucleotide Studies

(1) The original impetus in this whole field came as mentioned from Nirenberg et al's work (1961). After testing that the system worked, they added polyuridylic acid (UUU ....) to the system.

The trichloroacetic acid (t.c.a.) insoluble fraction contained 800 times as much labelled Phenylamine as in the absence of poly U. Further the t.c.a. satisfied the usual tests which would be expected for polyphenyl-amine. Other amino acids did not respond significantly.

By using a method of washing ribosomes on Millipore paper to separate those bound with t-RNA from the ribosomes, Matthaei et al (1962) were able to show that Phenylamine-t RNA-ribosome complexes are formed as intermediates in the synthesis of polyphenyl amine in the presence of poly U.

Similarly they also found that polycytidylic acid promoted poly-proline incorporations. In view of the in vivo results which were generally known, the assignment of phe ⟷ UUU and pro ⟷ CCC was natural.

(11) Next mixed random copolymers were used to study amino acid incorporation by both t-RNA binding and polypeptide amino acid content methods. These polymers have two or more kinds of bases.

The experiment consists of adding 19 'cold' and one 'hot' (labelled) amino acid to a cell free system along with the copolymer (this ensures the availability of precursors for all possible polypeptides). The difference in blank and + copolymer incorporation or binding rates then tells us whether the copolymer is directing incorporation of the amino acids or not. Thus Nirenberg et al (1962) obtained data which is being reproduced below:

Table 1

Addition of copolymer (mu moles of base residue)	C^{14}- Amino Acyl C^{14} Val t-RNA	t-RNA bound (mu moles) C^{14}- Phenyl tRNA	C^{14} leu-t RNA
No addition	0.38	0.22	0.37
4.7 poly U	0.23	4.73	0.22
4.7 poly UC	2.65	1.93	0.24

The chief difficulties arise due to hydrogen binding either between different chains or between bases on the same chain. For example, poly U and poly A would stick to each other. Poly (AU) (AUAU ....) is inactive since it doubles with itself. Poly G and G rich polymers are also inactive due to self homing.

Efficiency also depends on chain lengths (long 10^2 nucleotide copolymers direct synthesis much better than short 0 - 11 nucleotide ones) and decreases rapidly with time due to polymer degradation by enzymes.

However, once it is seen that incorporation occurs, restrictions on the possible codons for the amino acid may immediately be imposed. These are of two types and are best illustrated by examples -

(i) Poly U and Poly C do not direct incorporation of amino acid X but poly (UC) does. Therefore X must be coded for by a mixture of U and C bases (qualitative approach).

(ii) Quantitative: In a random copolymer,

Frequency of triplet XYZ = Product of apriori frequency of individual nucleotides.

Thus, if Poly (U) 3:1 is used, then evidently

Frequency UUU = $(3/4)^3$ = 27/64

Frequency UUG = $(3/4)^2$ 1/4 = 9/64

Frequency GGU = $(1/4)^2$ (3/4) = 3/64, and

Frequency GGG = $(1/4)^3$ = 1/64

Thus, since we know UUU to code for Phenylamine we expect any amino acid coded by 2 UUG, 2GUU and 3G to incorporate in a ratio 1/3, 1/9, 1/27 to the incorporation of Phenyl-ame. Here 1U2G = UGG, GGU, GUG, and no information on sequences is given. This is however obtainable from studies using polymers with a known sequence.

Some preliminary results on sequences were obtained by a combination of random polymer and amino acid replacement data but these were incomplete and had to await confirmation by 'known sequences' methods since a firm assignment of a codon to at least one or two well located (easily replicable by other amino acids) amino acids was required. This method (Rychlek and Sorm, 1966) is best illustrated as given below:

Suppose we know that Val ↔ GUA and GGA ↔ Gly and CAA ↔ Glu. Then if val → Gly and Val → Glu are observed, then Glu ↔ GAA and Gly ↔ GGA immediately follow by the usual single base replacement argument (U → A, U → G respectively).

Defined Polymer Studies

(a) t-RNA Ribosome Binding

The peptide bond formation (peptidation), during polypeptide synthesis can be blocked by omitting peptidyl transferase, GTP-ATP generation systems etc. The t-RNA binding alone may then be studied.

Nirenberg and Leder (1964[8]) developed a method in which m-RNA of just 3 codons were used to direct t-RNA binding to ribosomes. They digested poly UG with pork liver nuclease giving UG, GU, GUU. UGU and UUG mixtures. Paper chromatography and paper electrophoresis were used in tandem to separate and purify these fractions. Their purity was assayed by digesting with pancreatic RNAase (1) and T_1 RNAase (2) in alteration.

For example -

$$(XYZ) \xrightarrow{(1)} XY + Z \quad \text{where } Z = U$$

$$(XY) \xrightarrow{(2)} X + Y \quad \text{where } X = G;\ Y = U$$

$$(XYZ) \xrightarrow{(2)} X + YZ \quad \text{where } X = Cp$$

$$(YZ) \xrightarrow{(1)} Y + Z \quad \text{where } Y = U,\ Z = U$$

$$\Longrightarrow XYZ = GUU.$$

The prepared tri nucleotides were more than 98% pure. On conducting amino acyl t RNA-ribosome binding assays with radioactive precursors, they obtained the data given below:

Table 2

Addition of copolymer (mU mols of base residue)	C^{14} Amino Acyl C^{14} Val t-RNA	t-RNA bound (UUU moles) C^{14} Phenyl t-RNA	C^{14} Leu t-RNA
No addition	0.4	0.22	0.62
4.7 GUU	1.11	0.27	0.66
4.7 UGG	0.4	0.25	0.66
4.7 UUG	0.37	0.26	0.44

This indicates clearly that GUU is a codon for Valine, while UUG and UGG do not code for Phenylalanine. This method while very direct may lead to errors due to various reasons. For example, UU is not a Leucine codon (column 4, row 4). Finally dinucleotides gave no difference at all indicating a triplet code.

(b) Polypeptide Synthesis using Defined Sequence Copolymers

This is the most reliable in vitro method of decoding, however it is not always applicable. The major achievement of the technique on which this method is based on is the construction of polypeptides with a repeat sequence such as

ABABAB ... or ABCABCABC ... or ABCDABCDABCD .....

These are referred to as poly-r-AB, poly-r-AAG and poly-r-ABBB respectively. The method of preparation consists of the preparation of di or tri nucleotides, by chemical methods, of either the ribo or deoxyribo form. These are then polymerised by using a DNA polymerase on the oligorthonucleotide or by using a DNA polymerase to form a DNA strand and then copying this using a DNA polymerase with precursors for only one of the two strands to ensure that only one is copied. The pioneer work in the use of such chains was done by Khorana et al (1965).

(1) The polymer UCUCUCUC .... (poly-r-UC) with chain length of the order of 110 bases was added to an in vitro E. coli system and found to direct the synthesis of polypeptide chain with an alternating sequence of Serine and Leucine. This sequence was established from the -

(a) 1:1 molar ratio of the acids

(b) Both acids were required for synthesis

(c) Acid hydrolysis gave a chemically pure dipeptide ser-leu (tested by chromatography etc.). In another experiment, by Khorana et al (1966), they used poly-r-AAG to stimulate polypeptide synthesis and found that 3 distinct polypeptides, poly-lysine, poly-arginine and poly-glutamate were produced.

The first experiment implies an odd codon size and assuming a triplet size gives UCU and CUC as the codons. Since

UCU triplets promote cor- t RNA binding, CUC ↔ leu, a conclusion that could not be reached by the triplet method.

The second experiment can only be explained using a triplet, non-overlapping, commaless code with initial point chosen randomly and on these assumptions immediately gives AAG, AGA, GAA as the codons for the 3 acids. AAG ↔ lys., GAA ↔ glu gives AGA ↔ arg; a conclusion which could not be reached by the triplet t RNA binding experiments which gave AGA → asp.

Using a combination of the above methods, the table 3 has been has been arrived at. It only remains to discuss punctuation and universality. The code has been extensively confirmed in vivo by experiments on amino acid replacements, intra codon recombination, frame shift mutation etc.

Punctuation

To confine the discussion to a reasonable length and since punctuation is still a matter of considerable discussion, hence, a detailed consideration of experimental results is called for, if any clear cut pattern is to emerge. The results for various investigations are stated briefly only with no pretentions to comprehensiveness.

(1) The codons UAA, UAG, UGA have been shown not to code for any amino acid and are hence called nonsense codons. On the occurrence of such a codon translation halts and the m-RNA is released from ribosome. These codons thus function as

effective stop signals. There is a wide variety of in vivo and in vitro evidence to support this view. This is called C-terminal punctuation and is fairly well understood.

(ii) N-terminal Punctuation

The problem of a start signal however still remains. This is a considerably more complex matter and while by 1970 the mechanism in the lower (prokaryotic) organisms was fairly well understood, but that in eukaryotic organisms was still shrouded in difficulties.

The position as regards prokaryotes is as follows. It was found that in E. coli proteins a considerable number ended (began!) with Methionine. Further it was found that E. coli can formylate the amino group of Methionine when it is attached to one of the two kinds of met-t RNA (f) which is present in smaller amounts than the other (m). Now, if because of amino group blocking no amino acid can precede formyl Methionine in an amino acid sequence, hence it is a 'natural' starter. Some proteins may have their terminal foryl Methionines removed by enzymes adapted for the purpose. This is supported by studies of transcription of viral m-RNA's in E. coli in vitro systems; the terminal amino acids are found to be N-foryl met although the natural protein sequence follows this group. This is explained by the removal enzyme theory. It may be noted that no 'blocked' amino acyl t-RNA's are found in mammalian systems. Further investigations have shown

that the Methionine codon AUG recognises formyl Methionine and that at least two other codons GUA and GUG are also capable of doing this at least in E. coli. The problem of their dual role (they also code for Valine) can at least be formally answered by the hypothesis that they are read as start signals only if the preceding signal is a nonsense or stop codon, otherwise they are read as internal codons. Very briefly this can be stated as: A N-formyl met t-RNA can occupy the site P on the ribosome when it is empty due to the occurrence of a nonsense codon at the previous position and this initiates the synthesis of a new polypeptide. However, the whole matter even in E. coli is considerably more complex and cannot be discussed here.

The Wobble Hypothesis

Third place codon degeneracy, the existence of only 30-40 t-RNAs as opposed to 61 sense codons, the recognition by pure t-RNA's of several codons and the occurrence of unusual bases in t-RNA's have all lent support to Crick's Wobble hypothesis. This states that the base in the 3rd position of the anticodon (5') is not as spatially confined as the other two and this allows it to form a hydrogen bond, with more than one kind of base. Generally speaking purine-pyrimidine base pairing is favoured. The rule which is followed is:

Base in Anti-codon	Base in codon
G	U or C
C	G
A	U
U	A or G
I (Inosine)	A, U or C

Reproduced from Watson (Molecular Biology of the Gene - 1976)

Although this permits one base to recognise several others, the binding efficiencies display a spread. Thus this mechanism may be involved in rate control.

Universality and Evolution of the Code

The first approach, in the absence of direct data, to the study of the evolution of the code must naturally be a comparative one between organisms of various species.

This comparison has been performed in a number of cases and the conclusion that the code is universal is supported in all but one or two details. Three methods were used:

(a) Cell free system incorporation stimulated by known sequence polynucleotides.
(b) Protein synthesis in vivo induced by foreign DNA, and
(c) Amino acid replacement data.

(a) Speyer et al (1962) investigated whether organisms with unusual DNA composition obeyed different codes. A cell free

system from Alcaligenes faecalis (66% G+C) was used to test 22 triplets and found to specify the same amino acids.

Marshall (1967) tested the readings of 50 codons in cell free systems prepared from E. coli, Xenopus Laevis (atoad) and guinea pig tissues. In the large majority of c ses only quantitative differences (binding efficiency) were observed and these were most likely due to the limitations of the triplet method. The major differences found were;

(i) UGA - nonsense in E. coli is Cystine in guinea pig and other vertebrates, and

(ii) vertebrates lack a codon for formyl Methionine.

(b) Genetic loci may be transferred between species and viruses introduced into unusual hosts. In each case investigated the transfer is functional. Since different codon assignments, but a similar code nature would lead to chaos due to the presence of host t-RNA the virus would have to specify t-RNA degrading enzymes and specify its own t-RNA's. This is difficult in view of the small size of most viral chromosomes (DNA or RNA).

(c) The amino acid replacements in such widely diverse organisms such as E. coli, TMV (coat protein) and humans (haemoglobin) indicate strongly that the coding is the same.

Let us now try to analyse evolution in this context. There are two main aspects to consider this;

(a) DNA composition evolution.

(b) The evolution of the code itself.

(a) The main facts to be accounted for is the wide variation in DNA composition between different species and the fact that within a given taxonomic class the variation is quite small. Other effects such as redundancy are also observed. Thus in higher organisms (but not bacteria) some segments are repeated up to 10^5 times.

Thus the variation of (G+C) content among vertebrates is small, the composition being clustered around 42%. Invertebrates show a somewhat larger variation from 34-46%. Higher plants are again bunched at 40%, but with greater scatter than vertebrates. Algae have a high (G+C) content (50-70%) with a few groups of lower content. Fungi group about equally in 3 ranges, 36-38%, 42-44%, 50-52%. Protozoa are largely of low (G+C) content with a few species of high (G+C) content. Bacteria show the most remarkable spread from 25-75%, the number in any given percentile is about equal except that there are pronounced dips at 38-46%, 58-64%. Viruses generally have less than 50% (G+C).

Sueoka (1961) and Freese (1961) have proposed stochastic theories based on the assumption that the base composition changes slowly as a consequence of random single base mutations which integrate to produce a cumulative effect. Thus Sueoka estimates that if the initial G+C mole

Amino Acid Symbols

Amino Acid	Three letter symbol
Alanine	Ala
Arginine	Arg
Asparagine	Asn
Aspartic Acid	Asp
Asn and/or Asp	Asx
Cysteine	Cys
Glutamine	Gln
Glutamic Acid	Glu
Gln and/or Glu	Glx
Glycine	Gly
Histidine	His
Isoleucine	Ile
Leucine	Leu
Lysine	Lys
Methionine	Met
Phenylalanine	Phe
Proline	Pro
Serine	Ser
Threonine	Thr
Tryptophan	Trp
Tyrosine	Tyr
Valine	Val

Note: Twenty amino acids constitute the standard set found in all proteins. A few other amino acids occur infrequently in proteins but it is suspected that they originate as one of the standard set and become chemically modified after they have been incorporated into a polypeptide chain.

TABLE 3

THE GENETIC CODE

First Position 5' end	Second Position U	C	A	G	Third Position 3' end
	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term	Term	A
	Leu	Ser	Term	Trp	G
	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
A	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Note: Given the position of the bases in a codon, it is possible to find the corresponding amino acid. For example the codon 5' AUG 3' on mRNA specifies methionine, whereas CAU specifies histidine. UAA, UAG and UGA are stop signals. AUG is a part of the initiation signal, in addition to coding for internal methionines.

fraction is 0.30 and $AT \underset{v}{\overset{u}{\rightleftharpoons}} GC$ (u, v being mutation rates per generation), then if $v/u \sim p.4$, then about 10^7 generations are required to pass from $p_0 = 0.3$ to $p_n = 0.4$.

For bacteria with a generation time of one hour, this corresponds to 2000 years, while with generation time of one year, it is 20 million years. Thus it follows that bacteria starting from a 'common ancestor' would diverge in DNA composition much more rapidly than say vertebrates. In fact he predicts the composition heterogeneity resulting to be

$$\sigma^2 = (\text{Standard deviation})^2 = \frac{p(1-p)}{\text{No. of base pairs/molecule}}$$

This however, gives too small a value, even for bacteria, but he introduced several broadening effects. Again such stochastic theories are not being dealt with, but the question of DNA heterogeneity in the content of Information Theory will be considered in some details.

Evolution of the Code

There are three salient facts to be noted about the code shown in table 3.

(i) It is effectively universal.

(ii) It exhibits a regular degeneracy (Goldberg and Wittes - 1966). Twice degenerate words (amino acid or term) are all of the form ab pur or ab pyr. Four times degenerate amino acids have formula abx, where x is any base.

Arginine, Leucine and Serine are all 6 degenerate. Leucine has codons CUX and UU purine. Arginine has codons CGA and AG purine. Thus both have the form nb purine. Serine on the other hand has UCX and AG pyrimidine.

(iii) There are definite correlations between amino acid chemical structure and codon pattern.

(iv) There are two exceptions:

(a) Methionine-Isoleucine, where Isoleucine is 3 degenerate while Methionine is not and Tryptophan - Termination (UGA) both of which are non-degenerate. Also UGA is term in prokaryotes but Cystine in vertebrates.

It has been argued by Crick that given an established dictionary no change in it is possible as a change in codon assignment would replace one amino acid by another in all proteins in which the first codon is used to specify the first amino acid. This therefore possesses a high probability of lethality. Quantitative estimates involving quite reasonable estimates on the probability of an amino acid occurring in a critical locus lead to practical certainty of the lethality of dictionary changes. Some controversy is possible especially in relation to the more infrequent acids such as Methionine, Tryptophan, Cysteine and the rare terminal codon UGA. Thus rule (iv) tends to support the conclusion that minor changes involve codon splitting for eg.

AJ^{A}_{G} [Met] ——— AJA Ileu
AJG Met

UG^{A}_{G} [Tryp] ——— UGA Term or Cysteine
UGG Tryptophan

These codon splittings have occurred since rule (i) does not apply to such codons as we would apriori expect. There are also differences in stop start signals. For example, UGA is Cys in vertebrates, AUG does not code for formyl Methionine in vertebrates.

Thus, a little change has taken place in the code since it was established, and it appears to be a finished product. It has been suggested that its establishment was with the evolution of the original metabolic systems from enzymeless to catalysed ones.

Theories of code evolution (mechanistic and stochastic) as they exist today are highly dubious, speculative and only one approach (Goldberg and Wittes - 1966) will be described. This theory seems to be fairly coherent even if purely formal.

The pattern of degeneracy observed as outlined above (i to iv) is such as to minimise the deletrious effects of mutations. That is this involves the 'selection hypothesis', that the form of the code was arrived at by an 'optimization' procedure relative to random base change mutations. Thus Goldberg and Wittes point out that the code is such as to increase the probability that a base pair change will result in

a codon coding for the same amino acid. Thus one may expect pur → pur as more likely than pur → pyr. The twice degenerate amino acids are unaffected by third place replacements pur → pur or pyr → pyr. Similarly the 4 and 6 times degenerate amino acids (of which there are 5 and 3 respectively) are completely unaffected by third place replacements. However the latter, 6 degenerate is invariant under even more replacements.

The third place degeneracy has several advantages as translation errors are most frequent there and given the weak pairing at the third place, there exists a possibility of reducing the number of t-RNA's needed and so on.

Further Goldberg et al (1966) note that the code is such that mutations tend to produce a similar amino acid to the one originally coded for, thus reducing the probability of a lethal change in protein (enzyme) conformation and therefore in its activity. A more interesting hypothesis is that of Woese (1967) namely that translation rather than mutation errors were reduced by the evolutionary optimization of the code so as to reduce the effect of errors produced by the primitive translation mechanism which must have been only functional group (rather than amino acid) and prone to catastrophic error propagation.

••••

References

Crick, F.H.C., L. Barnett, S. Brenner and R.J. Watts-Tobin, Nature, 192, 1227 (1961).

Crick, F.H.C. Nature, 227, 561 (1970).

Freese, E., "The Molecular Mechanisms of Mutations", Proc. 6th Int. Cong. Biochem. Symp. 1, Pergamon Press, London (1961).

Goldberg, A.L., and R.E. Hittes, Science, 153, 420 (1966).

Grunberg-Manago, M., and S. Ochoa, J. Amer. Chem. Soc., 77, 3165 (1955).

Leder, P., and M.W. Nirenberg, Proc. Nath. Acad. Sci., 52, 420 (1964).

_____, Proc. Nath Acad. Sci., 52, 1521 (1964).

Matthaei, J.H., O.W. Jones, R.G. Martin, and M.W. Nirenberg, Proc. Nath. Acad. Sci., 48, 666 (1962).

Khorana, H.G., S. Nishimura, and D.S. Jones, J. Mol. Biol., 13, 283 (1965).

Nirenberg, M.W., and J.H. Matthaei, Proc. Nath Acad. Sci., 47, 1588 (1961).

Speyer, J.F., P. Lengyel, C. Basilio and S. Ochoa, Proc. Nath Acad. Sci., 48, 63 (1962).

_____, Proc. Natl. Acad. Sc., 48, 441 (1962).

Sueoka, N., J. Mol. Biol. 3, 31 (1961).

Watson, J.D., and F.H.C. Crick, Nature, 171, 737 (1953).

Watson, J.D., in Molecular Biology of the Gene (1976 edn.).

Watson, J.D. in "Molecular Biology of the Gene", W.A. Benjamin Inc., New York (1976 edn.)

Woese, C.R., Prog. Nucleic Acid Res. Mol Biol, 7, 107 (1967).

Yanofsky, C., "Gene Structure and Protein Structure", Harvey Lectures, 61, 145 (1967).

Yčas, M., "The Biological Code" North Holland, New York (1969).

Note: This was the basic work consulted for the first part of this chapter.

****

CHAPTER VII

NUMERICAL TAXONOMY

Taxonomy is the science of classifying species into groups (taxa) according to qualitative or quantative similarities or dissimilarities among them. Numerical Taxonomy, as the name suggests, comprises the quantitative methods of classification.

Numerical Taxonomy may be defined as the numerical evolution of the affinity or similarity between taxonomic units and the ordering of these units into taxa on the basis of their affinities. It is the aim of Numerical Taxonomy to develop methods by means of which different scientists, working quite independently, will and must arrive at identical estimates of the affinity between two organisms, given the same characters on which to base their judgements.

Classification is one of the fundamental concerns of science. Facts and objects must be arranged in an orderly fashion before their unifying principle can be discovered and used as the basis of prediction. Many phenomenon occur in such variety and profusion that unless some system is created among them they would be unlikely to provide any useful information.

