EVOLUTIONARY AND MOLECULAR TAXONOMY

E. H. Harley
Department of Chemical Pathology, University of Cape Town, South Africa.

Keywords: Evolution, taxonomy, genetics, systematics, DNA, genome, mitochondria, chloroplast, parsimony, maximum likelihood, neighbor joining.

Contents

Summary

Over the last two decades, the use of molecular methods, especially DNA sequence data, has had a profound influence on taxonomy. The information is derived from mutations, which are slowly but constantly accumulating in lineages of organisms over time. Since the way in which DNA sequences evolve is generally well understood, and since huge amounts of comparative information can be provided by DNA sequence data, these methods have had a major effect on accelerating taxonomic approaches away from a simply descriptive and towards a more evolutionary, or phylogenetic approach. The DNA sequences, which are commonly used, occur either in the nucleus of the cell, or in organelles, such as mitochondria or chloroplasts. Sequence data are being used at all hierarchical levels of phylogenetic analysis, from the species level to the relationship between bacterial and eukaryotic kingdoms, which is enabled by the very wide range of sequence conservation observed in different genes. There are a number of different

methods of phylogenetic analysis available, including distance measurements, parsimony, and maximum likelihood, with added statistics to provide measures of support for topological features. The stage is now being reached in which the increasingly large amounts of sequence information produced for some data sets, as well as the greater numbers of taxa often included in such sets, are taxing the abilities of computational methods to handle the data effectively. Nevertheless this is an example of an "*embaras de richesse*" and will be a spur to the development of progressively more powerful and sophisticated methods of analysis.

1. Introduction

Taxonomy is the biological discipline producing the classification of the diversity found in the biological domain. Its purpose is to classify the structure of the living world by constructing hierarchies consisting of progressively more general groupings of organisms. The criteria for inclusion in such groups originally depended mostly on descriptive morphology. More recently the evolutionary history, or phylogeny, of a group of organisms, has come to play a major role in identifying the relationships between organisms, and the reconstruction of this phylogenetic history is a main function of systematics, the study of biological diversity.

The accumulation, over the last few decades, of vast quantities of molecular data from an ever-increasing range of plant, animal, and bacterial species is having a huge impact on phylogenetic analysis. Molecular data come in various forms, but DNA sequence data now account for effectively all new information. Molecular data have significant differences from traditional morphological data. These not only create new opportunities for developing novel analytical approaches, but provide new problems as well, and both of these factors require adjustments to the previously established methods of phylogenetic analysis.

2. Basic Concepts

To understand these new approaches, it is necessary to have a basic grasp of the concepts underlying inheritance, genetics, and the way the molecules of inheritance change and evolve. The biochemical essence of a living cell, be it a bacterium, or a cell from a plant or a vertebrate animal, is an interconnecting set of biochemical pathways which constitute the metabolism of the organism.

2.1. Metabolism, Proteins, and Nucleic Acids

Metabolism converts food substances into useful cell constructs such as proteins, as well as supplying energy for maintaining the various biochemical processes of the living cell. These metabolic reactions are instigated, or catalyzed, by a vast array of enzymes - protein molecules designed to facilitate specific metabolic reactions. The information for the construction of these enzyme proteins, as well as information for higher level functions such as embryonic development, is stored in nucleic acids which, except in the case of a few viruses, consists of deoxyribonucleic acid (DNA). The structure of this most fundamental of all biological molecules was discovered less than half a century ago by Watson and Crick. It has two main attributes: replication, which is the ability to be copied so that the complete set of inheritance information is passed on to each of two progeny cells, and the ability to direct the sequence of amino acids in proteins. Proteins are made up of building blocks of some 20 amino acids, whereas DNA consists of a long chain built up from only four biochemical units referred to variously as nucleotides, bases or residues. These bases are adenine, guanine, thymine, and cytosine, commonly abbreviated to A, G, T, and C respectively. The DNA normally exists as a double helix, with the two chains running in opposite directions relative to each other, with every A in one chain paired with a T in the other, and every G paired with a C. More generally the purine bases A and G, which are double ringed structures, pair with their corresponding pyrimidine (single ring) structures, T and C respectively. Sets of three bases, termed codons, code for the various amino acids, for example CCA codes for the amino acid proline, AUG for methionine, etc. The complete set of these codons, which contain the information necessary to make a protein, together with some information in the DNA sequence, which controls when and how much of the protein should be made, constitutes a gene. The total amount of DNA possessed by an organism constitutes the genome of that organism and is clearly very large: 3 billion bases (or more correctly base-pairs, since each DNA molecules is a double helix consisting of two strands wound round each other) for the human genome, and underscores the magnitude of the recent achievement of the complete sequencing of the human genome, whereby the order of bases in the DNA is now fully defined. Knowledge of the structure of the human genome is only half the story however: interpreting how it functions at many different levels will take a great deal of time longer.