Difference between Classification and Identification

When a set of unordered objects has been grouped on the basis of like properties, biologists call this classification. Once a classification has been established the allocation of additional unidentified objects to the correct class is known as identification.

The purpose of taxonomy is to group the objects to be classified into "natural taxa".

The Nature and Properties of Classifications

(a) The Basic Axiom

A population consists of elements, each of which can be individually described by reference to a pre-determined list of relevant characteristics. This population is sub-divided into sets of elements. These sets for the sub-division, should fulfill certain requirements to rank as a classification. The requirements are:

(i) Within each of the many membered set, there must be atleast one other member, which shares at least one relevant characteristic with another such member.

(ii) To be the member of a particular set is not considered to be a relevant characteristic.

(iii) Each member of one set must differ from a member of another set by atleast one relevant characteristic.

(b) Mono-thetic and Polythetic Classifications

Classifications based on one or only a few characters are generally called as "Monothetic". This means that all the objects allocated to one class must share the character or characters under consideration.

Classifications based on many characteristics are called "Polythetic". They do not require any one character or property to be universal for a class. In such cases a given "Taxon" or class is established because it contains a substantial portion of the characters employed in the classification. Assignments to the taxon is not on the basis of a single property but on the aggregate of properties and any pairs of members of the class will not necessarily share every character.

(c) Maximisations

1. Principles of Maximisation

The basic axioms as discussed above define a large number of alternative classifications, and hence a further constraint is needed to select from among them. This has to be done in such a way that differences within sets are to be minimised and the differences between sets are to be maximised. There are two methods of doing this.

(a) Self Structuring Methods

(i) A function of the relevant characteristics is defined between pairs of elements.

(ii) An element may be either a member of a population or an entire set. If it is a set, then this set may be defined by any one of its members or by all of the members or by an element constructed from all of its members.

(iii) Sets are to be constructed so that the function is minimum or maximum within them, maximum or minimum between them or both.

(b) Derived Structuring Methods

(i) A function is defined between pairs of relevant characters over a given set of members.

(ii) A characteristic or a group of characteristics is found for which the function or a derivative of the function is maximal.

(iii) Sets of members are defined in relation to the characteristics so selected.

2. Equal Weighting

Sokal and Sneath (1963) accepted the "Adansonism" postulate that every character is of equal weight.

In computing the similarity between the two taxonomic entities, numerical taxonomy treats all taxonomic characters

an of equal value and importance. If one cannot decide how to weigh the features, one may give them equal weight; unless one proposes to allocate weight on irrational grounds. For example, even if the entire genetic constitution of a form were known it would be impossible to find a basis for weighing the genetic units since they have no fixed adaptive significance. Neither can one weigh characters according to their constancy with a certain taxon, because their constancy cannot be determined until one has defined the limits of taxon.

Thus the aim of taxonomy is to yield taxonomic groups which bring together organisms with the highest proportion of similar attributes.

(D) Hierarchical and non-hierarchical classification

These are of great advantage to the taxonomist since they enable to compare taxa at any desired level. The vast majority of existing numerical methods are hierarchical in nature. However, it is possible that each level in division is associated with some measure which shall fall as the hierarchy descends.

Natural taxa

Gilmour has discussed the properties as the sort of classification which can be recognised as natural. Such classifications may be termed as general in distinction to

special classifications. This implies that in constructing general classifications, characters have equal weight, and that taxa are based on correlations between features. These two postulates are called "Adansonian".

Members of a natural taxon are mutually more highly related to one another than they are to non-members. This leads one to try to define what taxonomic relationships are. Conventional taxonomists wish to equate taxonomic relationships with evolutionary relationships, but numerical taxonomists have pointed out that taxonomic relations are of three types.

1. Phenetic Relationships

Phenetic relationships are based on overall similarity among the objects to be classified.

Taxonomic relationships are to be evaluated purely on the basis of the resemblance existing in the material at hand. The relationships are therefore static or as Michener called them "Phenetic". They do not take into account the mode of origin of the observed resemblances or the rate at which such resemblances have increased or decreased in the past.

The restriction of taxonomic procedure to phenetic evidence is necessary for three distinct reasons:

(i) Phylogenetic speculation is not compatible with the stated aims of objectivity and repeatability.

(ii) The available fossil record is so fragmentary that the phylogeny of the vast majority of taxa is unknown.

(iii) Then, even when the fossil evidence is available, this evidence must be first interpreted in a strictly phenetic manner. The criteria for choosing the ancestral forms in a phylogeny are phenetic criteria and are based on the phenetic relationship between putative ancestor and descendant.

2. Cladistic Relationships

Cladistic relationships are based on common lines of descent. Close cladistic relationships generally implies close phenetic similarity but this is not the case always. Difference in evolutionary rates may give rise to lineages that diverged long ago but appear more similar than a subsequently diverged pair of stems, one or both of which has undergone rapid evolution.

Cladistics involves the formulation of hypothesis of phylogenetic relationship (defined by recency of common ancestory). It seeks to identify sister groups i.e. taxa or sets of taxa, which are closely related to each other than either is to any other taxon, among all under consideration. Generally sister taxa are pairs. This is a methodological

device which works and should be applied even if -

(i) the taxa under consideration are two species which have an ancestor - descendant relationship and hence are not like sister species in the strict sense of an evolutionary bifurcation and/or

(ii) the taxa under consideration have closer relatives unknown to the systematist.

Statements of relation based on this approach specifically avoid the question of ancestor - descendant relationships. This is not of course to say that ancestor-descendant relationship does not exist - any theory of evolutionary mechanisms maintains they must. But phylogenetic affinity can only be adjudged on the basis of the shared possession of derived character states among sister taxa. We can reject a taxon as an ancestor only if it possesses atleast one non-shared specialisation. This means that the cladistic method of analysis leads to a determination of relationships among taxa based on their shared derived characters, irrespective of similarities based on ancestral retentions or convergence, or on geography or stratigraphy. Such hypothesis provides a minimal statement which does not stress ancestor-descendant links, but which is open to test through re-interpretation of old, or acquisition of new data on character state distribution or ontogeny.

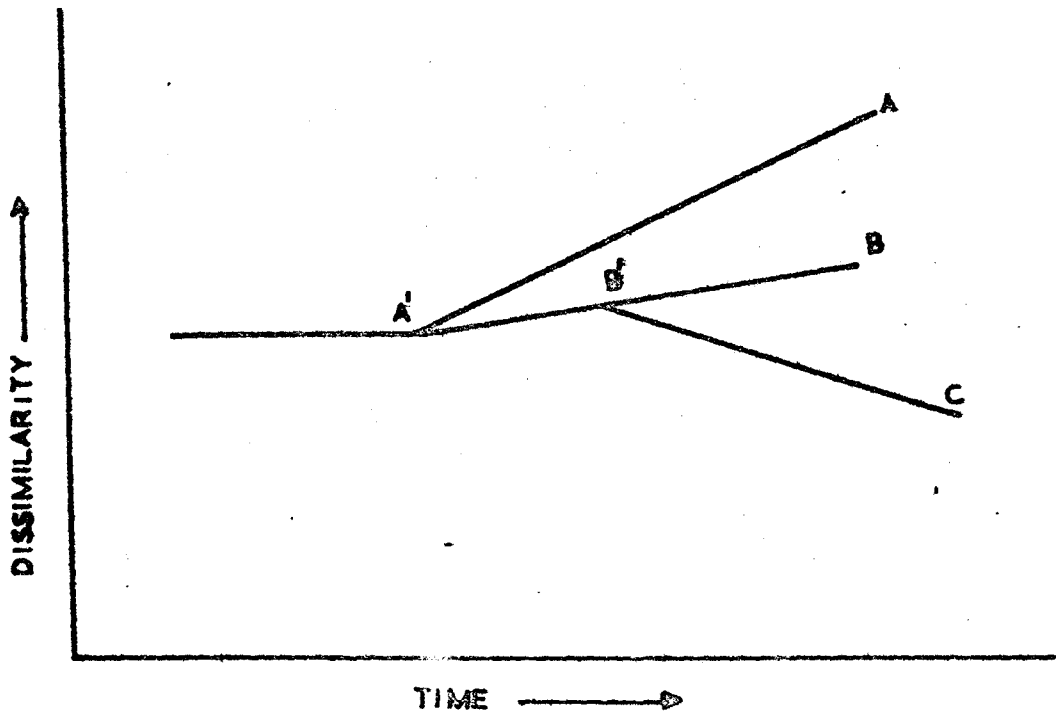


Fig 1. Taxonomic relationships viewed from three different aspects.

3. Chronistic Relationships

Chronistic or temporal relationships are among various evolutionary branches. Chronistic relationships for most organisms are known scantily if at all, and are generally inferred from phenetic evidence.

There is yet another classification which exists - "The Phylogenetic Classification".

Phylogenetic Classifications

The phylogenetic classifications of conventional taxonomy are usually based on an undefined mixture of phenetic and cladistic relationships. They often represent an overall similarity among the classified organisms disguised in evolutionary terminology.

Taxonomic relationships can be viewed from three distinct aspects. With reference to Fig. 1, we have the relationships as:

(i) Phenetically the organism B is more closely related to organism A, than it is to C, even though C evolved much later than A as a branch of stem B.

(ii) Cladistically the organisms B and C are closer to each other than either is to A, since they have an ancestor 'B' in common before either has a common ancestor 'A' with A.

(iii) Chronistically the organisms, A, B and C are closer to one another than any of them is to 'B', since they occupy the same time horizon.

From the above mentioned facts we see that Numerical Taxonomists propose to base classifications entirely on resemblance, defining natural classification as those yielding taxa whose members are in some sense more similar to one another than they are to members of other taxa. Thus the classification based on a variety of characters will be of general utility to biology as a whole, whereas a classification resting on only a few characters is less likely to be generally useful, except for the special purposes relevant to the chosen characters.

The nature of similarity is of course a fundamental problem of taxonomy. This ancient philosophical problem has become acute in a variety of fields because of the introduction of automata for classification and identification. One of the underlying principles of numerical taxonomy is that quantification of degrees of similarity is possible. Similarities can be established only on the basis of homologous or corresponding characters. Homology as interpreted by numerical taxonomists is the existing over all similarity in structure rather than similarity due to common ancestry, although this may often be the underlying cause. To describe such essential similarity due to common ancestory, one needs to base it on numerous "unit characters" of the structures to be compared. Numerical Taxonomists regard unit characters as those that cannot be sub-divided into logically or empirically independent characters. This is a complex subject, since the same set of biological

characters can be described in innumerable slightly varying ways. All these descriptions may not be used, yet one cannot avoid redundancy by choosing the best ones.

Another problem that arises is how many characters have to be chosen for describing phenetic similarities? Is there an asymptotic similarity among organisms that is approached when more and more characters are measured, or will each additional set of characters contribute a new dimension to similarity, making the taxonomic structure of a group inherently unstable?

The philosophical origins of the present development in taxonomy derive from the work of "Michel Adanson" an eighteenth century "French Botanist" who rejected a priori-assumption on the importance of different characters and proposed basing natural taxa on his essentially phenetic concept of affinity.

The objects to be classified are called "Operational Taxonomic Units". They may be individuals as such, individual representing species or higher ranking taxa such as "genera" or "families of plants and animals" or statistical abstractions of the higher ranking taxonomic groups.

Classification by Numerical Taxonomy are based on many numerically recorded characters. These may be measurements that are appropriately represented numerically, or they may be coded in such a way that the difference between them are proportional to their dissimilarity.

Choice of the Characters

In order to estimate resemblance between organisms one sets up a table of data in matrix form.

Characters ↓ / Taxa →	1	2	3	4
1	8	9	7	2
2	9	9	9	8
3	2	2	1	2
4	6	8	2	3
5	5	3	2	1
6	7	6	3	1

Columns are the organisms to be investigated and rows are the characters.

Characters however should be unit characters or if they are multiple, they should be broken into unit characters. Unit characters may be defined as those characters which cannot be further divided into logically or empirically independent characters.

The ruling idea behind this is that each state should contribute one new item of information.

Secondly, the unit characters must be meaningful. One should avoid logical correlations and must include any property which is a logical consequence of another. For example, both the diameter as well as the radius of a

circular organ will be of no use. Partial dependence of characters must also be included.

Then, if one has the evidence that more than one factor determines the two correlated features, whether the evidence comes from within the study or without it, one would then be including both characters, but otherwise only one.

Another way of representing similarity is the distance between OTU's in a multi-dimensional space. Suppose the similarity between all possible pairs taken from four objects is to be estimated on the basis of three characters. Each OTU is plotted into the three dimensional space according to its state or value for the three characters. Those objects that are very similar will be plotted close to each other while the dissimilar ones will be considerably farther apart. The computation of such straight line distances is quite simple then. In any real case, however there will of course be more than these characters and a multi-dimensional space would be necessary.

The results of a numerical classification are usually represented by means of a phenogram. These tree like diagrams indicate the similarity between OTU's or stems bearing more than one OTU along one axis. Because phenograms collapse multi-dimensional relationships into two dimensions, there is an appreciable distortion of the original relationship as can also be seen from a similarity matrix.

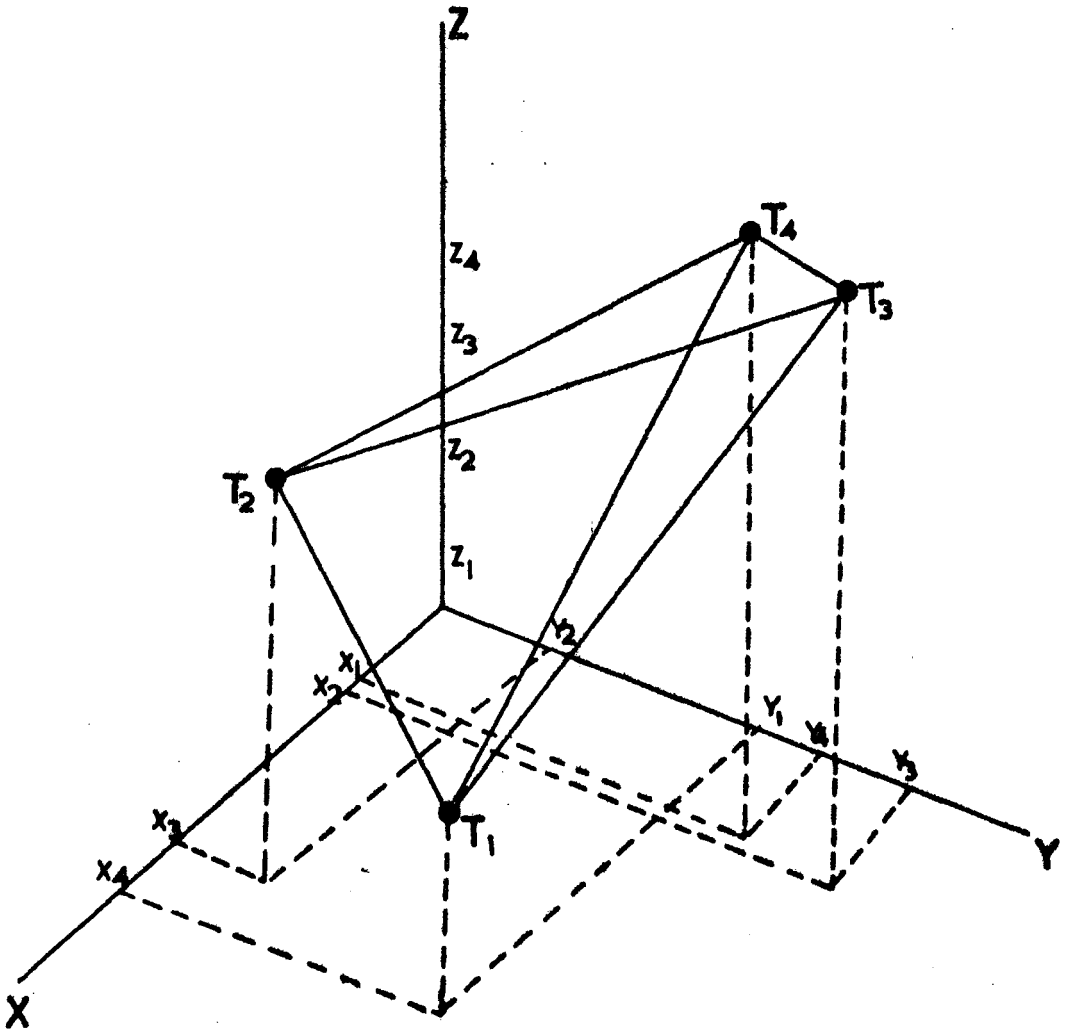


Fig 2. Similarity can be represented as the distance between the objects to be classified called OTUs in a multi dimensional space. In this example the similarity between all possible pairs taken from four objects is estimated on the basis of three characters, which are represented by the three co-ordinate axes X, Y and Z. Similar objects are plotted closer to one another than the dissimilar ones.

Computation of Resemblance

Three kinds of coefficients have been developed for resemblance or similarity.

(i) Coefficient of association - They refer to characters divided into only two classes, such as positive and negative characters. The data can be tabulated in a 2x2 table.

(ii) Correlation Coefficient used by Sokal and Michener (will be discussed later).

(iii) A third method discussed by Sokal measures taxonomic distance. It uses the convention of a multi-dimensional space with one dimension for each characters. (Fig. 2)

$$\delta^2_{24} = (x_2 - x_4)^2 + (y_2 - y_4)^2 + (z_2 - z_4)^2$$

Dendogram

A dendogram is any tree like diagram for representing relationships among genes, organisms, populations, species or other biological meaningful O.T. Units. The dendogram itself arises from a single source or root and branches progressively until it terminates in a collection of buds, the contemporary OTU.

Each non root point has one parent and in the class of dendograms each non-terminating point has exactly

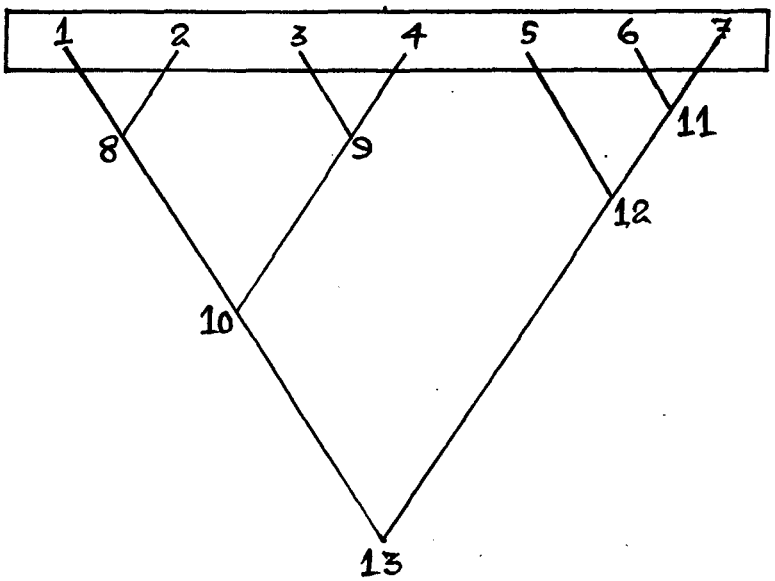


Fig 3. A Dendogram.

two off-springs. There are three regions in a dendogram, exterior, interior and the root. The exterior is the set of all contemporary Operational Taxonomic Units. The root being the most ancestral point on the dendogram. The interior is the set of all remaining points. Each member of the exterior has exactly one immediate ancestor (or parent) and no descendants. The root has exactly two immediate descendants and no ancestors. Each member of the interior has exactly one parent and two offsprings. A dendogram with n contemporary OTUs always has one root, n exterior points and n-2 interior points.

Various approaches to this Dendogram Problem

There are two theories which can be used for reconstructing the phylogeny of OTUs and have also received rigorous mathematical foundations. The first is a theory of Camin and Sokal (1965). The conditions of this theory as modified by Estabrook (1965) (who gave the theory its mathematical foundation) may be summarised as follows:

(i) discrete character states for each OTU are known.
(ii) the ancestral order of appearance or relative primitiveness of these character states is known and is reversible.
(iii) the ultimate ancestral state happened only once, and
(iv) evolution is as parsimonious as possible.

This hypothesis however cannot be applied to studies using molecular character states in which phylogeny is inferred from amino acid sequence, D.N.A. hybridization etc, simply because there is hardly any reason to believe that any nucleotide is intrinsically more primitive than another in the sense that one can often order the "primitiveness" of classical morphological characters.

The second theory (Moore 1971) postulates that the more ancient the common ancestor for a pair of OTUs, the greater is the genetic distance for that pair of OTUs. Thus, the degrees of amino-acid sequences homology are represented from knowledge of the genetic code as minimum mutation distances (Fitch, Margoliash - 1967). If molecular evolution has occurred at a uniform or nearly uniform rate in all lines of descent, then the unweighted pair group method of Sokal and Michener (1958); the complete linkage method of Sorenson (1948), the single linkage method of Sneath (1957), the divisive analogs of Moore (1972) of these agglomerative methods can produce correct dendograms, that is phylogenetic trees from dissimilarity matrices.

Fitch and Margoliash (1967) method for Constructing Phylogenetic Trees from amino acid sequence data

They produced a tree first, by the unweighted pair groups method using the dissimilarity matrix of minimum mutation distances from all pairwise comparison of aligned

amino acid sequences in the set, then these original input minimum mutation distances are reused;

(i) to estimate how the lineages descending from a common ancestor vary among themselves, with respect to the amount of mutational change, and
(ii) to calculate a reconstructed mutation distance for each pair of OTUs in the set.

Then the reconstructed mutation distances are compared to the input minimum mutation distance and as a measure of the deviation between these reconstructed and original distances, the average per cent standard deviation coefficient of Fitch and Margoliash (1967) is determined for the entire tree.