2.2. Nuclear and Organellar Genomes

In animals and plants, the DNA resides in the nucleus of the cells (the nuclear genome). Very important for taxonomic purposes however, is another genome found in the cells of virtually all animals, the mitochondrial genome. Mitochondria are small organelles specialized to undertake several crucial biochemical processes in animal (and plant) cells, in particular the utilization of atmospheric oxygen to supply energy. They have their origin in bacteria, which were free-living some billion or so years ago, until captured by the ancestor of all living animals and plants. Since then they have lost, or transferred to the nuclear genome, the vast majority of their DNA, and remain in animals as only a small circular genome of some 16 000 base pairs. Plant cells have yet another, third genome, in the chloroplasts, which is a feature of all green plants. This also originated from a micro-organism, in this case a blue-green alga, whose ability to photosynthesize has now been co-opted by its host to perform this most fundamental of metabolic functions in the plant world.

The importance of these two organellar genomes for molecular taxonomy is out of all proportion to their small size, and is attributable to a number of factors, which include ease of isolation in the laboratory, their rate of evolution relative to the nuclear genome, and a different mode of inheritance.

2.3. Inheritance of Genomes

The DNA in the nuclear genome is packaged in discrete units called chromosomes. The human genome, for example, has 23 pairs of these, with one randomly chosen member of each pair being passed on to an offspring where they associate with the 23 chromosomes from the other parent to reconstitute the full genome complement of 23 pairs of chromosomes. This bi-parental inheritance does not apply to organellar genomes. The mitochondrial genome is passed on to the next generation via the female, with the sperm passing little or no mitochondrial DNA to the next generation. Similarly the chloroplast genome in plants is inherited uni-parentally, in flowering plants, for example, passing to the next generation only through the seed parent. These different modes of inheritance have significant implications for molecular taxonomy (see below).

2.4. Mutation and Molecular Evolution

Despite some highly sophisticated biochemical mechanisms to minimize errors when DNA is being copied for transmission of copies to the next generation of cells or individuals, such replication errors do occur. It has been estimated that this error, or mutation, rate in the human nuclear genome is about one mistake for every ten billion (10^{10}) nucleotide bases copied. The most common form of mutation (point mutation), is the replacement of one type of base with another, for example, the replacement of an A with a G. If this mutation occurs in a gene coding for a protein it may result, by changing a codon, in a different amino acid being exchanged for the usual one at that position in the protein. If this occurs, it is frequently the case that the enzymatic or structural function of that protein is impaired, in which case, the organism inheriting this change is likely to be affected in some deleterious manner. Inherited disorders in humans, such as haemophilia, albinism, or cystic fibrosis arise in this fashion. Since organisms possessing such deleterious mutations tend to be less fit than others in their population they, or their progeny, are less likely to survive and the mutation is in this way removed from the population as a whole. However, many mutations either have no, or only a very minor, effect on fitness. These are termed neutral (or nearly neutral) mutations and so can persist for long periods in populations, and such a mutation may even, through the random shuffling of genes at each generation, increase in frequency to the extent that it becomes fixed, meaning that it completely replaces the original (ancestral) gene in the population. Less frequently, the mutation may actually confer some advantage on the individuals possessing it, in which case selection will operate to accelerate its sweep to fixation in the population. Neutral changes are surprisingly frequent in a normal population: for example it has been estimated that one in about every 200 base pairs is different between any two human individuals. The measurement of the amount of such variation in populations and the difference in frequency of such differences between populations are of great value in the field of population genetics. In molecular systematics they are also of great value, but systematics operates at a different level from that of population genetics. It examines groups of taxa which are generally no longer interbreeding, and compares mutations which in some cases represent the ancestral form and in others the mutated, or derived, form which has become fixed in a lineage.

An example of such a situation is shown in Figure.1. Here it is assumed that an ancestral taxon possesses a C at some position in its genome. Over the course of time, perhaps some millions of years, populations of this ancestral taxon diverge and differentiate. The result is a number of populations which do not interbreed, due either to spacial separation, or a failure to recognize the other population as suitable for mating, or are unable to produce fertile offspring if mating does occur. These populations can then be looked upon as separate taxa, perhaps even separate species (the precise definition of "species" is a complex issue, see *The Species Concept, and Infraspecific Taxa*). During the evolution of one lineage, a mutation occurred in an individual, which changed the state of the nucleotide at that position from a C to a T (short, bold-faced arrow in Figure 1). This T became fixed in that lineage over the course of time, either by chance if it was neutral in its effect, or perhaps by selection if it caused some advantage to those individuals that possessed it. This change was then inherited by two taxa that descended, after a subsequent period of time, from the population with this fixed mutation.