The Additive Hypothesis

The evolutionary hypothesis which bears a close resemblance to what is happening in the data sets is the "additive hypothesis" of Cavalli Sforza and Edwards (1967). The additive hypothesis states that in a true phylogeny, there is a certain plus number of mutations which are fixed between any branch point on the phylogenetic tree and each of its immediate descendants. Furthermore, the numerical value in a comparison matrix corresponding to any pair of contemporary OTU is proportional to the number of mutations fixed since the time those two OTUs shared a common ancestor.

However, for actual amino acid sequence data sets, the additive hypothesis never holds in the strict sense. Firstly, because multiple mutations at the same nucleotide positions increase with time and since these multiple changes are not revealed by the pairwise comparisons of contemporary amino acid sequence; the minimum mutation distances between very anciently separated sequences, and are much more grossly underestimated (teleost and mammalian α^{t} hemoglobin chain), than between more recently separated sequences (e.g. such as between different mammalian α^{t} hemoglobin chains).

If classifications are to be established on overall similarity, numerical taxonomy is required to put the procedures on an operational and quantitative basis. Recent development of Numerical Taxonomy starts with the simultaneous publication of papers by Peter H.A. Sneath a British microbiologist and by Charles D. Michener and by Robert R. Sokal in 1957.

The existing methods of quantitative classification are based on -

1. The G.C. percentage of the species.
2. D.N.A. hybridisation experiments.
3. Doublet Frequencies of D.N.A. molecules, and
4. Amino Acid sequence of proteins.

The G-C percentage of the species

The principle underlying the classification based on the G-C % is that organisms whose G-C % lie within a certain arbitrary range form a group. A serious drawback of this method is that since the G-C content of bacteria can lie anywhere between 25% and 75% approximately, one can get anomalous grouping (e.g. a plant and a bacterium being grouped together).

The D.N.A. Hybridisation Experiment

The method of D.N.A. hybridisation is to take single strands of the D.N.A. (obtained by denaturation) of two species and to study the rate and degree of reassociation between the two strands. A high rate and degree of reassociation would mean that two D.N.A. have a large number of nucleotide sequences in common and hence the two species are closely related.

The Doublet Frequencies of D.N.A. Molecules

In the method using doublet frequencies of various D.N.A. molecules, two types of analysis are used:

(a) The Principle co-ordinated Analysis, and
(b) The Cluster Analysis.

Principle Coordinate Analysis

The aim of principle co-ordinate analysis is to

reduce a system of n points in a s- dimensional space to a system of n points in a lower dimensional space with the minimum amount of loss of information.

Cluster Analysis

Cluster analysis (Sokal and Sneath - 1963) is concerned with a grouping of the species, arbitrarily selecting a level on the scale of similarity coefficient for the pairing of two species. For example, one could pair two species that have the minimum distance (in some units) between them compared to all other possible pairs. The next step is to calculate the distance between this pair and all other species and groups of species. The species (or groups of species) closest to the pair would then be grouped with the pair and so on.

The Amino Acid Sequence of Protein

In the classification based on proteins the amino acid sequences of the homologous proteins of various species (i.e. proteins that are common to all the species) are used. One such protein whose amino acid sequence has been established for a large number of species is the protein of Cytochrome C, a complex substance found in the cellular organelles - mitochondria in higher plants and animals, where it plays a role in biological oxidation.

The distance between organisms is defined in terms of the number of changes in the nucleotides in the gene of one of the organisms so as to produce identical proteins. These distances are then used to classify the organisms.

Comparison of the above Quantitative Classifications

In all the above schemes, the relation between various species is exhibited through what is known as a phylogenetic tree. This is a diagram consisting of nodes and branches, where the nodes represent points of evolutionary division of two species having a common ancestor and the length of the branches (in some units) represent the degree of divergence. For any such tree there is a "Point of Earliest Time" and radiating from this point, time increases along branches, with protein sequences from present day organisms at the ends of the branches. The location of the "point of earliest time" - the connection to the trunk of the tree cannot be inferred directly from the amino-acid sequences but must be estimated from other evidence.

In the first three methods, the classification is done at the level of D.N.A. mols. rather than at the protein level as in the fourth method. Since D.N.A. stores more information than the proteins (part of the information is lost due to degeneracy of code) and since part of the D.N.A. molecule is used for coding t-RNA, r-D.N.A. etc., the first three methods are more suitable for classification.

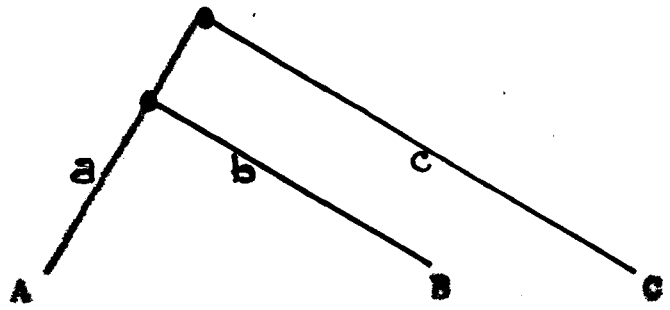
Thus Numerical Taxonomy is the quantitative classification of organisms into groups. Classifications can be performed at the level of protein (its amino acid sequence) and at the level of the D.N.A.

1. Classification at the Protein Level

(a) The principle behind this method used by Fitch and Margoliash (1967) can be understood by considering three hypothetical proteins, A, B and C with the following mutation distances:

	B	C
A	24	28
B		32

Now, which pair has to be joined first? By arbitrarily accepting that pair which has the minimum mutation distance between the two members, in the above case between A and B. So we have a tree which looks like:



where a, b, c represents the lengths of the branches.

The next probability is to get the value of a, b and c.

This can be done by solving the simple equations as

$$a+b = 32, \quad a+c = 28, \quad b+c = 32$$

giving $a = 10$, $b = 14$, and $c = 18$.

When information of more than three proteins is utilized, the basic procedure is the same; except that initially each protein forms a subset of its own. One then simply joins two subsets to create a single but more comprehensive subset. A phylogenetic tree is nothing but a graphical representation of the order in which the subsets are joined.

(b) Another alternative method of assigning proteins to sets A and B would be to choose a mutation distance which is not greater by some arbitrary amount than the minimum mutation distance and thus one can get an alternative phylogenetic tree. The best tree is one which has the minimum per cent standard deviation defined by

$$\sum_{i<j} \left[\frac{(i, j) - (j, i)}{(i, j)} \times 100 \right]^2$$

where (i, j) is the input mutation distance between species i and j, and (j, i) is the mutation distance obtained from the tree by reconstruction.

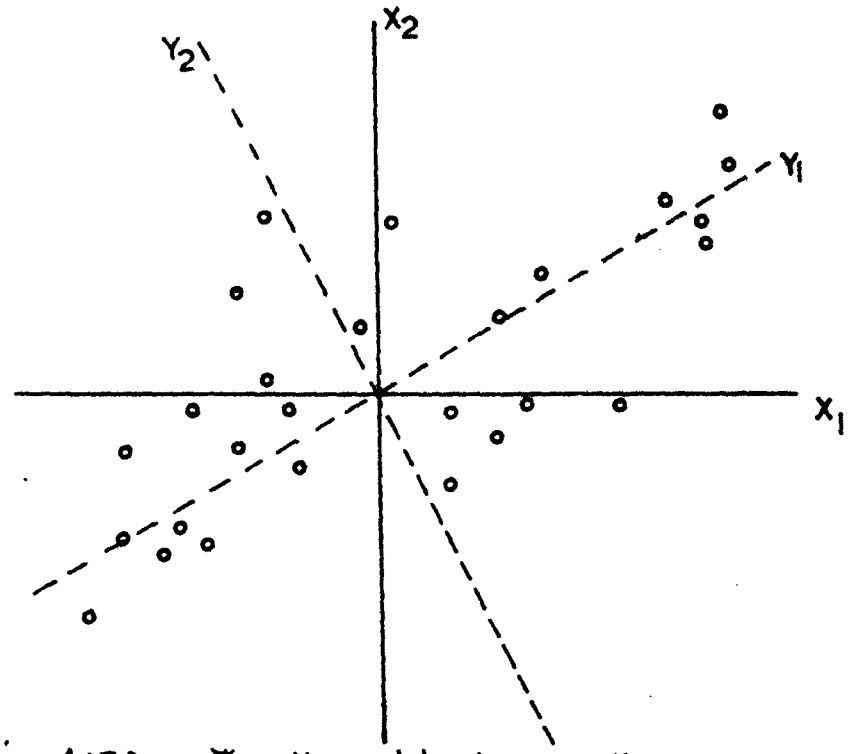


Fig 4(a) In this diagram the principal component axes are Y_1 and Y_2 for the swarm of points whose original co-ordinates were given as X_1 and X_2.

Certain anomalies were found in the results of the above analysis.

1. One was that had the frequencies of the codons been known, a more accurate estimate of the mutation distances would have been obtained.

2. Classification at the level of D.N.A.

(a) Principal Component Analysis

This is the method by which a data can be obtained by measuring the amounts of s species in each of n sample units by a scatter diagram of n points in an s - dimensional co-ordinate frame.

The most straight forward method of ordination is to project the original s-space onto a space of fewer dimensions in such a way that the arrangement of the point suffers the least possible distortion.

Supposing that one has to project a swarm of points in the plane onto a line to obtain a linear ordination. The distortion will be a minimum if the line is oriented so as to preserve as far as possible the spacing of the points. For example, Figure (4) shows results that might be obtained by sampling vegetation made up of only two species of plants. The points represent the quantities of species 1 (measured along the x_1 axis) and of species 2 (on the x_2 axis) in each of n = 27 quadrats. The origin of the

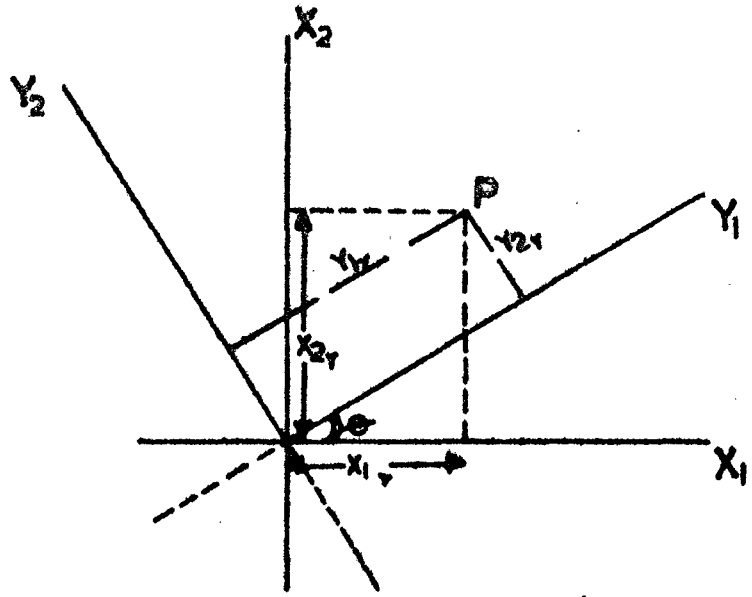


Fig 4(b). Illustration of relationship among the co-ordinate axes.

co-ordinates at $(\bar{x}_1, \bar{x}_2)$, the mean quantities of the species averaged over all quadrats.

A new orthogonal axes is obtained by subjecting the axes to a rigid rotation through an angle, indicated in the figure by the y1 and y2 axis. Then the y1 axis is the line $x1/\cos\theta = x2/\cos\phi$ where $\phi = \tan y2 - \theta$. ($\cos\theta$ and $\cos\phi$ are the direction cosines of the line).

A point with co-ordinates (x_{1r}, x_{2r}) relative to the old axes now has co-ordinates (y_{1r}, y_{2r}) relative to the new axes. A linear ordination of the points may be obtained by projecting them on the y_1 axis and one may say that the distortion produced by the ordination is least when the sum of squares of the y_1 values $\sum_{r=1}^{n} y_{1r}^2$ is a maximum. The next thing is to find out that value of θ which will give this result.

Considering fig 4(b), the point P has co-ordinates (x_{1r}, x_{2r}) in the original co-ordinates and (y_{1r}, y_{2r}) in the original co-ordinates and (y_{1r}, y_{2r}) in the new one.

$$(OP)^2 = x_{1r}^2 + x_{2r}^2 = y_{1r}^2 + y_{2r}^2$$

$$\text{or } \sum_1^n x_{1r}^2 + \sum_1^n x_{2r}^2 = \sum_1^n y_{1r}^2 + \sum_1^n y_{2r}^2$$

It also follows that since the origin is at the centroid of the swarm of points, therefore

$$\text{Var}(x_1) + \text{Var}(x_2) = \text{Var}(y_1) + \text{Var}(y_2) \quad \ldots\ldots \quad (1)$$

since $\bar{x}_1 = \bar{x}_2 = \bar{y}_1 = \bar{y}_2 = 0$

Figure 4(b) also shows that

$y_{1r} = x_{1r} \cos\theta + x_{2r} \sin\theta$ and

$y_{2r} = -x_{1r} \sin\theta + x_{2r} \cos\theta$

In matrix form this may be written as

$$\begin{vmatrix} y_{1r} \\ y_{2r} \end{vmatrix} = \begin{vmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{vmatrix} \begin{vmatrix} x_{1r} \\ x_{2r} \end{vmatrix}$$

where r = 1, 2, 3, 4, ... n.

or as $\vec{Y} = \vec{U}\ \vec{X}$ (2)

where $U = \begin{vmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{vmatrix}$

and $\vec{X}$ and $\vec{Y}$ are (2xn) data matrix.

$$\vec{X} = \begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{vmatrix} \quad \text{and}$$

$$\vec{Y} \begin{vmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \end{vmatrix}$$

$\vec{U}$ is also orthogonal i.e. $\vec{U}\,\vec{U}' = \vec{U}'\,\vec{U} = \vec{I}$

The covariance matrix of the x's is

$$\Sigma x = \frac{1}{n} X X^1 = \frac{1}{n} \begin{vmatrix} \sum_1^n x_{1r}^2 & \sum_1^n x_{1r} x_{2r} \\ \sum_1^n x_{1r} x_{2r} & \sum_1^n x_{2r}^2 \end{vmatrix}$$

$$= \begin{vmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) \\ \text{Cov}(x_1, x_2) & \text{Var}(x_2) \end{vmatrix} \qquad (3)$$

One has to find out that value of θ which will make $\sum y_{1r}^2 = n \text{ Var}(y_1)$ a maximum. From equation (1) it is obvious that maximising Var (y_1) implies minimizing Var (y_2).

From equation (2), $\vec{Y}\vec{Y^1} = \vec{J}\vec{X}\vec{X^1}\vec{J^1}$ (4)

or $$\begin{vmatrix} \text{Var}(y_1) & \text{Cov}(y_1, y_2) \\ \text{Cov}(y_1, y_2) & \text{Var}(y_2) \end{vmatrix}$$

$$= \begin{vmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} \begin{vmatrix} x_1 & x_2 \end{vmatrix} \begin{vmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{vmatrix} \qquad (5)$$

Where for convenience the (2xn) data matrix is replaced by the vector (x_1, x_2)

From equation (5), $\text{Var}(y_1) = x_1^2 \cos^2\theta + x_1 x_2 \sin 2\theta + x_2^2 \sin^2\theta$

The condition for Var (y_1) to be maximum is

$\frac{d}{d\theta}$ Var $(y_1) = 0$

i.e. $(-x_1^2 + x_2^2)\cos\theta \sin\theta + x_1 x_2 \cos 2\theta = 0$

From equation (5) it is also clear that

$$\text{Cov}(y_1, y_2) = (-x_1^2 + x_2^2)\cos\theta \sin\theta + x_1 x_2 (\cos^2\theta - \sin^2\theta)$$

$$= (-x_1^2 + x_2^2)\cos\theta \sin\theta + x_1 x_2 \cos 2\theta$$

Thus Var (y_1) is a maximum and Var (y_2) a minimum when Cov $(y_1, y_2) = 0$; the new variates y_1 and y_2 are uncorrelated.

Therefore when Var (y_1) is a maximum, equation (4) becomes

$$\vec{U}\vec{\Sigma x}\vec{U^1} = \begin{vmatrix} 1 & 0 \\ 0 & 2 \end{vmatrix} = \vec{\Lambda} \text{ or } \vec{U}\vec{\Sigma x} = \vec{\Lambda}\vec{U} \qquad (6)$$

where $\lambda i = \text{Var}(yi)$ for $i = 1, 2$.

It now becomes clear that

(1) λ_1 and λ_2 are the latent roots of the symmetric matrix Σx.

(2) They are the roots of the determinantal equation

$$\left| \vec{\Sigma x} - \vec{\lambda I} \right| = 0 \quad \text{and}$$

(3) the rows of U are the latent vectors of $\vec{\Sigma x}$.

Thus, θ may be found by solving the equation

$$\text{Var}(x_1)\cos\theta + \text{Cov}(x_1 x_2)\sin\theta = \lambda_1 \cos\theta.$$

Now, instead of defining any point by the co-ordinates (x_1, x_2) we can treat it as the point (y_1, y_2) where

$y_1 = x_1 \cos\theta + x_2 \sin\theta$ and

$y_2 = -x_1 \sin\theta + x_2 \cos\theta$

These two new variates are thus a linear combination of the original variates (measured quantities of the species).

Inferences

(1) The new variates are so defined that y_1, which is known as the first principal component, has a maximum possible variance.

(2) The best linear ordination of the points is obtained by projecting them onto the y_1 axis, the first principal axis.

(3) In this simple two dimensional case the direction of the y_2 axis (the second principal axis) is given once that of the first is known, since it must be orthogonal to it.

(4) Also the two new variates y_1 and y_2 are uncorrelated.

Generalisation of the above result

Considering a s - dimensional case, the objective would be to find the rigid rotation of the original axes, or

equivalently the linear combination of the original variate values (the x's), that will yield derived variates (the y's) with the following properties -

(a) The variance of y_1's is to be as great as possible.

(b) The variance of the y_2's is to be as great as possible, subject to the restriction that the y_2 axis must be orthogonal to the y_1 axis. The variates y_1 and y_2 are un-correlated.

(c) The variance of the y's is to be as great as possible, subject to the restriction that the y_3 axis must be orthogonal to the y_1 and y_2 axes. There are no correlations among the variates.

(d) and so on.

(e) The final axis, the y_s axis is to be orthogonal to all the (s-1) axes already fixed.

The s direction cosines of each of the s principal axes are given by the elements of the s x s matrix $\vec{U} = (a_{ij})$ with i, j = 1, 2, 3 .... s. These may be obtained by solving the matrix equation $\vec{U}\vec{\Sigma x}\vec{U'} = \vec{\Lambda}$ (similar as equation 6)

Here $\vec{\Sigma x}$ is the s x s covariance matrix of the x's and

$\vec{\Lambda}$ is the diagonal matrix whose elements are

$\lambda_j = \text{Var}(y_i)$ i = 1, 2, 3 ... s (latent roots of $\vec{\Sigma x}$)

The direction cosines of the jth principal axes are the elements of the jth row of $\vec{U}$ (i.e. the jth latent vector of $\vec{\Sigma x}$.)

In terms of the original coordinates, the yj axis (the jth principal axis) is the line $x_1/u_{j1} = x_2/u_{j2} = \ldots = x_s/u_{js}$

The jth derived variate is

$$y_j = u_{j1} x_1 + u_{j2} x_2 + \ldots + u_{js} x_s$$

Thus the coordinates of the rth point (r = 1, 2 .... n) are given by

$$y_{jr} = u_{j1} x_{1r} + u_{j2} x_{2r} + \ldots\ldots\ldots + u_{js} x_{sr}$$

where j = 1, 2, ... s.

The transformed variates have been ranked such that

$$\text{Var}(y_1) > \text{Var}(y_2) > \ldots > \text{Var}(y_s)$$

Then depending on the information of the original data, one can disregard the variates y_{k+1}, y_{k+2} ... y_s (i.e. those with small variates) for some chosen k and retain only the k variates with largest variances, the data can be ordinated in a space of k dimensions. In general practice it has been found that the first few latent roots of $\Sigma\vec{x}$ account for a large proportion of the total variance.

Example

(a) Orloci (1966) analysing a data on the vegetation of sand dunes and dune slacks considered the most 101 frequent

species, found that the first three principal components accounted for more than 40% of the total variance.

(b) Likewise, Greig, Smith Austin and Whitmore (1967) presented a three dimensional ordination of forest types in the British Solomon Islands Protectorate.

In both the examples quoted it was found that clusters of points recognizable in two or three dimensional plots of the vegetation samples could be associated with environmental differences.

Cluster Analysis

For a precise classification, a variety of numerical clustering procedures have been developed, and these are carried out on the computer after the similarity matrix has been calculated. The analysis being done for the classification of species into groups. Different methods yield different results depending on the underlying "similarity structure" of the objects to be clustered. These several methods involve two considerations each.

1. A measure of inter group likeness is defined. (The so called similarity co-efficients).

Suppose two groups i and j fuse to form a single group k, then one can have three types of measures -

a) (i) - measures which define a property of a group.
b) (i,j) - measures which define a resemblance or difference

between two groups.

c) (i,j,k) - measures which define some difference between the original two groups considering jointly and that formed by their fusion.

The "squared euclidean distance" is defined as follows:

Considering the example of the doublet frequencies and the original 10x11 matrix in which the elements represent deviations of the doublet frequencies from random. Then the Squared Euclidean Distance between two species i and j is defined as:

$$d_{ij} = \sum_{t=1}^{10} (x_{ti} - x_{tj})^2 \qquad (6)$$

2. Secondly, the chosen measure has to be incorporated into a "sorting strategy" whereby groups of elements are extracted. Considering two groups (i, and (j) with n_i and n_j n_j elements respectively and with an inter group distance as (i, j) measure denoted by d_{ij}.

Supposing that d_{ij} is the smallest measure remaining in the system to be considered so that (i) and (j) fuse to form a new group (k) with $n_k = (n_i + n_j)$ elements and consider a third group h.

$$\text{Then } d_{hk} = \alpha_i \, dh_i + \alpha_j \, dh_j + \beta d_{ij} + \gamma \left| d_{hi} - d_{hj} \right| \qquad (7)$$

where the parameters α_i, α_j, β and γ determine the nature of the sorting strategy.

The sorting strategy that is inserted here is the flexible sorting strategy of Lance and Williams. This is the system derived from eqn (6) by the quadruple constraint.

$$\alpha_i + \alpha_j + \beta = 1 \; ; \; \alpha_i = \alpha_j \; ; \; \beta < 1 \; ; \; \gamma = 0$$

The flexibility lies in its space distorting properties depending upon the value of β. In order that the strategy be as close to space conserving as possible β is taken to be -0.25.

$$\text{Then } \alpha_i = \alpha_j = 0.625$$

$$\text{Thus } d_{hk} = 0.625 \; (d_{hi} + d_{hj}) - 0.25 \, d_{ij} \qquad (8)$$

Since d_{hi}, d_{hj} and d_{ij} are all known, d_{hk} can be calculated.

The details of the method of obtaining a dendogram of the sample of 11 species based on the 10 doublet frequencies would then be something as follows:

The computer is programmed to calculate the distances between all possible pairs of entities in the sample.

The computer output consists of an (11 x 11) symmetric matrix of distances between all possible pairs of entities. This matrix say MI is examined carefully and the smallest distance in each column is noted.

Now if it so happens that say $d_{12} = d_{21}$ is the smallest distance in the first two columns. This means that species 1 is closer to species 2 than to any of the other species. Therefore, species 1 and 2 form a group 1. If the same is true for some other pairs of species then these pairs also form groups. All other species which do not fall into such pairs remain as they are without forming groups.

The next step is to construct Matrix II whose elements represent inter group distances, group element distances and the inter element distances. The first two types of distances are calculated from MI using equation (8) whereas the third type of distances are the same as they were in Matrix I.

Again the smallest distance in each column of MII are underlined and new groups are formed in two ways;

(i) by the fusion of two elements, or
(ii) by the fusion of a third element into an already existing group.

The entire procedure as above is repeated again to get Matrix III and so on and so forth until all elements are members of some group or the other.

The final step in Cluster Analysis is the construction of the dendogram which is just a visual method of stating that species 1 and 2 combine at a level of similarity represented by the squared euclidean distance between them

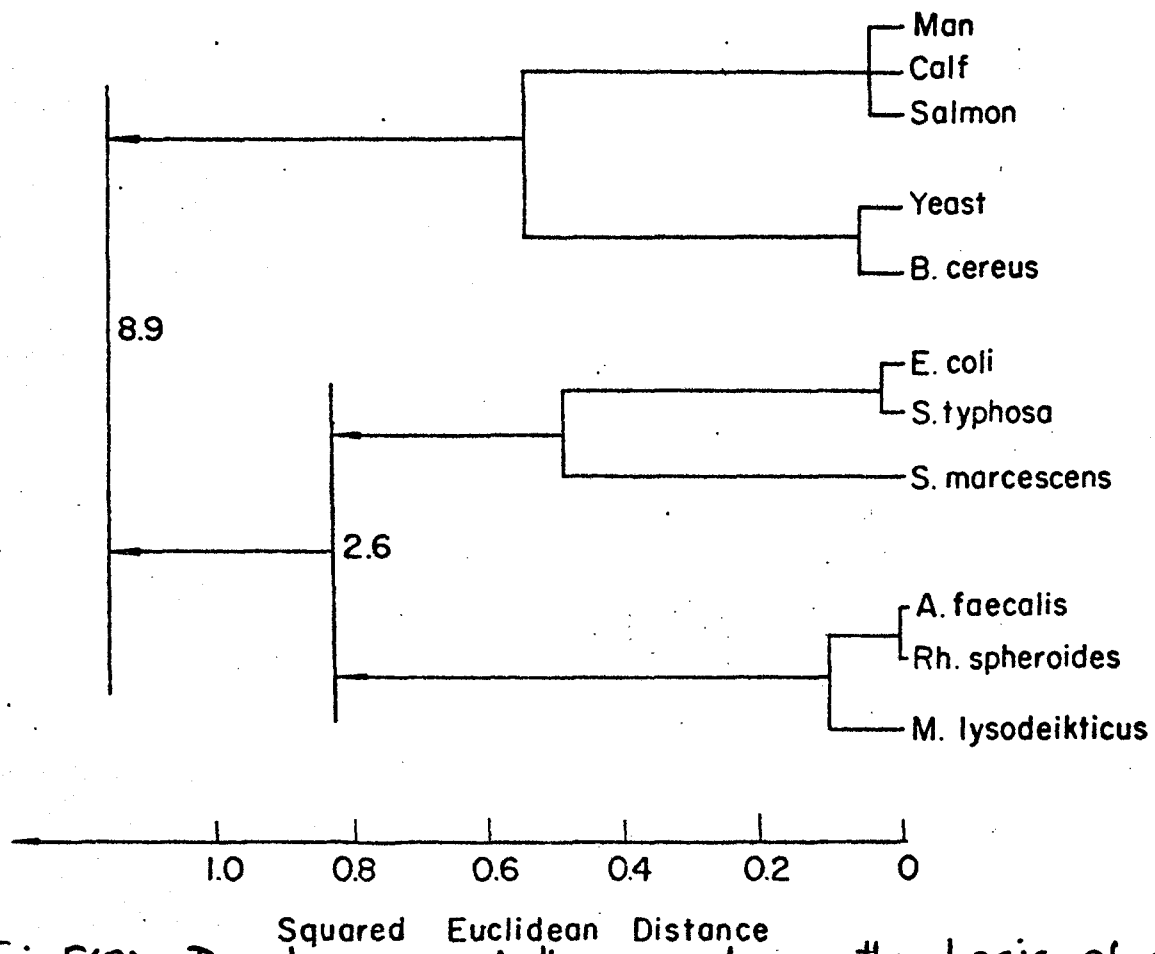


Fig 5(a). Dendogram of the sample on the basis of doublet frequencies. Data are entered as deviations of the doublet frequencies from random expectation.

and so on. The dendograms based on doublet and codon frequencies are shown in the figure (5a) and (5b) respectively. The length of the horizontal lines represent the distances between species. The lengths of the vertical lines have no meaning at all.

These results of the Cluster Analysis based on both the codon frequencies and the doublet frequencies show an anomaly in the fact that B. Cereus falls into the same group as the mammal rather than being grouped with the bacteria. Though undesirable, this anomaly is not entirely unexpected in view of the considerable number of extrapolations, assumptions, approximations and estimations that have been made in arriving at some of the data used.

The data for both methods of classification were entered as deviations of the codon frequencies or doublet frequencies from random expectation.

Conclusion

The controversy about "Numerical Taxonomy" will continue for some time to come until a new "synthetic theory" of "Taxonomy" accepting what is soundest from various schools becomes established. The revolution that the computer has brought in Taxonomy has only just begun.

Controversy has arisen between numerical taxonomists and supporters of traditional taxonomic practices and principles. In the early days of Modern Science and even today, the classification is based on a single property or characteristic, the

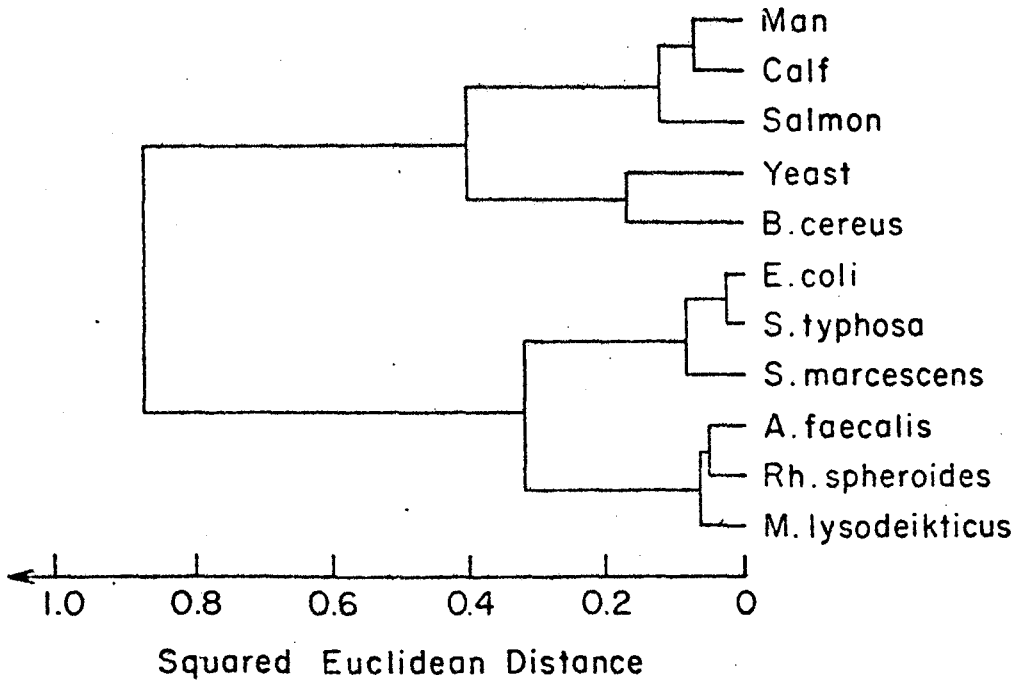


Fig 5.(b). Dendogram of the sample on the basis of codon frequencies. Data are entered as deviations of the codon frequencies from random expectation.

choice of which might be quite arbitrary. For example, (i) metals being divided into conductors and non-conductors; (ii) organisms into unicellular ones and multicellular ones etc.

Moreover since we do not yet have measures of similarity between different genetic codes, we are forced to rest to the morphological and physiological characters employed in conventional taxonomy. Recently, it has been found that though different types of characters in a taxonomic study may be correlated, this correlation is not sufficiently strong for a classification based on one set of characters (e.g. external characters) to agree fully with a classification based on a second set (internal characters).

Numerical Taxonomists working in Biological Taxonomy are continually surprised and impressed by the applicability of their principles in numerous sciences and other fields of human activity. This knowledge is spreading throughout the biological, medical, geological and social sciences, as well as the humanities. The broad spectrum of application of Numerical Taxonomy should not surprise, because after all it is "the precise categorization of human experience which is one of the foundations, for a scientific understanding of the "universe".

...

References

Bellett, A.J.D., J. Mol. Biol., 27, 107 (1967).

Dayhoff, M.O., Sc. Am., 221, 87 (July 1969).

Fitch, W.M., and E. Margoliash, Science, 155, 279 (1967).

Gilmour, J.S.L., Nature, 139, 1040 (1937).

Gilmour, J.S.L., Nature, 168, 400 (1951).

Goel, N.S., J. Theor. Biol., 16, 440 (1967).

Huber, J.E., and D.E. Koshland, Jr., J. Mol. Biol., 50, 617 (1970).

Krzywicki, A., and P.P. Slonimski, J. Theor. Biol., 21, 305 (1968).

Lewis, G.N., and W.T. Williams, Nature, 212, 218 (1966).

Pielou, E., in "Introduction to Mathematical Ecology", Wiley-Interscience (1969), pp. 250-62.

Smeath, P.H.A., and R.R. Sokal, Nature, 193, 855 (1962).

Sokal, R.R., and P.H.A. Smeath, in Principles of Numerical Taxonomy, W.H. Freeman and Co., San Francisco (1963).

Sokal, R.R., Sc. Am., 215, 106 (December 1966).

Subba Rao, G., Doctoral Thesis, University of Rochester, Rochester, New York (1971).

/This thesis was the main literature consulted, for a great part of this chapter./

Evolving Genes and Proteins, edited by Vernon Bryson and Henry J. Vogel.

....

CHAPTER IV

INFORMATION THEORY

Exploration of the practical and theoretical consequences led to the discovery that biochemical specificity of proteins is carried by the exact order of twenty amino acid residues. The suggestion of Watson and Crick (1953) that the genetical information is carried by the exact order of four kinds of nucleotide pairs provides a molecular vehicle for the genetic control of protein specificity. Information Theory in its simplest form is the study of coded messages, the message being the coding of the needed amino acid sequences for protein synthesis. The information content of DNA or the code for genetic information means that there is a specific set of directions available in the structure of DNA molecules, which specifies the primary structure of every protein synthesized by the particular cell involved.

The Central Dogma - a basic postulate of Molecular Biology governs the storage transfer and reception of information in DNA, RNA and protein. The mathematical discipline that resulted from the inquiries of Shannon (1949) who found a definition for Information is known as Information Theory. The mathematical theorems of Information Theory have already been proved, nevertheless Information Theory in Biology has a character of its own.

The Basic Concepts

(A) Information Theory finds its place in biological thought through its ability to deal quantitatively with organisation and specificity. Lord Kelvin said: "When you can measure what you are speaking about and express it in numbers, you know something about it, but when you cannot measure it in numbers, your knowledge is of a meagre and unsatisfactory kind."

The very first thing to be done in Information Theory is to choose a finite set of symbols which are called alphabets. From this alphabet, the set of all possible series of length N chosen at random in accordance with a given statistical structure is constructed. Each such series is a message. The statistical structure is reflected by the fact that some symbols may be chosen more often than others. Some symbols may often, or always be followed by certain others. This means that there may be an inter symbol influence which may extend some distance along the series. This random selection seems odd, because according to the Sequence Hypothesis, we are more interested in a series which reflects an exact order of nucleotides or amino acids. The protein sequences of length N, which carry specificity, are embedded in the ensemble of all amino acid sequences of length N. The sequences which carry specificity are a tiny fraction of the ensemble.

On the other hand, the Central Dogma regards the genetic system as a communication system in so far as its operation is concerned. So we are concerned with the system and its ability to function. If the system can record, transfer, replicate etc., any given sequence chosen at random, it can also do the same for those sequences that carry specificity. We know only the statistical structure of the ensemble of sequences and not the order of the particular sequence being transmitted. The system should however handle any sequence of the ensemble. This very realisation of dealing with the statistical properties of the whole ensemble of all possible messages in order to deal with those few members of the ensemble which do carry specificity is due to Wiesner (1949), Fano (1961) and further developed by Shannon (1949).

Gamow (1954) was however the first to see that this control implied the existence of a four letter to twenty letter code. Thus by following the logical consequences of purely biological or perhaps biochemical problems one is led directly to a problem purely mathematical in character.

Let us now review the methods of Information Theory, which have been applied previously by many other authors, to the genetic message coded by the nucleotide base sequence.

B. Some Useful Mathematical Formulae

1. Conditional Probability

Conditional probability is the probability of an

event A, which has been modified by the occurrence of another event B referred to as $P[A/B]$.

The intersection of two events A and B represents the probability that both of them occur.

(a) If the two events are independent, then $P[A]$. $P[B] = P[A \cap B]$

(b) If they are not independent, then $P[A]$. $P[B/A] = P[A \cap B]$ also describes the conditional probability.

2. Source Entropy

Let there be a source S emitting x_i symbols with probability p_i. Then in accordance with the requirement that an improbable message has a high information content and vice-versa, certain quantities have been defined. One should remember that joint information is additive while probability is multiplicative.

(a) The Self Information S_I of an event s_i of probability p_i is defined as $S_I(s_i) = -\log_2 p_i$ where the base 2 determines the unit. The units of this function are the binary digits, needed for representation of a given event, and are called BITS. The bit is a technical unit of the amount of information and not a small piece of information. The unit is called nat when the log is to the base e.

(b) Shannon and Wiener (1949) defined the average information content per symbol of a source S as the Information Function. This function should satisfy three properties:

(i) H which is the measure of information, is continuous in p_i.

(ii) If there are n p_i's and if all $p_i = 1/N$; then H is increasing in n monotonically.

(iii) If the choice is composed of several successive choices, the original H is the weighted sum of these choices.

Shannon (1949) has shown that the only function of p_i which satisfies these conditions is H (S) =

$$-K\sum_{i=1}^{n} p_i \log p_i \qquad \ldots \qquad (1)$$

where K is a positive constant, n is the number of symbols s_i with probability p_i. H is thus always greater than or equal to zero. If we set K = 1 and take the log to the base 2, we have

$$H = -\sum_{i}^{n} p_i \log_2 p_i \qquad \ldots \qquad (2)$$

At $p_i = p_j$, the maximum of this function $H^{max} = \log n$ ..

From equations (1) and (2), it is clear that we are not dealing with the p_i themselves, but rather with a function of p_i, which is conserved during the recording, transmission, reception and duplication of the genetic message and which is proportional to the length of the genetic message. The conservative and additive properties of H and

the fact that it is well defined makes this quantity a useful answer to the purpose of quantifying the notion of Information in Molecular Biology.

3. The conditional probability that event x occurs given that y, z .... occurs is denoted as $p(x/y, z \ldots) = p(x, y, z \ldots)/p(y, z)$ where p (x, y, z) is the joint probability of x, y and z.

As the information transfer system is not being considered as a whole, but only an error free translation is assumed, hence a detailed consideration of Information Channels is being omitted.

Markov Processes

A Markov process is a mathematical characterisation of a series of dependent trials. Thus an m^{th} order Markov source is one such that the probability of the k^{th} trial yielding on a given symbol x_j is dependent on the outcome of the previous m trials.

The process is stationary if $P[x_j(k)/x_1(k-1) \ldots xm\,(k-m)]$ is independent of time.

In a Markov Chain if

(i) m = o, this is a random occurrence case

(ii) m = 1, this corresponds to nearest neighbour correlation (doublet frequency nonrandom)

(iii) m = 2, second nearest neighbour correlation corresponding to triplet frequencies (nonrandom).

Markov Source Entropy

(a) First Order Markov Chain

The information gained on moving one step ahead from state a_i is

$$H = -\sum_{j=1}^{q} p_{ij} \log_2 p_{ij}$$

$$= -\sum_{j=1}^{q} p(a_j/a_i) \log_2 p(a_j/a_i)$$

Hence the source entropy per symbol is

$$H(S) = H(\overline{a_i}) = \sum_{ij}^{q} p(a_i)\, p_{ij} \log_2 p_{ij}$$

where q is the number of distinct symbols.

Second Order Markov Chain

In a similar way a second order Markov source has entropy

$$H(S) = -\sum_{ij} p_{ij} \sum_{k} P(k/ij) \log_2 P(k/ij).$$

In this chapter, we are mainly concerned with rigorous development of Information Theory as applied to DNA and RNA. This work was initiated by Gatlin, following Apter and

Wolpert's critical survey of previous attempts to apply this technique.

Apter and Wolpert (1965), criticised earlier attempts of Dancoff (1953) and Raven (1961) in this field as faulty, meaningless and trivial.

Earlier Raven (1961) had estimated that there were $3X10^9$ bases in mammalian nucleus (DNA). As there are four kinds of equiprobable bases, each base carries $H = \sum_i p_i \log_2 p_i$

$$H = \sum_i \frac{1}{4} \log_2 \frac{1}{4} = \frac{1}{4} \log_2 4$$

$$= 2 \text{ bits.}$$

Thus the total information content is 6×10^9 bits. This estimate is completely arbitrary.

Dancoff (1953) speculated on information content starting from estimates of significant molecules, orientations etc. He arrived at 5×10^6 bits of total information content. Here the objection is against the choice of the building blocks, for one could even choose electrons and protons to calculate information per molecule.

Thus Raven's estimate is merely the maximum information encodable in a 3×10^9 letter sequence of four letters with equal probabilities. This can function at the most as an upper bound to the number of proteins coded.

Apter and Wolpert (1965) concluded that -

1. Information Theory is useful only when information storage and transfer are a challenge to the organism's capabilities.

2. Only the information encoded is relevant, since only that is involved in the characteristically cellular process itself.

3. When considering information storage in DNA the information transmission system in its generality is not being considered but only the encoding itself. The question is simply of translation and not of information transfer and noise.

A rigorous application of Information Theory to taxonomic and evolutionary problems was made by Gatlin (1966, 1968). She took into account the works done by Jones et al (1961) and Kinchin (1958). She recognised the connection of the transition probability matrix (nearest neighbour experiments by Jones) with the Markov Source Entropy. This enables a calculation of the Information Content of DNA per symbol for a given DNA. This results in a numerical ordering of organisms.

We are using Gatlin's notation for convenience.

Alphabet $A = \{A, T, C, G\}$ source emits sequence $(x_{t-2}, x_{t-1}, x_{to}, x_{t1} \ldots\ldots)$ of symbols x_{ti} at times t_i.

If symbols α_{ti} emitted at times t_i $(i = 1, n)$ then the set of all such sequences defines a cylinder C of length n (Kinchin - 1958).

Eg: - C = (x_{t-1}, x_{to} | A_{t1}, C_{t2}, G_{t3}, x_{t4} ....)

where κ_{ti} are the sequential symbols and n = 3.

For a given source every cylinder C has a definite probability of emission $p(C)_n$. Evidently there are 4^n cylinders of length n. The entropy of such a cylinder is 4^n.

The set of all 4^n different sequences of length n of the above form defines a set of elementary events on a finitely additive probability space. The entropy of such a set is then

$$H_n = - \sum_{i=1}^{4n} p(C_i) \log p(C_i)$$

where $H_n^{max} = \log_2 4^n = 2n$.

Gatlin has taken each different base sequence as distinct. However, this point of view will have to be abandoned when this method is refined in relation to the coded polypeptide chain (Smith - 1969, Hasegawa and Yano - 1975).

The deviation of observed uncertainty from maximum uncertainty per cylinder is defined as Information Density.

$$I_n = H_n^{max} - H_n = \log_2 4^n - H_n^{obs.}$$

Smith abandoned $H_n^{max} = \log 4^n$ also, as our knowledge includes the Genetic Code and hence reduces the maximum

uncertainty H_n^{max}.

It is argued that from the point of view of the receiver, I_n is a measure of the average information stored per cylinder. The average information per symbol of the cylinder is then I_n/n (rate of information transfer per step of encoding).

As n increases, I_n also increases and so I_n/n converges rapidly as n increases and approaches a non-zero limit.