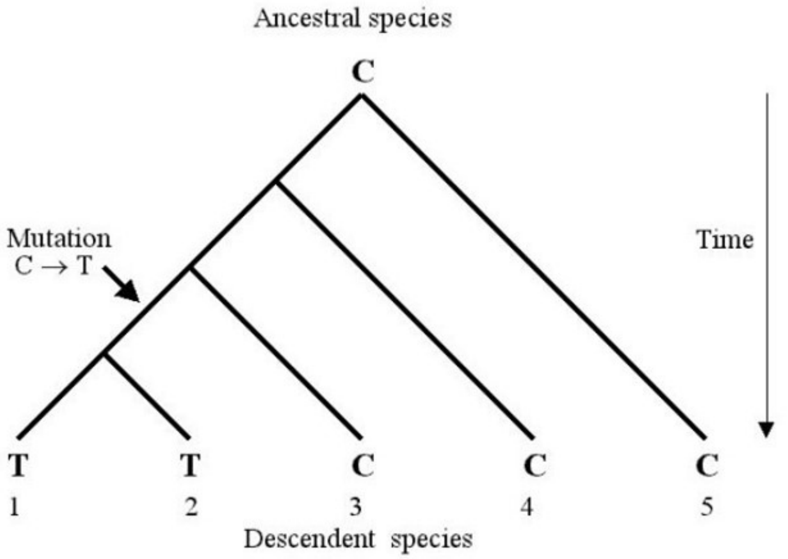

Figure 1. The evolutionary distribution of a single point mutation.

As can be appreciated from the evolutionary scenario or phylogeny, depicted in the form of a dendrogram or tree in Figure.1, two processes operate which give rise to measurable effects in the taxa constituting such a phylogeny: taxa which have a longer period of evolutionary separation, such as the two taxa on the rightmost part of the tree, will tend to have more positions mutated and show more changes relative to each other than those with shorter periods of evolutionary time separating them. Secondly, taxa which share the same character state (in this case either C or T) are likely to share either the same remote, ancestral status or, when they share the mutation occurring in a recent common ancestor, the same close sibling status. It is the measurement by an appropriate method of large numbers of such molecular changes in a group of taxa that constitutes the data set for subsequent construction of a molecular taxonomy.

Some regions of the nuclear genome evolve, or accumulate mutations, at a faster rate than others. These regions often separate those parts of the DNA that code for proteins. They can therefore be the favored regions for studies on a closely related set of taxa, because they accumulate new information relatively quickly. Genes coding for very conserved proteins, or for ribosomal RNA, do not easily tolerate mutations, and so change relatively slowly. These will be the genes of choice for the study of more distantly related taxonomic groups, such as a study of the relationship of animal phyla.

It might be thought that the more change that can be measured, the more useful it would be in a phylogenetic study, but this is not always the case. Since in DNA there are only four different states at a sequence position, A, G, C, or T, there is a significant possibility of parallel mutations occurring, in which the same change (e.g., a C to a T) occurs at that sequence position in two separate lineages, so that although these two taxa share the same character state, this has not been derived from a single mutational event. Back mutations can also occur, where a base mutates to a different state, but then some time later mutates back to its original state. Parallel and back mutations, illustrated in Figures 2a and 2b respectively, can contribute misinformation (homoplasy) in a data set, which increases in proportion to the total amount of mutation that has taken place. Such homoplasy can be a major problem for phylogenetic analysis in very divergent sets of taxa, or when using areas of DNA that have especially rapid rates of evolution.