Gatlin defines $\lim_{n\to\infty} H_n/n \equiv H \equiv$ Source Entropy

The limit $n\to\infty$ is justified only if the convergence is fairly rapid so that infinite DNA chains are not assumed.

Similarly $\lim_{n\to\infty} I_n/n \equiv I \equiv$ Source Information per symbol.

Finally the usual Markov measures have been defined.

a) The average entropy per symbol

$H_M = -\sum_{jk} P_j P_{jk} \log P_{jk}$. This is removed within the DNA to give an Information storage per symbol of $I_M = H_M^{max} - H_M^{obs}$

b) The probability of a first order Markov cylinder of length n is $P(C) = P(x_1)\, P(x_1/x_2)\, P(x_2/x_3)$ ---- $P(x_{n-1}/x_n)$.

Method of Calculations

The transitional probability matrix can be calculated from the nearest neighbour data. She gives an example of the calculations for Micrococcus lysodeikticus using the data of Josse et al (1961).

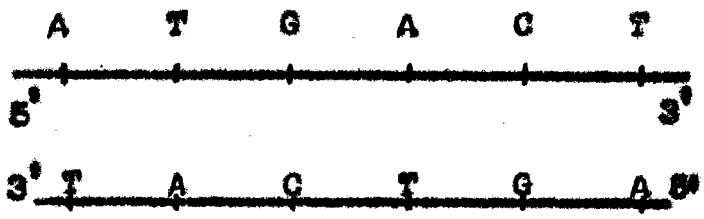
TABLE 1

Doublet Frequency Matrix as obtained by Josse and Kornberg for M. lysodeikticus*

	5'	A	T	C	G ← Given Base
Conditional base ↓	3'				Row Sum
	A	0.019	0.022	0.057	0.049	0.147
	T	0.011	0.017	0.063	0.053	0.144
	C	0.052	0.050	0.113	0.139	0.354
	G	0.065	0.056	0.121	0.112	0.354
Column Sum P(A) P(T) etc.		0.147	0.145	0.354	0.354	

* In the above data, the given base is always in the 5' and the conditional in the 3' position.

The transition probability is calculated as $p(x/_y) = p(x, y)/p(y)$ as only one strand is read, with the provision that it be read only from 5' to 3'. This is because the data of Josse have the given base (y) always in position 5'. This corresponds to anti parallel DNA chains.



Now $p(T/G) = p(T, G)/p(G) = .053/.384 = .153$ (3'–5' / 3'–5')

and so on.

The probability matrix is given in Table 2 below:

	5' A	T	C	G ← Given Base	
	3'				Row sum
	A	0.129	0.152	0.161	0.138	0.580
Conditional base ↓	T	0.075	0.117	0.178	0.153	0.523
	C	0.354	0.345	0.319	0.393	1.311
	G	0.442	0.386	0.342	0.316	1.487
Column sum		1.00	1.00	1.00	1.00	

Thus the column sum is 1, because $\sum_i P(i/j)=1$ always.

However a 3' – 5' reading of the matrix is invalid and it would mean $\sum_i P(A/A) = 1$, which we see is not true. The sum of the rows indicate the validity of this point.

As $H_n = -\sum_{i=1}^{n} p(C_i) \log p(C_i)$

1) The observed uncertainty using base doublet frequencies

$$H_2 = - \sum_{i=1}^{4} p(C_i) \log_2 p(C_i)$$

From Table 1

$$H_2 = - [.019 \log_2 .019 + .022 \log_2 .022 \cdots]$$

$$= 3.727 \text{ bits}$$

$$I_2 = \log_2 4^2 - 3.727 = .271 \text{ bits}$$

and $I_2/2 = .271/2 = 0.1355$ bits

11) To calculate $I_3/3$ etc. the probability of the triplet must be calculated.

The convention $3' = 5'$ is established.

The transition probability P_{jk} = Conditional probability of k given j, but as in Jones et al data, the given base is always in the $3'$ position.

Hence $P_{jk}^{5' \leftarrow 3'} = P\ (k^{3'}_{j} \rightarrow)^{5'}$

Similarly

$$p\ (\overleftarrow{x_1\ x_2\ x_3}) = p\ (x_1)\ p\ (\overleftarrow{x_1\ x_2})\ p(\overleftarrow{x_2\ x_3})$$

$$p\ (\overrightarrow{x_1\ x_2\ x_3}) = p\ (\overrightarrow{x_3 / x_2})\ p\ (\overrightarrow{x_2 / x_1})\ p\ (x_1)$$

Thus $p\ (\overleftarrow{GCA})^{5' \leftarrow 3'} = p\ (G)\ p\ (\overleftarrow{GC})\ p\ (\overleftarrow{CA})$

$$= (.354)\ (.399)\ (.161)$$

$$= 0.0224$$

Similarly $p\,(ACG)^{5' \leftarrow 3'} = p\,(GCA)^{3' \rightarrow 5'}$

$= p\,(A)\; p(\overleftarrow{AC})\; p\,(CG)$

$= (.147)\;(.354)\;(.342) = .0133$ bits

Then $H_3 = -[.0224 \log_2 .0224 \text{ --- } + 64 \text{ times}]$

$= 5.584$ bits

And $I_3 = \log_2 4^3 - 5.584 = 3 \log_2 2 - 5.584$

$= 3 \times 2 - 5.584 = 0.412$ bits

And $I_3/3 = 0.1373$ bits.

and so on and so forth.

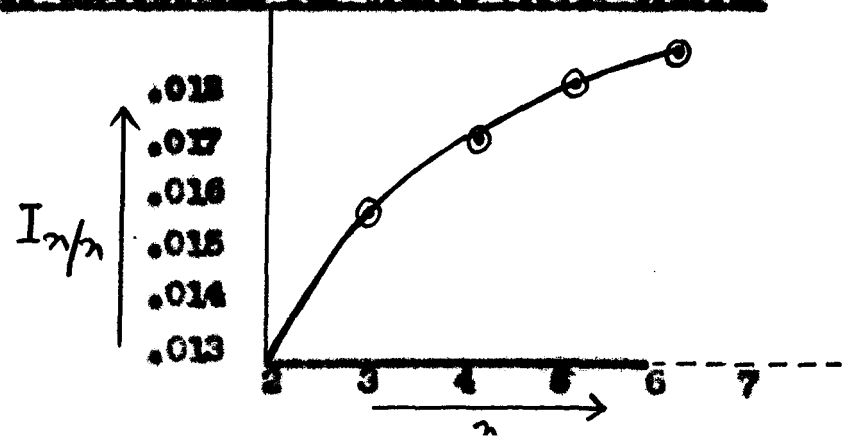
iii) Finally $I_M = 4_{\#} - H_M$

where $H_M = -\sum_{j=1}^{4} \sum_{k=1}^{4} p_j\, p_{jk}\, \log_2 p_{jk}$

Results: A typical result obtained was as shown for Mouse Thymus DNA. I_n/n was found to be bounded above by I_M. The results were arranged by values of $I_2/2$ --- $I_6/6$ and I_M; and they were found to group by taxonomic and similar criteria. The occurring was invariant whichever I_n/n was used. (Fig 1)

Fig. 1

Source Information per Symbol Versus Symbols

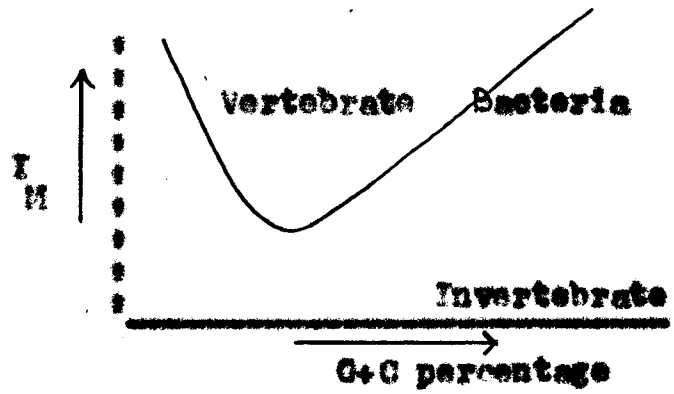


Base Composition (G+C percentage) Dependence

I_M was plotted against G+C composition. The points for the bacterial data fell on the curve shown below. This is just the usual H curve. (Fig. 2)

Fig. 2

Information Content Versus G+C Content



The curve corresponds to the fact that -

1) For Bacteria G+C varies from 25% to 75%.

11) Gatlin (1972) has shown that for simple systems the base doublet frequency is a linear function of G+C composition.

$$P_{jk} = C_{jk} \cdot \lambda$$

And under this condition, the bacterial curve should have the form

$$I = H_M^{max} - \sum_{jk} P_j P_{jk} \log P_{jk}$$

$$= H_M^{max} - \sum_{jk} P_j (C_{jk}\lambda) \log (C_{jk}\lambda)$$

$$= \log_2 4 - \sum_{jk} P_j C_{jk} (\lambda \log \lambda + \lambda \log C_{jk})$$

$$I = 2 - \sum_{jk} P_j C_{jk} \lambda [\log\lambda + \log C_{jk}]$$

Thus at maximum entropy, one should get minimum information and so on. This is a valid point.

iii) From the above derivations, it can be inferred that at any given base composition, the information measure induces an ordering among living organisms with the character $I_{vert} > I_{bact} > I_{invert}$ because it reflects average information per symbol.

Discussion on Gatlins' Work

1. The above mentioned results are dubious since bacteria are given a higher position than invertebrates. The degeneracy of the genetic code is not taken into account also at any time. The genetic message must be seen in relation to the functional efficiency as reflected in the polypeptide message and hence degenerate codons must be counted as one. This has been later taken into account by Smith (1969). Masaud and Hasegawa (1975) also obtained some important results using the Information Theory.

2. L.L. Catlin (1968) has obtained other interesting results for unicellular organisms which reflect the elegance of this technique. She recognised that this is because the inter-generic λ in unicellular organisms is large, while the intra-cellular is negligible as compared to the vertebrates, which are closely grouped around 42% in λ. But this value is the average of the compositions of DNA from various cell organelles with quite different functions. And so it, as an average of another average reflects phenotypic characteristic much more crudely.

Thus if I_{total} = No. of bases X I_M,

then the data is as given in Table 3.

Table 3

For E. coli Bacteriophages*

	PHAGE	I_M		$I_{total} \times 10^4$ bits		Remarks
GI	T_2	0.060		1.30		Lysogenic (Cell ruptuting very virulent)
	T_4	0.0635	Mean .06	1.82	Mean 1.39	
	T_6	0.0581				
GIII	T_5	0.0473		0.85		Degrades host DNA but does not rupture cell wall.
	T_1	0.185		0.13		Non Virulent
	λ^+	0.0108	Mean .0175	0.13	Mean .13	
GII	ϕX-174	0.0242		0.13		
	E. Coli B	0.0217	Mean .0188			Much greater than all of above.
	E. Coli-Ba	0.0160				

* Reproduced from Catlin - 1968.

From Table 3, it can be inferred that

Virulent Semi Virulent Non Virulent

I_M G1 > I_M G3 > I_M (*E. coli*) > I_M G2

while I_{tot} (*E. coli*) > I_{tot} G1 > I_{tot} G3 > I_{tot} G2

This ordering is readily comprehensible on the following considerations.

1(a) The high I_M, I_{tot} of GI viruses corresponds to the carrying of sufficient information to code for all enzymes, coat - proteins etc. and hence further synthetic activities of host DNA may be dispersed with and after a suitable number of reproductions using the host Ribosomal and Mitochondrial (ATP) system; the host is destroyed.

(b) In GII, I_M and I_{tot} are both small. This is because they are less self sufficient and need host DNA to code for t - RNAs etc. Hence these phages are non-virulent "parasites".

(c) T5 (GIII) occupies an intermediate position and correspondingly degrades host DNA only but does not breakdown the cell wall completely.

(d) These observations become even more comprehensible when it is noted that I_M (G I) > I_M (*E. coli*) (by about 3 - 5 times). As I_M corresponds to the rate of information "emission" it is followed that even though

I_{tot} (*E. coli*) > I_{tot} (C I), the CII phages will always win. (This is because of the reasonable assumptions of polypeptide chain length for CII and *E. coli*).

This is so because CII viruses synthesize their required proteins (such as enzymes for cell wall degradation) approximately five times as fast as *E. coli*.

It has been inferred that this is the reason why the 'COMFORT' of the injected cell as an environment for the CI phages steadily decreases and hence provides all the more reason for them to disrupt the cell and look for fresh hosts. Conversely the *E. coli* is better able to compete with the CII viruses. This arrangement is thus stable from an evolutionary point of view.

Reason for Failure of Gatlin's Work

L.L. Gatlin did not take into account the highly degenerate nature of the genetic code; simply because it was not known completely at that time. Another defect was the use of I_n and the ignorance of next to nearest neighbour correlations.

Smith and Gatlin both used double stranded data for DNA.

Later in 1969, Smith took into account the degeneracy of the genetic code.

But in 1975, all the above three failures were taken into account by Hasegawa and Yano.

Smith's Work : (1969)

Smith argued that since the message concerned amino acid sequence, so degenerate codons should not be taken as distinct words but regarded as a single word. He thus defined "Informational Uncertainity"-H as $H = -\sum_{l} p_l \log p_l$

where $\sum_{l} p_l = 1$, p_l represents the probability of the l^{th} amino acid, ochre or amber codons. The sum is thus over the codons corresponding to the l^{th} word. Thus l runs over the twenty amino acids and Ochre and Amber taken as two distinct words, that is, in all there are twenty-two words.

The informational density I is defined as

$$I = H^{max} - H^{obs} = H + \sum_{l} p_l \log p_l$$

Again as the present day knowledge includes the degeneracy of the code, Smith stated that H^{max} cannot be taken as the maximum entropy of a random chain but in fact it must be maximised in the appropriate variables.

This means that $H^{max} \neq \log n$

Smith constructed both a 'Random' as well as 'First Order Markov Model'.

Random Model

The probability of the l^{th} amino acid being commanded for, is given as $p_l = \sum_{d} p(i)\, p(j)\, p(k)$

where p(i) = base percentage of codon (nucleotide) i.

$p(j)$ = j^{th} letter apriori probability; and the sum is over all codons coding for the same amino acid (degenerate codons for 1)

There are three constraints,

1) $\Sigma p_1 = 1$

2) $[A] = [T]$, and

3) $[G] = [C]$

This leaves p_1 to be the function of just one parameter and that is $\lambda = [G + C]$ percentage.

Example:

$$P_{cys} = P(T)\,P(G)\,P(T) + P(T)\,P(G)\,P(C)$$

$$= P(T)\,P(G)\,[P(T) + P(C)]$$

Now $\lambda = [G] + [C]$

From constraint (3), $\lambda = 2[G]$

$$\text{or } [G] = \lambda/2 \qquad (1)$$

Again $2[A] + 2[G] = 1$

$$\text{or } [A] + [G] = 1/2$$

Using (1) $[A] = 1/2 - \lambda/2 = \left(\frac{1-\lambda}{2}\right)$ (2)

$$P_{cys} = (1 - \lambda/2)\ (\lambda/2)\ [(1 - \lambda)/2 + \lambda/2]$$

$$= \frac{(1-\lambda)\lambda}{2^2}$$

First Order Markov Chain

The probability of the 1^{th} amino acid being commanded for, is given as $p_1 = \sum_{deg} p(1)\, p_{ij}\, p_{jk}$

where p_{ij}, p_{jk} may be calculated as in Gatlin's paper; from nearest neighbour frequency data (base doublet).

There are three constraints again.

1) $\sum_i p_{ij} = 1$ for all i.

2) $[A] = [T]$, and

3) $[G] = [C]$

The sum is again over the codon degeneracies as before.

$P(1)$ is the codon's first letter a priori probability

p_{ij} is the conditional prob. that the second letter j follows the first letter i. These p_{ij}'s form an organisms genetic code transition matrix.

Smith argued that the 16 element transition matrix was reduced to 12 independent parameters by $\sum_j p_{ij} = 1$ (4 constraints)

Again because of the base pairing due to the other two constraints, we introduce

DF (A, A) = DF (T, T) etc.

Six such equations are obtained, since the base doublet experiments do not distinguish the polarity or chain while only one chain is read with definite polarity. These reduce the

number of free transition matrix parameters to only seven, of which one is still lambda $\lambda = [G] + [C]$.

Smith defined "Informational Densities" for double stranded DNA as

$$I_R^{DNA} = HR_{max}(\lambda) - HM(6,\lambda)^{obs.}$$

$$I_M^{DNA} = HM_{max}(6,\lambda) - HM(6,\lambda)^{obs}$$

1) The reference points HR_{max}, HM_{max} were calculated for each organisms with λ taken to be equal to an experimentally measured value. Maximization for $HM(6,\lambda)$ was done by a Taylor series expansion - a mini-maximum random search method. $HM(6,\lambda)^{obs}$ was calculated from doublet frequency data. His results confirm the same taxonomic clustering by using I_R and I_M that Gatlin had demonstrated.

According to his calculations, $\lambda = 36.2$ to 70.8 for E. coli phages.

$I_R = 0.02$ to 0.04, and

$I_M = 0.06$ to 0.08

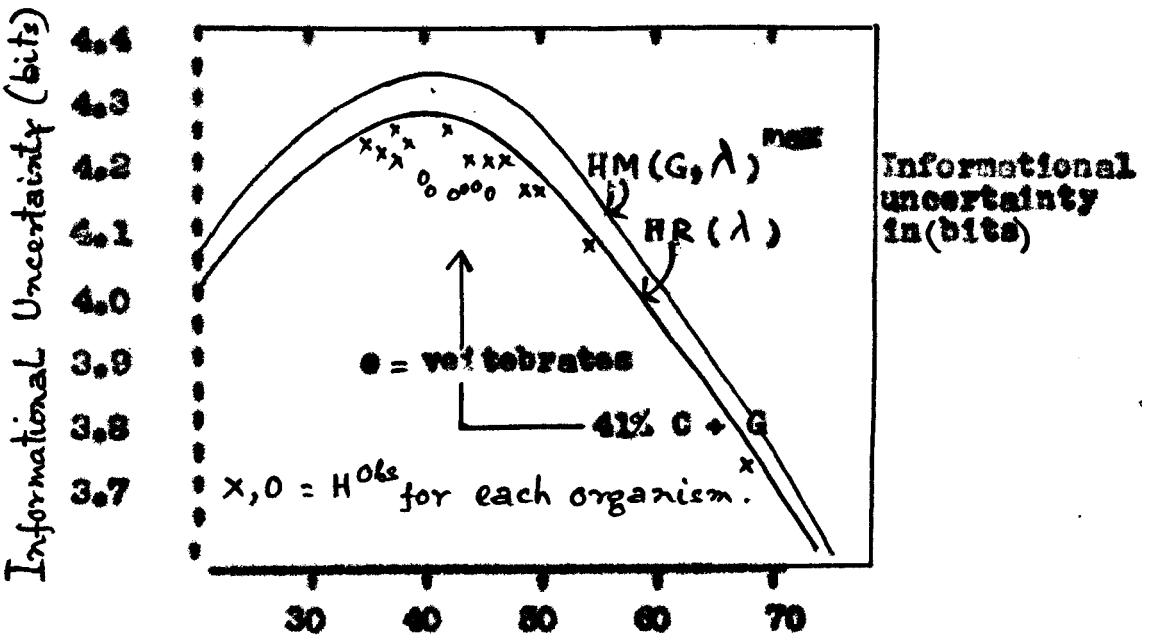
Thus it can be seen that I_M is a better measure.

2) Results for $H_R(\lambda)$ and $H_M(6,\lambda)$ are much more interesting. He showed that both have an absolute maximum at $\lambda = 415$. This is also the value at which the values for all the vertebrates cluster. Smith thereupon concluded that

evolution seems to work by coding "higher organisms" information more "effectively", by working towards a maximum in $H_M(n,\lambda)$. However, this result was not obtained by Gatlin who had ignored codon degeneracy. All organisms were found to have H below the maximum curves. (Refer Fig. 3)

Fig. 3*

<u>The Informational Uncertainty for each Organism analysed with a plot of $H_M(n,\lambda)^{max}$ as a function of λ.</u>



* Reproduced from Smith - 1969

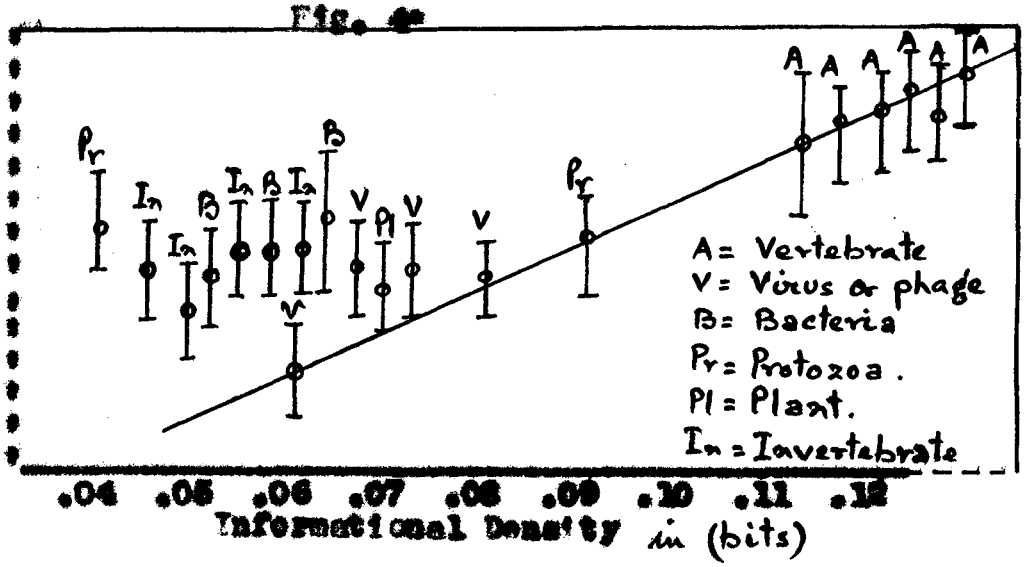
8. Then Smith introduced an evolutionary measure parameter e, by assuming that evolution begins from a random base distribution, whence the root mean square deviation between observed and random transition matrix elements, λs

$$e = \left[\sum_{ij}^{4} \frac{(TM_{ij} - TR_{ij})^2}{4} \right]^{\frac{1}{2}}$$

The greater the deviation, the higher is the organism.

On plotting e versus I_M^{DNA} and fitting the vertebrate points to a straight line, he got the graph shown in Fig. 4

Fig. 4*



* The Informational Density versus the evolutionary measure e, fitted with the proposed line of maximum selection efficiency. (Reproduced from Smith - 1960)

Conclusion from Smith's work

Inference

Thus all the lower organisms were found above the line which thus seems to act as the limit to the information density available by a given value of e* (deviation from random) and this is what the higher animals have achieved.

Masami Hasegawa and Yano's Paper of 1975

The article "The Genetic Code and the Entropy of Protein" (1975) of Hasegawa and Yano is the most well argued and coherent of this series and presents the most clear cut conclusions. They made the following observations, about the work done previously in this field.

1) As pointed out earlier by G.B. Weiss (1970), the genetic code used by T.F. Smith (1969) was not correct. Smith considered that UGA codes for Tryptophan, while today — — it is accepted as a chain terminating codon. Thus Tryptophan has one, not two, code words. Also since UAG (ochre), UGA and UAA (amber), all three, code for chain terminating codons, hence they should all be grouped as one word. The 64 codons then code for 21 and not 22 distinct words.

2) They questioned the use of Information Density $I = H^{max} - H^{obs}$ as an evolutionary measure.
Their objection can be understood by illustrating an example.

T C A G T C A G T C A G --- ≠

A C T C A G T C A G T C --- ≠

The above example codes for four amino acids at the most; in fact only three as TCA = AGT = Serine, GTC = Valine and CAG = Glutamine. This would give a high value of Information Density I and hence it ought to belong to a

higher rank than any existing organism which is evidently observed.

$$H^{max} = -\sum_{i=1}^{21} P_i \log_2 P_i$$

$$= -21 \times \frac{1}{21} \log_2 \frac{1}{21} = \log_2 21$$

$$H^{obs} = -\sum_{i=1}^{4} P_i \log_2 P_i = -4 \times 1/4 \log 1/4$$

$$= \log_2 4$$

Thus $I = H^{max} - H^{obs} = \log_2 21 - \log_2 4$

Evidently value of I is a maximum as $\log_2 21 \gg \log_2 4$

On studying this, it can be concluded, that in spite of this objection, Gatlin and Smith found I to be a useful information. This is because, while assuming deviation from maximum uncertainty to be a measure of the information transmitted to the polypeptide chain, they maximised H at the given λ . Thus by creating an individual reference point for each organism they removed the objection of high H^{max}.

Also, both Hasegawa and Yano chose to regard H itself as a measure of evolution. This was also done by Smith earlier. Hasegawa and Yano argued that H can be regarded as a measure per amino acid of the variety of amino acids of the proteins of the organisms. As higher

organisms will have diverse proteins to carry out their functions, hence they will require a large variety of amino acid sequence per amino acid and will thus lead to the maximisation of the uncertainty itself.

Other errors of Smith in calculations, were evaluated by Gatlin, who also argues for $\lambda = 42\%$, as being an attempt to diversify proteins. However, according to her, the biased doublet frequencies are the result of evolution as a game theoretical optimisation procedure for reducing error in a noisy channel. Hasegawa and Yano question this, since the analogy with human receivers, who use Shannon-Weiner redundancy (biased doublet frequency according to Gatlin) to reduce error by transmitting highly unique combinations, cannot be extended to ribosomal complex. Instead, they consider the biased doublet and triplet frequency as beneficial in encoding a greater variety of proteins.

King (1972) removed another defect of Smith's theory by determining the base composition of single stranded m-RNA, that would optimise the diversity of predicted amino acid frequencies as read by the Genetic Code. However he used a "Random Base Model".

Work of Hasegawa and Yano

With the above objections in mind, they considered the m-RNA base sequence as a second order stationary Markovv Chain. (They took into account both the nearest neighbours and next to nearest neighbour correlations). They optimised

base doublet and triplet frequency towards maximal H. They used the latest code [Table 3 of Chapter II].

Second Order Markov Chain

This assumption means that a base probability in an m RNA sequence depends only on the preceding two bases along the strands. As this base sequence is assumed to be stationary, the probabilities of the four bases P(i), the doublet frequencies ≡ P(i,j), the triplet frequencies P(i,j,k) in which i,j,k ∈ (U,C,A,G) are constant in any part of the strand.

Consequently $P(i,j) = \sum_{k \in U,C,A,G} P(k,i,j)$

$= \sum_{k} P(i,j,k)$

giving 16 independent constraints.

The normalisation condition

$\sum_{ijk} \beta_{ijk} = 1$ gives 16 constraints in all

They have given equation for determining doublet frequency from triplet frequencies.

Similarly $P(i) = \sum_{j} P(i,j) = \sum_{j} P(j,i)$

There are 64 Pijk's but only 48 are independent.

P(j/i) = P(i,j)/P(i)

P(k/ij) = P(i,j,k)/ P(i,j)

Entropies

The entropies have been defined as

a) $H^{2m}_{mRNA} = - \sum_{ij} P(i,j) \sum_{k} P(k/i,j) \log_2 P(k/ij)$

b) $H^{M}_{mRNA} = - \sum_{i} P(i) \sum_{j} P(j/i) \log_2 P(j/i)$

c) $H^{R}_{mRNA} = - \sum_{i} P(i) \log_2 P(i)$

d) The probability of the I^{th} kind of amino acid P(I), and probability of amino acid doublet I, J , P(I,J) are given as

$$P(I) = \sum_{ijk} P(i,j,k)$$

$$P(I,J) = \sum_{ijk} \sum_{lmn} P(i,j,k,l,m,n)$$

I degenerate ↑ ↑ degenerate J

So the entropy of amino acid sequence considered as a simple Markov Chain is defined as

$$H^{M}_{AA} = - \sum_{I} P(I) \sum_{J} P(J/I) \log_2 P(J/I)$$

The entropy of amino acid sequence considered as a Random Chain is $H^{R}_{AA} = - \sum_{I} P(I) \log_2 P(I)$

The summation is over 21 codons.

e) The probability of the MMMA sequence (i,j,k,l,m,n) is given by $P(i,j,k,l,m,n) = \frac{P(i,j,k)\ P(j,k,l)\ P(k,l,m)\ P(l,m,n)}{P(j,k)\ P(k,l)\ P(l,m)}$

Next, Hasegawa and Yano maximised H_{AA}^{N} values by a random search method in the phase space spanned by the 48 independent P(i,j,k)'s.

The 64 values of P(i,j,k) are assumed to be random Rijk ($-1 < Rijk < 1$). The upper limit of the displacement is denoted by δ and P(ijk) is altered as follows -

$$P^{1}(ijk) \longrightarrow P(ijk) + Rijk\ \delta$$

If $P^{1}(i,j,k) < 0$, then Rijk is generated again and again, until a positive $P^{1}(ijk)$ is obtained. This procedure is carried out for the 64 P(i,j,k)'s so that they satisfy the constraints calculating H_{AA}^{M}. If this value is greater than the previous value, then the new value is adopted, and the procedure is repeated. The new random numbers are used on the original valuesof P(i,j,k).

contd....

Table 4

Trinucleotide Frequencies of the H^{M}_{AA} Maximum RNA*

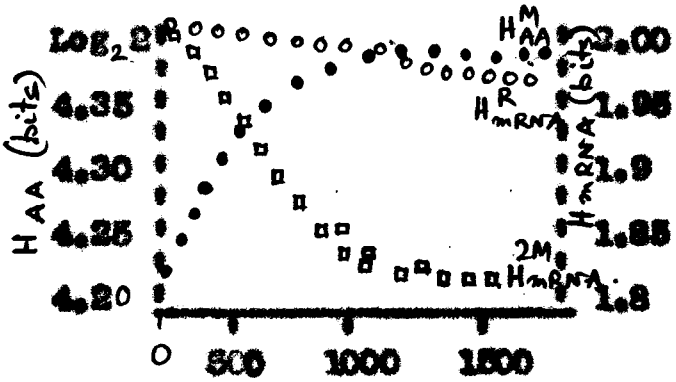
1st ↓ / 2nd →	U	C	A	G	
U	0.0299	0.0002	0.0217	0.0230	U
	0.0123	0.0046	0.0121	0.0901	C
	0.0136	0.0152	0.0122	0.0364	A
	0.0234	0.005	0.0011	0.0323	G
C	0.003	0.0075	0.0293	0.0037	U
	0.0	0.0141	0.0143	0.0031	C
	0.0098	0.0232	0.027	0.0054	A
	0.0188	0.0036	0.0304	0.0	G
A	0.0306	0.0089	0.0283	0.0123	U
	0.0038	0.0171	0.0182	0.0188	C
	0.0189	0.0302	0.0368	0.0645	A
	0.0475	0.0	0.0219	0.018	G
G	0.0187	0.0071	0.027	0.0135	U
	0.003	0.0127	0.0146	0.0113	C
	0.0108	0.0229	0.0261	0.0206	A
	0.024	0.0085	0.0192	0.0124	G

* Calculations were done by HITAC 5020E (University of Tokyo) and FACOM 230 - 60 of Hokkaido University.

Reproduced from Hasegawa and Yano - 1975.

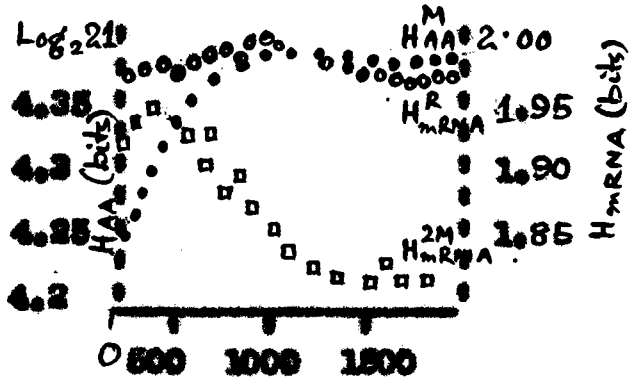
Results

Fig. 5(a)



No. of Repetitions (Random)

Fig. 5(b)



(Spleen)

[Reproduced from Nagegawa and Yano - 1975]

1. In fig. 5(a) and (b), the process of maximising H as a function of repetitions has been shown for $\delta = .02/\frac{3}{4}$. Two initial points were taken for the two models.

a) Completely Random Base Sequence.

b) Human Spleen DNA.

H^{R}_{mRNA} and H^{2M}_{mRNA} were also calculated with the same Rijk's.

2. The final states were uniquely determined for whatever initial conditions are carried out.

Final Values

$$H^{R}_{AA} = 4.376 \qquad H^{M}_{AA} = 4.367$$

$H^{R}_{mRNA} = 1.97$ $H^{M}_{mRNA} = 1.866$

$H^{2M}_{mRNA} = 1.813$

P(U) = 0.263, P(C) = 0.1757

P(A) = 0.3171 P(G) = 0.2442

P(C) + P(G) = 0.4199 = 41.99%or 42%.

The corresponding tri - nucleotide frequency are given in Table 4.

Points, Predictions and Comparisons

1. The nucleotide frequency are highly biased (Table 4).
2. The GC composition is 42%, the same value around which vertebrate DNA compositions vary. Measurements for 98 species gave 36.7% to 41.8% for main composition of DNA. So this result is better than that of Smith - 43% (1968) and King - 44.4% (1972).
3. The large difference in H^{R}, H^{M} H^{2M}_{RNA} are due to the highly biased triplet frequency.

   UCU, CUC, CGG and ACG are nearly absent in the maximal H_{AA} state. These codons should therefore be very rare in higher organisms. Since in DNA observations, TCT, CTC, CGG, ACG are arranged with AGA, GAG etc. these may be observed, but 'CGG, CCG' and 'ACG, CGT' are both rare and so these triplets will be rare.

4. Hasegawa and Yano (1975) on comparing their result with those obtained by Goel et. al found out the following agreements in spite of the objections about the inaccuracy of the data used and non-uniqueness of the solutions. These similarities are -

a) TTT (Phenylalanine codon) is used in preference to TTC by vertebrates. P (UUU) = 0.0299; P(UUC) = 0.0123.

b) TCG, CCG, ACG, GCG are rare in vertebrates. Here their frequencies are 0.005; 0.0036; 0.00; 0.0035 respectively as compared to 10^{-1} on the average.

c) They calculated the deviations of di - nucleotide frequency P(ij) from random expectations

$$d.f.r. = \frac{P(ij)}{P(i)P(j)} - 1$$

for human spleen DNA, E. coli DNA and the H^{M}_{AA} max. DNA. All average with the complementary strand and it was also found that

i) The doublet pattern of human DNA is much closer to H^{M}_{AA} maximum DNA than that of E. coli DNA.

ii) In particular the CG frequency of both are very scarce. This is a feature shared by all higher animals and is interpretable as a strategy to increase protein diversity.

The biased base composition is a result of the biased code. The code is universally degenerate.

i) Leu, Arg, Ser are 6 - D.

ii) Met and Tryp are non degenerate.

iii) Four kinds of bases are not used evenly.

iv) There is a correlation between GC content and its degree of degeneracy. Sets of codons [i.e. AXd, d = U,C, AG; AXe (e=A,G) or AXb (b=U,C)] which code for the same amino acid with maximum GC content in the first two position are at least four fold degenerate. Those without GC are two fold degenerate. Now the greater binding energy of the GC pair may obviate the reading of the third nucleotide in those triplets, containing only G and C in the first two positions, thus requiring complete degeneracy to be relinquished to such codons.

Again, an aminoacid with high degeneracy are GC rich, the GC composition should be limited so that amino acids are uniformly distributed. Thus λ = 42% is optimal in this sense.

The GC parity is now understood as -

when H^{M}_{AA} is maximum [G + C] = < [A + T]

Arginine is (CGd d = U, C, A, G and AGe (e = A,G) are 6 - D while Alanine (GCd), Glycine (GGd) and Proline (CCd) all have G or C in the first and second positions and are only 4 D. If they are to appear evenly, Arginine must prefer to use AGe, rather than CGd, so that the insufficient

nucleotide G or C may be used for Ala, Pro (because of 42% optimization).

And for this reason, only GC is scarce in vertebrates.

This is clearly seen, since P(GG) = 0.0122 and P(AGe) = 0.0375 i.e. thrice as often while one would expect P(GGG) = 2 P(AGe).

The observations on vertebrate DNA's confirm these ratios.

Conclusion

One can remark in connection with the grammar that generates the functional set of poly-peptide communications. A contradiction exists between amino acid sequence and 3-D structure and the protein function.

The absence of amino acid correlations in all investigations upto date is another anomaly.

Smith tested the predicted amino acid correlation with the random ones. Predicted correlation -

$$P_{\xi n} = \sum_{if} P_{\xi}(f)\ P_{f}(1)\ P_{n}(1) / \sum_{\xi} P_{\xi}(1)$$

where $P_{\xi}(f)$ is the probability of amino acid codon ending in base f.

P_n is the probability of n codon starting with 1, P_{f1} is the $f \rightarrow 1$ transition matrix element.

Smith did not find any difference.

Similarly Hasegawa and Yano got H^{M}_{AA} = 4.367 and H^{R}_{AA} = 4.376 i.e. no nearest neighbour amino acid correlation.

We feel that all such attempts will always exhibit negative results since a stationary chain is assumed, and the apriori correlations which lead to function are averaged out. This would however involve a time independent Markov Chain. Such a possibility has not been tackled up to date due to the non-availability of suitable data.

.....

References

Apter, M.J. and L. Wolpert, J. Theor. Biol, 8, 244 (1965)

Gamow, G., Nature, 173, 318 (1954).

Gatlin, L.L., J. Theor. Biol., 5, 360 (1963).

Gatlin, L.L., J. Theor. Biol., 10, 281 (1966).

Gatlin, L.L., J. Theor. Biol., 18, 181 (1968).

Goel, N., G. Subba Rao, M. Yeas, Brenneman and King, J. Theor. Biol, 35.

Josse, J., A.D. Kaiser and A. Kornberg, J. Biol. Chem., 236, 864 (1961).

Nasemi, Masegnur and Take-Aki-Yano, Math. Bio Sciences, 24, 169 (1975).

Smith, T.F., Math. Bio Sciences, 4, 179 (1969).

Smith, T.F., Math. Bio Sciences, 8, 293 (1970).

Watson, J.D., and F.H.C. Crick, Nature, 171, 737 (1953).

Weiss, G.H., Math Bio Sciences, 8, 291 (1970).

Yockey, H.P., J. Theor. Biol., 46, 369 (1974).

Yockey, H.P., J. Theor. Biol., 67, 345 (1977).

Anfinsen, C.B., "Informational Molecules", Academic Press, New York (1959).

Gamow, G., "Symposium on Information Theory in Biology", Pergamon Press, London (1958).

Jukes, T.H., "Molecules and Evolution", Columbia University Press, New York (1966).

Kinchin, A.I., "Mathematical Foundations of Information Theory", Dover, New York (1953).

Ratner, V.A., "Application of Mathematics to the Code and Protein Structure", Progress in Theor. Biol., vol. 3.

Shannon, C.E., "Mathematical Theory of Communication", University of Illinois Press, Urbana (1949).

Wiener, N., "Cybernetics", John Wiley and Sons, New York (revised edition, 1961).

*****

CHAPTER V

RESULTS AND DISCUSSIONS

We know very well that the transfer of information from DNA into protein involves the monomer units as the units of the code. There are four different kinds of bases in the DNA molecule. These bases determine uniquely twenty amino acids along the polypeptide chain.

We are concerned with the information content. As has been stated earlier (Chapter IV), Shannon-Wiener (1949), defined information content as the amount of selective information.

If m be a classification with categories 1 and associated probability P_1, then the information content of m is designated as $H(m)$ and is given by

$$H(m) = -\sum_{1}^{m} P_1 \log_2 P_1 \qquad (1)$$

where $\sum_{1}^{m} P_1 = 1$; the sum is over all the m distinct word probabilities P_1.

We also know that the biological code is degenerate. A detailed discussion has been presented in Chapter IV. Goel et al (1972) point out that if codon frequencies are just functions of DNA composition, then they have no specific

functions. If the codon frequencies of the different organisms are not attributable to variations in DNA composition, then we could think that the synonymous codons have different functions. And so the codon frequency could also be used to Numerical Taxonomy. Further Goel et al (1972) also assumed that the sum of codon frequencies for all the codons for an amino acid are equal to the frequencies of the amino acid.

Further, the organisation of the Genetic Code is such that most mutations result in either a synonymous code word or in a code word for a chemically similar amino acid. There are heritable changes in the genetic material, hence the mutations are true. Changes in amino acid sequences occur much more slowly than changes in total DNA. Changes in DNA which cause changes in proteins are held in check by natural selection to a far greater degree than are those which do not.

One is interested in the frequency of occurrence of amino acid in the genetic code as compared with their average levels in proteins. The code has 61 codons for amino acids. A protein with one amino acid per codon would have the following composition:

$$ala_4,\ arg_6,\ asn_2,\ asp_2,\ cys_2,\ gln_2,\ glu_2,$$

$$gly_2,\ his_2,\ ile_3,\ leu_6,\ lys_2,\ met,\ phe_2,$$

$$pro_4,\ ser_6,\ thr_4,\ tyr_2,\ trp,\ val_4.$$

Let us now know the way in which the 20 acids and terminator are coded by the natural genetic code of nucleotide triplets. If these 21 categories are equi-probable and coded optimally, then

$$H(y) = -\sum_{j}^{21} 1/21 \log_2 1/21 = \log_2 21$$

$$= 4.4 \text{ bits per amino acid.}$$

The maximum value of H(y) in a protein sequence where there are no terminators is

$$H(y) = -\sum_{j}^{20} 1/20 \log_2 1/20 = \log_2 20$$

$$= 4.322 \text{ bits per amino acid.}$$

Now, each amino acid is coded by a triplet of quarternary digits U, C, A, G. There are thus 64 different codons, each requiring some bits for its specification.

If the application to Molecular biology, degeneracy is absent in only two cases, met and trp, which have each a single codon. This term represents the redundant information due to the code degeneracy. This is the information which cannot be transferred from RNA to protein.

The maximum value of H(x) for codons is

$$H(x) = -\sum_{1}^{64} 1/64 \log_2 1/64 = \log_2 64$$

$$= 6 \text{ bits per codon.}$$

If we consider a given protein chain where there are no terminators, then i runs to 61 and not to 64. The P_i for the codons UAA, UAG and UGA are set to zero.

$$H(x) = -\sum_{1}^{61} 1/61 \log_2 1/61 = \log_2 61$$

= 5.931 bits per codon.

Calculation of Information Content

We are calculating the Information Content (H) taking into account

i) the GC content (following the method adopted by Smith - 1969)
ii) the codon frequencies (as calculated by Ceel et al. - 1972)

The GC Content Method

We are using the random base model for calculating the information content from the knowledge of the GC content.

We know that the RNA (of known G + C content) is a completely random sequence of bases. By this we mean that the frequency of a particular codon is just the product of the individual frequencies of the three bases that make up the codon.

This is not strictly true in our case because of our assumption that the frequencies of nonsense codons is zero, i.e., (UGA) = (UAA) = (UAG) = 0. In order to incorporate this

assumption we are calculating the frequencies of the 64 codons first by the method mentioned above.

Secondly, we are also calculating the frequencies of the remaining 61 codons (sense-codons) by the above method. The sum of these 61 frequencies will obviously be less than unity. In order to maintain the normalization (sum of all 64 codon frequencies should be unity) we divide each of the above codon frequencies by the sum of all the -

(i) 64 codon frequencies in the first case; and
(ii) 61 codon frequencies in the second case.