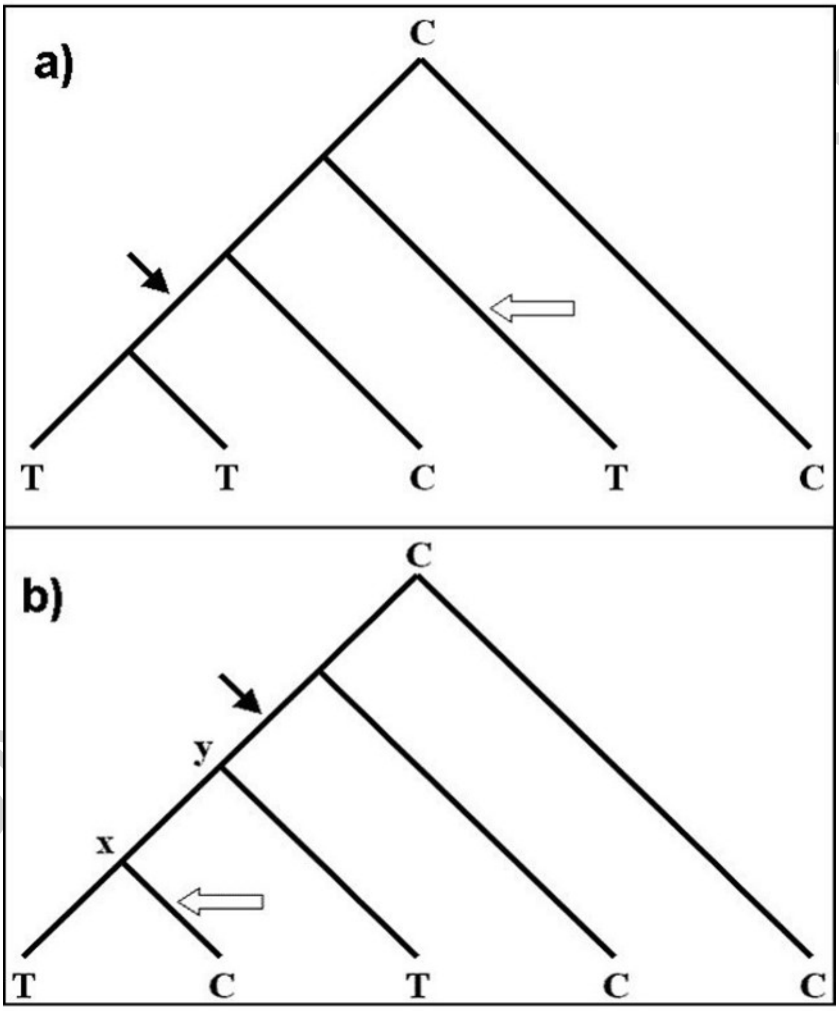

Figure 2. Examples of parallel and back mutations: a) Example of a parallel mutation. The same example as in Figure 1, but there is an additional parallel mutation depicted by the open arrow; b) Example of a back mutation. A similar example to that in Figure 1, although the mutation of the ancestral C to a T takes place at an earlier stage. There is an additional mutation back to the ancestral state depicted by the open arrow.

A final factor, relevant to the use of these mutations in phylogenetic reconstructions, is that bases do not change with equal frequency to any of the other three bases. A purine base (A or G) is more likely to mutate to the other purine base than to a pyrimidine base, and the same applies to the pyrimidine bases C or T. These more probable events are termed transitions, and the rarer purine to pyrimidine inter-conversions are termed transversions.

Mutations other than point mutations can be used by some of the methods of phylogenetic analysis. Deletions or insertions (indels) of one or more bases can occur, and since these are less likely than single base substitutions to occur in identical fashion twice, they are less subject to homoplasy. Major rearrangements of genes, especially in organelle DNA, can provide useful characters for higher level taxonomy, for example between classes or orders of vertebrates or mollusks.

3. Molecular Methods used in Systematics

3.1. Allozymes

The use of DNA itself to provide molecular data for taxonomic purposes is a relatively recent event. Molecular approaches were first introduced in the 1960s in the form of protein electrophoresis. It was found that the protein constituents of serum or tissue extracts could be separated under the influence of electrical fields since different proteins contain different proportions of charged amino-acids. However, this would by itself simply provide a smear consisting of thousands of different proteins indistinguishable from each other in the crush. The advent of special enzymatic or immunological staining methods to allow visualization of single protein species provided a breakthrough. It showed that even in individuals of the same species, proteins were sometimes seen to consist of two, or occasionally more, electrophoretically separable and distinguishable forms. These forms, or allozymes, were the consequence of a mutation in the ancestral gene coding for the protein at some time in the past such that both the ancestral form and the more recent mutated form were now both present in the same species. This provided the first molecular measure of within and between species diversity, and is still a widely used technique. It is however, generally of more use in the field of population genetics than in taxonomy or phylogenetics, which is mostly concerned with taxonomic diversity above the species level, (see *Populations, Species and Communities*).