Example: The probability P_1 of a word (amino acid or ochre or amber) is given by the random model as $P_1 = \sum_d P(i)\, P(j)\, P(k)$
where d = degeneracy of P_1 codons

There are 3 constraints.

(i) $\sum P_1 = 1$ (ii) $[A] = [T]$ and (iii) $[C] = [G]$

leaving one parameter lambda.

$$\lambda = [C] + [G]$$

Also $2[A] + 2[C] = 1$

so $[A] + [C] = \frac{1}{2}$

Thus $[C] = \lambda/2$

and $[A] = \frac{1}{2} - [C] = \frac{1}{2} - \lambda = [T]$

To the probability that amino acid cysteine (2 - degenerate) is called for is given as

$$P_{cys} = P(T)\,P(G)\,P(T) + P(T)\,P(G)\,P(C)$$

$$= P(T)\,P(G)\,[P(T) + P(C)]$$

$$P_{cys} = \left(\frac{1-\lambda}{2}\right)\left(\frac{\lambda}{2}\right)\left[\frac{1-\lambda}{2} + \frac{\lambda}{2}\right]$$

$$= \frac{(1-\lambda)\,\lambda}{2^3}$$

On substitution of the value of λ for the different species under consideration, we can calculate P_{cys}. In a similar manner we calculate the probability P_1 of other amino acids.

Finally the sum of all such probabilities gives the measure of information content for that particular species.

The species for which the Information Content has been calculated in this manner are:

1. Man (G+C content = 40.8%)
2. Calf (G+C content = 42.0%)
3. Salmon (G+C content = 42.0%)
4. Yeast (G+C content = 38.0%)
5. Bacillus Cereus (G+C content = 37.0%)
6. Escherichia coli (G+C content = 50.0%)
7. Salmonella typhosa (G+C content = 51.7%)
8. Serratia marcescens (G+C content = 59.0%)
9. Alcaligenes faecalis (G+C content = 67.2%)
10. Rhodopseudomonas spheroides (G+C content = 67.7%)
11. Micrococcus lysodeikticus (G+C content = 70.8%)

TABLE 1

Species	GC Content (per cent)	Information Content	
		64 codons	61 codons
(1)	(2)	(3)	(4)
B. cereus	37.0	4.4524426	4.2926286
Yeast	38.0	4.4632673	4.3002865
Human Spleen	40.8	4.5107673	4.3150135
Salmon, Calf	42.0	4.5159006	4.3181943
E. coli	50.0	4.5002639	4.2843471
S. typhosa	51.7	4.4868781	4.2808449
S. marcescens	59.0	4.4004524	4.2028935
A. faecalis	67.2	4.1854073	3.9023557
Rh. sphaeroides	67.7	4.1701291	3.9819116
M. lysodeikticus	70.8	4.0649899	3.5144781

N.B.: The graphs have been plotted taking into account two decimal points only.

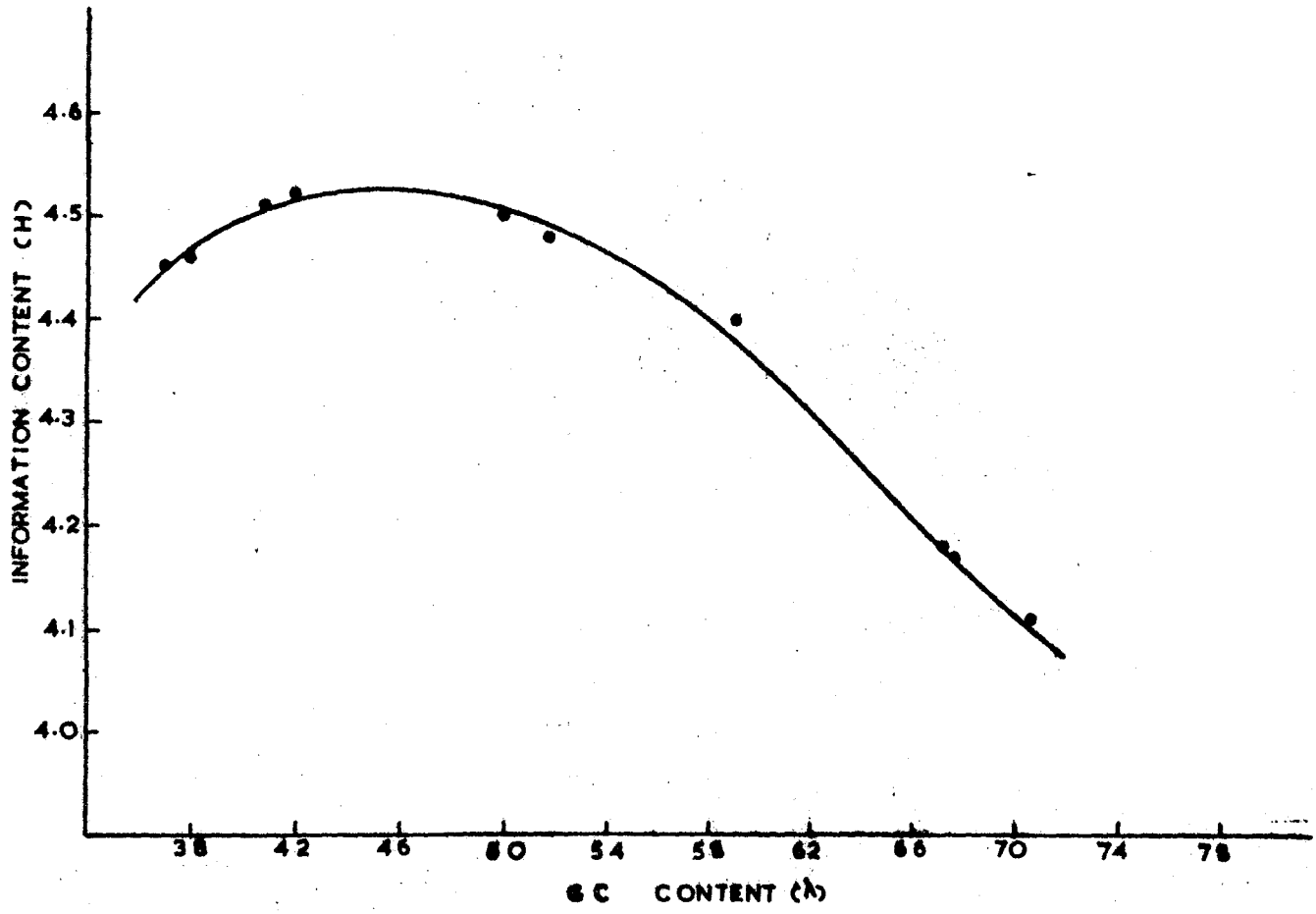


Fig 1. RANDOM MODEL.
The curve shows the variation of H with λ, as calculated for all the 64 Codons.
(Following Smith, 1969).

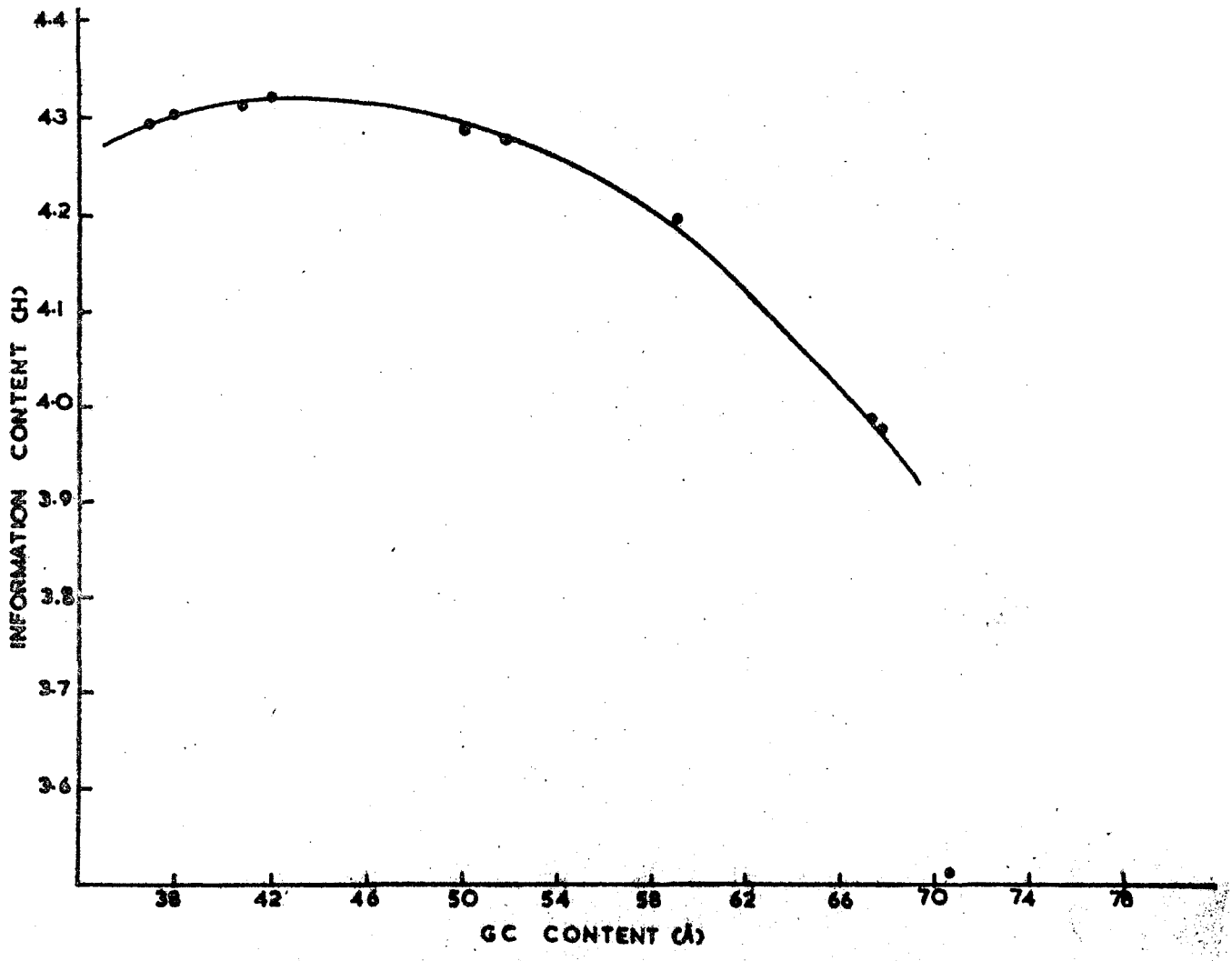


Fig. 2. RANDOM MODEL.
The curve shows the variation of H with λ, calculated only for the sense (61) codons.
(Following Smith, 1969)

This data on the base composition of the various DNA molecules are the results of chemical determination (Wyatt - 1951). They have also been calculated by using the doublet frequencies (Josse et al., 1961) and from buoyant density determination (Geel, 1967).

A graph is then plotted between the G+C (λ) content and the Information Content (H). Table 1 represents the Information Content calculated for 64 codons and 61 codons respectively. Fig. 1 and Fig. 2 represents the variation of the Information Content with the base percentage content for the two cases respectively.

Codon Frequencies

Codon frequencies in DNA of various organisms (the species mentioned above) are estimated by Geel et al (1972). They use the data on the dinucleotide frequencies, the isostich frequencies and the average amino acid compositions.

We are not discussing the method by which the amino acid frequency data has been determined. However, it should be remembered that if all DNA is coding for protein, then the frequencies of the sums of codons of each amino acid are the same as the average frequencies of the corresponding amino acids in proteins. This average is calculated by giving equal weight to each species of protein, since each species and not the amount of a given protein made corresponds to a cistron in DNA.

TABLE - 2.

A Comparison of the Codon Frequencies of the DNAs of Different Species.

Codon(s)	Human Spleen	Calf Thymus	Salmon	Yeast	B.cereus	E.coli	S.typhosa	S.marcescens	A.faecalis	Rh.spheroides	M.lysodeikticus
(CCC)	0.0003	0.0050	0.0080	0.0156	0.0092	0.0180	0.0071	0.0195	0.0170	0.0270	0.0203
(CCA)+(CCG)	0.0051	0.0358	0.0231	0.0142	0.0001	0.0095	0.0253	0.0219	0.0267	0.0167	0.0240
(ACT)+(CCT)	0.0611	0.0881	0.0659	0.0603	0.0516	0.0339	0.0384	0.0156	0.0255	0.0003	0.0
(ACC)+(GCC)	0.0299	0.0174	0.0150	0.0419	0.0384	0.0404	0.0371	0.0615	0.0933	0.0933	0.1115
(TCA)+(CCA)+(ACA)+(GCA)	0.0815	0.1045	0.0991	0.1203	0.0583	0.0467	0.0352	0.0353	0.0008	0.0020	0.0008
(TCG)+(CCG)+(ACG)+(GCG)	0.0001	0.0	0.0	0.0244	0.0011	0.0538	0.0607	0.0959	0.1231	0.1310	0.1351
(TGT)+(TGC)	0.0230	0.0230	0.0230	0.0080	0.0188	0.0190	0.0190	0.0191	0.0193	0.0193	0.0193
(TGA)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(TGG)	0.0153	0.0153	0.0153	0.0160	0.0148	0.0150	0.0150	0.0151	0.0152	0.0152	0.0153
(CGT)+(CGC)	0.0	0.0012	0.0001	0.0021	0.0363	0.0452	0.0406	0.0490	0.0600	0.0542	0.0497
(CGA)	0.0001	0.0002	0.0006	0.0144	0.0001	0.0	0.0086	0.0045	0.0016	0.0006	0.0032
(CGG)	0.0	0.0001	0.0	0.0044	0.0005	0.0	0.0003	0.0001	0.0035	0.0109	0.0149
(AGT)+(AGC)	0.0290	0.0149	0.0254	0.0058	0.0043	0.0321	0.0397	0.0157	0.0104	0.0198	0.0057
(AGA)	0.0330	0.0141	0.0174	0.0001	0.0	0.0036	0.0009	0.0038	0.0001	0.0	0.0007
(AGG)	0.0033	0.0257	0.0238	0.0	0.0001	0.0002	0.0002	0.0	0.0	0.0	0.0001
(GGT)+(GGC)	0.0260	0.0335	0.0511	0.0604	0.0496	0.0670	0.0615	0.0720	0.0573	0.0679	0.0493
(GGA)	0.0130	0.0247	0.0179	0.0114	0.0158	0.0105	0.0097	0.0064	0.0149	0.0064	0.0012
(GGG)	0.0351	0.0166	0.0058	0.0	0.0062	0.0015	0.0083	0.0058	0.0168	0.0150	0.0406

TABLE-2 Cont'd

Codon(s)	Human Spleen	Calf Thymus	Salmon	Yeast	B.cereus	E.coli	S.typhosa	S.marcescens	A.faecalis	Rh.sphaeroides	M.lysodeikticus
(TAT)+(TAC)	0.0351	0.0345	0.0345	0.0279	0.0318	0.0260	0.0252	0.0219	0.0182	0.0179	0.0165
(TAA)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(TAG)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(TAC)+(CAC)+(AAC)+(GAC)	0.0334	0.0510	0.0555	0.0	0.0101	0.0492	0.0542	0.0803	0.1031	0.1024	0.1188
(TAT)+(CAT)+(AAT)+(GAT)	0.1334	0.1147	0.1102	0.1416	0.1506	0.0988	0.0922	0.0588	0.0278	0.0279	0.0085
(CAT)+(CAC)	0.0223	0.0222	0.0222	0.0229	0.0201	0.0190	0.0189	0.0182	0.0175	0.0175	0.0172
(CAA)	0.0407	0.0150	0.0141	0.0208	0.0344	0.0250	0.0165	0.0063	0.0006	0.0045	0.0
(CAG)	0.0006	0.0192	0.0272	0.0	0.0	0.0280	0.0366	0.0471	0.0531	0.0493	0.0539
(AAT)+(AAC)	0.0534	0.0535	0.0535	0.0449	0.0505	0.0510	0.0511	0.0514	0.0517	0.0517	0.0519
(AAA)	0.0485	0.0556	0.0275	0.0683	0.0863	0.0403	0.0295	0.0162	0.0028	0.0001	0.0001
(AAG)	0.0129	0.0049	0.0330	0.0164	0.0009	0.0117	0.0211	0.0286	0.0353	0.0376	0.0346
(GAT)+(GAC)	0.0560	0.0555	0.0555	0.0459	0.0583	0.0520	0.0512	0.0476	0.0435	0.0432	0.0417
(GAA)	0.0198	0.0230	0.0303	0.0522	0.0808	0.0429	0.0337	0.0199	0.0249	0.0007	0.0
(GAG)	0.0345	0.0379	0.0235	0.0027	0.0122	0.0271	0.0355	0.0459	0.0370	0.0610	0.0602
(TGC)+(CGC)+(AGC)+(GGC)	0.0114	0.0114	0.0	0.0	0.0325	0.0921	0.1006	0.1106	0.1034	0.1084	0.0974
(TGT)+(CGT)+(AGT)+(GGT)	0.0666	0.0612	0.0996	0.0763	0.0765	0.0712	0.0602	0.0452	0.0436	0.0528	0.0266

TABLE - 2 Cont'd

Codon(s)	Human Spleen	Calf Thymus	Salmon	Yeast	B.cereus	E.coli	S.typhosa	S.marcescens	A.faecalis	Rh.sphaeroides	M.lysodeikticus
(TTT)	0.0437	0.0377	0.0371	0.0020	0.0152	0.0307	0.0352	0.0285	0.0001	0.0	0.0001
(TTC)	0.0	0.0025	0.0	0.0299	0.0246	0.0043	0.0260	0.0031	0.0284	0.0283	0.0270
(TTA)+(TTG)	0.0117	0.0066	0.0118	0.0497	0.0548	0.0294	0.0150	0.0247	0.0	0.0215	0.0002
(CTT)	0.0537	0.0494	0.0052	0.0091	0.0424	0.0197	0.0195	0.0115	0.0007	0.0007	0.0
(CTC)	0.0096	0.0229	0.0284	0.0225	0.0192	0.0225	0.0243	0.0238	0.0557	0.0492	0.0498
(CTA)+(CTG)	0.0124	0.0020	0.0355	0.0	0.0001	0.0144	0.0001	0.0261	0.0297	0.0148	0.0362
(ATT)+(GTT)	0.0425	0.0481	0.0506	0.0545	0.0484	0.0179	0.0173	0.0068	0.0	0.0	0.0004
(ATC)+(GTC)	0.0002	0.0078	0.0139	0.0002	0.0141	0.0363	0.0348	0.0619	0.0250	0.0542	0.0623
(TTA)+(CTA)+(ATA)+(GTA)	0.0342	0.0033	0.0429	0.1315	0.0968	0.0503	0.0354	0.0224	0.0210	0.0062	0.0095
(TTG)+(CTG)+(GTG)	0.0628	0.0640	0.0545	0.0001	0.0368	0.0703	0.0572	0.0835	0.1009	0.0926	0.0784
(ATC)	0.0158	0.0155	0.0155	0.0160	0.0239	0.0210	0.0206	0.0190	0.0171	0.0170	0.0163
(TCT)	0.0172	0.0202	0.0192	0.0306	0.0	0.0054	0.0002	0.0171	0.0	0.0006	0.0006
(TCC)	0.0072	0.0137	0.0215	0.0063	0.0085	0.0002	0.0	0.0	0.0183	0.0048	0.0173
(TCA)+(TCG)	0.0217	0.0262	0.0089	0.0512	0.0130	0.0063	0.0041	0.0114	0.0156	0.0191	0.0208
(CCT)	0.0437	0.0086	0.0183	0.0220	0.0001	0.0115	0.0071	0.0001	0.0	0.0002	0.0004

The assumptions based on which the codon frequencies and H as a function of G + C content have been calculated is being summarised.

1. The genetic code is correct in its present form (Table 3, Chapter II).
2. The codons are distributed at random throughout the DNA molecule. This means that the frequency of two adjacent codons is the product of their individual frequencies.
3. The statistical distribution of codons is the same, both in the coding as well as in the non-coding parts of the DNA molecules.

The first assumption has been tested to be correct. Efforts to find non-randomness have failed (as the occurrence of an overlapping code has been ruled out - Weaver - 1963, Krzywicki et al. - 1966). This implies that the randomness of the nearest neighbour sequence of amino acids is true, at least for the work that we are doing.

The evidence for the third assumption is scarce.

Table 2 gives the codon frequencies of the species under consideration (Ceel et al. - 1972). The Information Content is calculated as mentioned earlier and is shown in Table 3 (Column 3). Fig. 3 represents the variation of the information content with the base percentage of the species when the codon frequencies of Ceel et al (1972) are used.

TABLE 3

INFORMATION CONTENT FOR

Species	G.C. Content percent	All the codon frequencies(49)	The unique codon frequencies (25)	Codons of linear combinations taken as equiprobable contributors (49)
(1)	(2)	(3)	(4)	(5)
B. cereus	37.0	4.7427549	3.358605437	4.8758775
Yeast	38.0	4.645786213	3.742736092	4.8820781
Human Spleen	40.8	4.9170481	3.832071449	5.0393298
Salmon	42.0	4.9823085	4.116594955	5.16454503071
Calf Thymus I	42.0	4.9428833	4.035319406	5.057715774
Calf Thymus II	42.0	4.8575881	4.035319406	4.911700803
E. coli	50.0	4.9982989	3.872848859	4.9178637
S. typhosa	51.7	5.044808006	3.934222521	5.163922
S. marcescens	59.0	4.956251592	3.817583918	5.1056427
A. faecalis	67.2	4.7292723	3.563791839	4.8666423
Rh. spheroides	67.7	4.6850614	3.385606624	4.7803377
M. lysodeikticus	70.8	4.6057486	3.394741689	4.6939052

N.B. The graphs have been plotted taking into account two digits after the decimal point.