3.2. DNA/DNA Hybridization

One of the earliest methods of inferring relative amounts of genetic divergence between species exploited the fact that the two strands, which comprise the DNA helix can be separated from each other by heating. On subsequent cooling, the strands, given time, find their respective counterparts and recombine or anneal. The temperature at which this annealing process takes place decreases if the DNA strands are not accurately matched, as will be the case in two different species, where the one DNA differs from the other by the various mutations which will have accumulated in both lineages since their separation from their common ancestor. The more distantly related any two taxa are, the lower will be the annealing temperature of a mixture of DNA, from both species, compared with that of either DNA preparation annealed to itself. A table can then be constructed consisting of the amount by which the annealing temperature is lowered in the mixture, compared to the mean of the two unmixed DNAs for all pairs of taxa in the analysis. These values constitute a measure of genetic distance and can be used to construct phylogenetic trees. Although the method has the advantage that it averages over the whole nuclear genome, it is technically difficult to apply and to obtain the high level of accuracy required to differentiate correctly between taxa with different times of divergence from a common ancestor. In addition, it suffers from a number of criticisms common to the use of measures of genetic distance for phylogenetic reconstruction (see below). As a consequence it is little used now.

3.3. Restriction Fragment Length Polymorphisms (RFLP)

Restriction enzymes are enzymes found in many bacteria, which have the property of cutting DNA very precisely at specific positions along the chain of nucleotide bases. For example the restriction enzyme Eco R1 (so named because it was isolated from the bacterium *Escherichia coli*) cuts wherever it finds the sequence ..GAATTC.. in a DNA molecule. This results in the production of a set of fragments in a DNA molecule with sizes unique to that taxon and which can easily be separated and visualized by gel electrophoretic techniques. A molecule the size of mitochondrial DNA will produce on average about 4 fragments with Eco R1, and rather more with restriction enzymes such as Hae III which recognizes a four base pair sequence (GGCC in the latter case) and so cut more frequently. The fragments produced by different taxa provide sets of characters that have been used for phylogenetic analysis, but because the fragments are not strictly independent characters (a new site created by a mutation will convert one fragment into two) this creates problems for phylogenetic analysis. In practice, the method has some value at the population genetic level, but is of little use for systematics at the species level or above.

-
-
-

Bibliography

Felsenstein J. (2000) PHYLIP Version 3.6 (the PHYLogeny Inference Package) [A package of programs for inferring phylogenies, available from http://evolution.genetics.washington.edu/ phylip.html].

Hillis D.M., Moritz C. and Mable B.K. (1996). *Molecular Systematics*, Sunderland: Sinauer Associates, Inc. [This book provides a comprehensive discussion of all aspects of the subject].

Kimura M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press. [A thought-provoking early work on molecular evolution, emphasizing the major role of selectively neutral mutations].

Nei M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press. [This book provides a good comparison of the various types of distance methods used in molecular taxonomy].

Sokal R.R. and Rohlf F.J. (1981). *Biometry*, 3rd Ed., 881 pp. W.H.Freeman & Co. New York. [This book has a comprehensive account of many of the statistical tests used in systematics, including permutations, g_1, bootstrapping etc.].

Swofford, D.L. (2000). PAUP4*: phylogenetic analysis using parsimony Version 4.1. Sinauer Associates Inc http://www.lms.si.edu/PAUP/. [A set of computer programs for implementing distance, parsimony, or maximum likelihood methods, for molecular data].

Biographical Sketch

Eric H. Harley graduated with medical degrees at Guy's Hospital, London, in 1963, thereafter graduating from his Ph.D., investigating fungal toxins, and an M.D. in molecular virology, both at University College Hospital, London. After a while at sea serving as a ship's surgeon, he took a post as a consultant chemical pathologist at the University of Cape Town, South Africa, where for many years his and his students' research interests centered on genetic and metabolic aspects of human and animal disease. Highlights of this period included defining the adaptive variation in humans of the control regions of the genome of a putative human cancer virus, BK, characterizing the inherited defect in the Dalmatian Coach Hound, and developing and exploiting novel methods for measuring metabolic co-operation between human cells in tissue cultures. Observing, as human populations soared and animal and plant populations correspondingly declined, that the latter were as much in need of consideration as humans, he turned his attention increasingly to the adaptation of these metabolic and genetic tools to addressing problems of conservation biology. Experience in human red cell metabolism has enabled him to define novel metabolic pathways in the red cells of the black rhinoceros, which suffers high mortality in captivity from haemolytic anemia. Experience in molecular biology has enabled extensive studies to be performed on the molecular systematics, and/or molecular population genetics of a wide range of animal and plant species, including elephant, rhinoceros, zebra, wild cats, tortoises, birds, amphibians, fish, grasses and orchids. Experience in cell culture has resulted in the establishment of one of the world's largest banks of cells cultured from African mammals, and an enjoyment of computer programming has led to the development of many programs to assist students in manipulating and analyzing molecular data. Hobbies include rock-climbing, programming, and orchid culture.