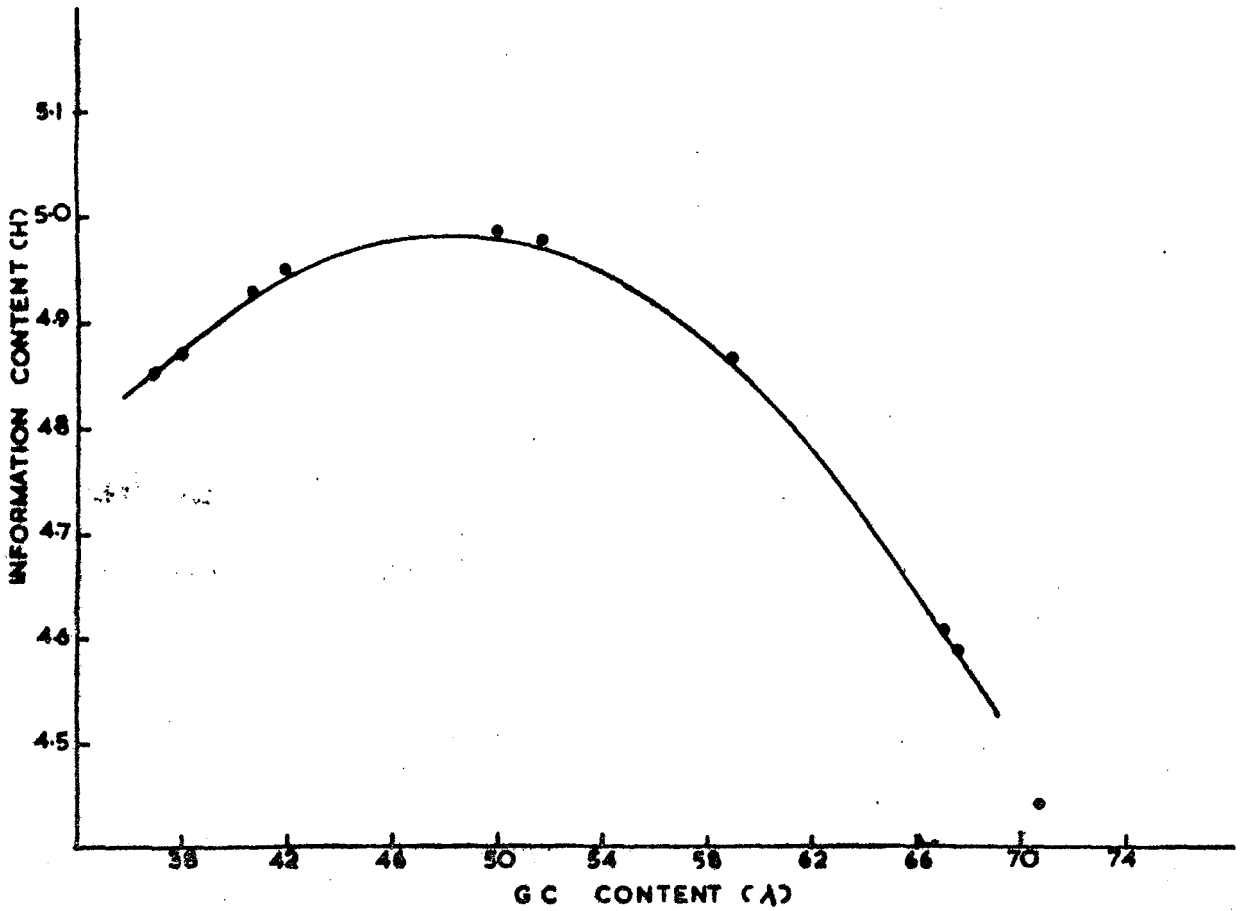


Fig. 3. The curve showing the variation of H with λ for the codons whose frequencies has been given by Goel et al.

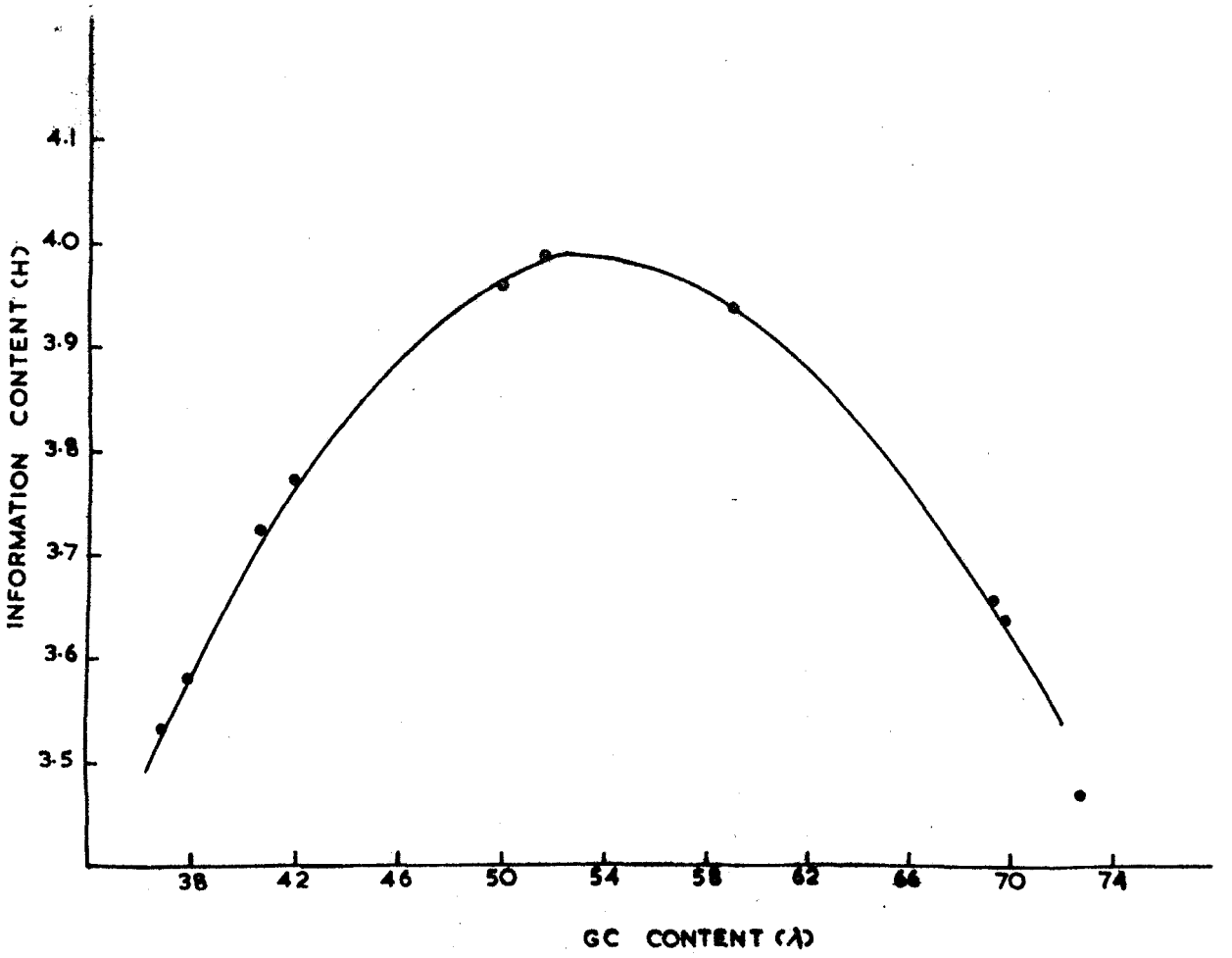


Fig 4. The curve showing variation of H with λ, when the codon frequencies are taken to be unique.

It is clear from Table 2, that some of the codon frequencies are given as linear combinations. We should remember that the biological code is degenerate and so the amino acid composition of proteins does not determine the codon frequencies uniquely. Twenty-five of the forty-nine codon frequencies are unique while the rest (twenty-four) have been given as linear combinations. This, probably introduces an error in the calculation of the information content. This also holds true in the light of the mathematical reality that

$$(a+b+c+d)\log(a+b+c+d) \neq a\log a + b\log b + c\log c + d\log d.$$

In order to eliminate this probable error we have next calculated the Information Content (H) for only the twenty-five unique codon frequencies. Table 3 (column 4) gives the measure of the Information Content in bits for the different organisms. Figure 4 represents the graphical representation.

Next, the codons of the linear combinations are assumed to contribute equally for the codon frequency. For example, for Human Spleen one of the linear combination is TTA + CTA + ATA + GTA = .0342,

If each of these four codons have contributed equally, then TTA = CTA = ATA = GTA = .0342/4 = .00855. In this way we have taken the individual codon of the linear combinations as equal contributors. It should also be

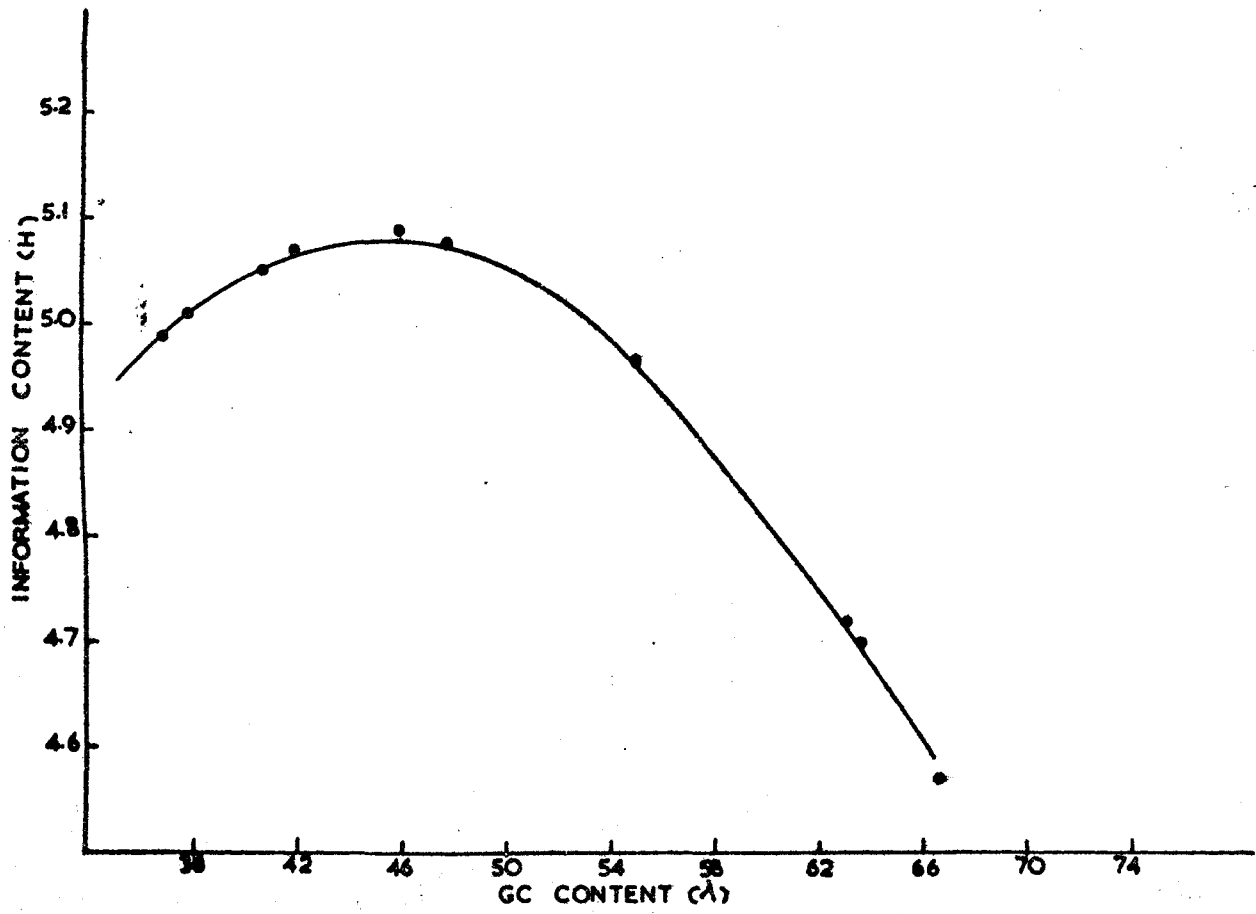


Fig 5. The Curve showing variation of H with λ, when each of the linear combi-nations is also considered as an equal Contributor (Equiprobable Codons).

noted that in twelve of the linear combinations, the difference is only in the first nucleotide place, while for the other twelve, it is in the third nucleotide position. We have also calculated H for the different organisms based on such type of data. Table 3 (column 5) gives the value of H for the different species and figure 5 the graphical representation.

Discussions

As far as Numerical Taxonomy is concerned, we did not do any classification based on either the cluster analysis or on the principle co-ordinate analysis. On the other hand, we have reproduced the results of G. Subba Rao (1971). The phylogenetic tree (Figure 5a, Chapter III) is based on the doublet frequencies and on the codon frequencies (Figure 5b, Chapter III) respectively. A study of the two trees reveal that B. cereus groups with the mammals rather than with the bacteria. Man and calf, and E. coli and S. typhosa have been grouped together. A comparative study of the two figures reveals that in the analysis based on codon frequencies, man and calf are closer to each other than to Salmon. In the classification based on doublet frequencies, all the three are equidistant. Thus we may say that the classifications based on codon frequencies is finer than that based on doublet frequencies to a certain extent.

She has used two more types of data. The first type of data are deviations of codon frequencies from the corresponding frequencies expected on the basis of the minimum amount of A, T, C and G required for the survival of the organism. The anomaly of *S. serevg* falling into the mammal group has been removed by cluster analysis and principle co-ordinate analysis. At the same time, some other undesirable features have cropped up. *E. coli* is shown to be closer to *S. marcescens* than to *S. typhosa*. *Rh. sphaeroides* is closer to *A. faecalis* than to *M. lysodeikticus*.

This means that the results of the classification depends on the type of data used to some extent. These results would have certainly improved if the sixty-four codon frequencies were unique.

The most important procedure in Numerical Taxonomy is the computation of the phenetic similarity. It begins with the collection of values for the characters of the taxonomic group which is to be classified. We could have ourselves constructed a tree following Margoliash and Fitch (1967) method of the minimum mutation distances, if the amino acid sequences of cytochrome c for these very species under consideration was known. This particular aspect is not being tackled at present.

Phylogenies and macromolecules provides an excellent account of the molecular aspects of evolution. They permit the establishment of phylogenies, which usually, but not

always, agree with the phylogenies based on morphology or other non molecular data. There are certain questions that still need elaborate answers. What are the reasons for the evolutionary change of macromolecules? If a molecule performs a highly specific function, then why does it change in the course of millions of years? Do these changes occur in spite of an opposing force of selection or are they the result of selection?

Large number of amino acid replacements accumulate in proteins during evolution. This becomes evident when homologous proteins in different species of living organisms are compared. This process results from point mutations in DNA.

The result of Margoliash and Fitch (1967) is based upon a quantitative determination of mutations. This suggests that ultimately this will be the most accurate method of constructing phylogenetic trees. These very same concepts (minimum mutation distance) have been used in a very natural way to calculate the Information Content of the biochemical function of a family of homologous proteins like haemoglobin or cytochrome c. (Yockey - 1977). The knowledge of Information Content for the species known may well contribute to a better understanding of the origin of life and of evolution.

In the Non-Darwinism model the amino acid composition should be strongly influenced by the Genetic Code, since

by hypothesis a significant proportion of the amino acids present have arisen by random mutation and drift.

In the Darwinian model, all evolutionary change depends upon natural selection. This predicts that there should be no relationship between amino acid frequency and the Genetic Code.

The average amino acid composition of proteins is determined by the Genetic Code. Amino acid composition is a product of molecular evolution. Proteins with novel functions evolve from pre-existing proteins by mutation. The novel proteins ultimately escape the control mechanism which determines the rate of production of their ancestors. The nucleotide sequence of mRNA is a code determining the amino acid sequences of proteins. It is also an established fact that the sequence of bases along the polymer varies from organism to organism (DNA source) and carries the hereditary information.

Coming back to the central idea of Information Content, we have been able to get results similar to those obtained by Smith (1969). The curves (Figure 1 and Figure 2) represent the information content (bits) against the GC contents for the sixty-four and sixty-one codons respectively. The maximum information content is at 42 per cent GC (Salmon) content. This is in accordance with Smith. It should be noted that our calculations are based on unconditional monomer frequencies (base - composition). The comparison

between the information content of the different organisms reveals the relationship between the information content and the taxonomic or evolutionary classifications. Closely following Smith, we can infer that evolution does code the more complex organisms more efficiently by working from the maximum in H (R, λ) i.e. the Random Model Case.

We have used triplet frequencies and comparisons at the phylogenetic level. The universal genetic code has a highly biased nature and so we can say that the DNA base sequences of higher organisms have great capacities to code various proteins. The GC composition of double stranded DNA is 42 per cent, the value around which higher organisms cluster. The base composition of vertebrate DNA is not exactly fixed. Figure 1 and Figure 2 which show a maximum for H at 42 per cent point out that different species but same GC content possesses same amount of Information Content. This means that vertebrates and mammals must be grouped together. This however is not true.

The graphs (Figure 3, Figure 4 and Figure 5) also represents the variation of H with λ, one for the codon frequencies calculated as such by Gool et al. (1972), the second for the twenty five unique codons and the third for the forty-one codons (linear combinations being treated as consisting of equal contributing codons). If we neglect S. limbosa, then in all the three cases, H maximises for Salmon (GC content = 42 per cent) as is evident from Table 3.

The graphs have however been plotted by a curve of the power series. A parabola of the second degree of $y = a + bx + cx^2$ is fitted to this data. This parabolic curve however indicates a maximum at 80 per cent for E. coli. This indicates that there is either some error in the experimentally arrived codon frequencies of Ceel et al. (1972) or we can make a very crude generalization. We can say that the information content tends to increase for all such organisms, in which the GC per cent tends to equal the AT per cent.

The Information Content of the twenty-five unique codons (Table 3, column 4) forms a major portion of the Information Content as calculated for the total codon frequencies (Table 3, column 3). This indicates that we are not committing a serious error in neglecting the linear combinations. This means that the linear combinations are perhaps not playing an important role in evolutionary process. The amino acids being coded by these twenty-five uniquely determined codons are pro, glu, lys, gln N, ser, met, leu, phe, gly, arg, tryp and term. Thus these code for ten of the twenty amino acids and the terminator.

There are at present only two experimentally known quantities relating to total base sequences.

(1) monomer frequencies (base composition)
(2) doublet frequencies (or nearest neighbour frequencies)

The first is an unconditional monomer probability. The second is an unconditional doublet probability, which has

been synthesized from the intersection of two monomer probabilities, one unconditional and the other conditional.

As has been stated by Smith earlier, we need to calculate higher order correlations for a more precise and accurate study of the subject. Till today, the main problem lies in obtaining a good mRNA statistical data - the needed test of the Information Content approach.

A scientific theory of evolution can only be a theory of descendance. Ideally, we should therefore be able to compare species along their direct derivation, which is impossible, as the species from which the present are derived have either disappeared or have themselves changed. At the same time common ancestory may not be the only possible basis for similarity in amino acid sequences. Quoting Margoliash and Fitch (1968) "Two proteins may be similar at a time, also, because having arisen from different ancestral origins they have tended to evolve to similar or identical functions in different lines of evolutionary descent and have therefore acquired the degree of similarity of structure required by this similarity of functions."

...

References

Fitch, W.M. and E. Margoliash, Science, 155, 279 (1967)

Goel, N.S., J. Theor. Biol., 16, 440 (1967).

Goel, N.S., C. Subba Rao, M. Yčas, Brenmerman, King, J. of Theor. Biol., 35, 399 (1972).

Josse, J., A.D. Kaiser and A. Kornberg, J. of Biol. Chem., 236, 864 (1961).

Krzywicki, A. and P.P. Glowiński, J. of Theor. Biol., 17, 136 (1967).

Shannon, C.E., in "Mathematical Theory of Communication", University of Illinois Press, Urbana, Illinois (1949).

Smith, T.F., Math. Bio Sciences, 4, 179 (1969).

Subba Rao, C., Doctoral Thesis, Rochester University, New York (1971).

Wiener, N., in "Cybernetics", John Wiley and Sons Inc., New York (rev. edn. 1961).

Yčas, M., in "The Biological Code", North Holland, New York (1969 rev. edn.)

Yockey, H.P., J. of Theor. Biol., 67, 345 (1977).

.....

Appendix I

Hardy-Weinberg Law

Let us suppose that there is a population composed entirely of the genotype AA and that the second population consists only of the genotype aa. Further suppose that the two populations are brought together and that the individuals interbreed at random. Finally it is assumed that there is no migration, no selection and no mutation. What will be the genotypes of the resulting population in the next and in all following generations? What will be the frequencies of the two genes A and a? These questions were answered by Hardy.

Hardy was able to show, with some simple algebra, that if the initial frequency of the AA genotype was p and if the initial frequency of aa was q and if these were the only two alleles involved, so that p+q = 1, then it would follow that the distribution of genotypes in the next and all subsequent generations will be p^2AA : 2pqAa : q^2 aa, when all the assumptions stated earlier hold true. This population has become known as the Hardy-Weinberg Law, and it is the foundation of population genetics.

Geneticists today state the Hardy-Weinberg formulation in a slightly different way. Hardy postulated that one population consisted only of the genotype AA. It follows also

that it consists only of the gene A. A similar argument can be stated for the second population; which consisted only of the genotype aa and so only of the gene a. The values p and q cannot be assigned to genotypes but to gene frequencies. This means that, in the first population, the frequency of the gene A = p, and a = q is the frequency of the allele in the second population. If the frequency of gene A = p, then the frequency of AA should be $p \times p = p^2$. In a similar manner, frequency of aa should be $q \times q = q^2$ and aA should be $2 \times p\ q$. The coefficient 2 is necessary because there are two ways in which aA can occur viz. a from one parent and A from the other parent or vice-versa.

*****