Classroom

In this section of *Resonance,* we invite readers to pose questions likely to be raised in a classroom situation. We may suggest strategies for dealing with them, or invite responses, or both. "Classroom" is equally a forum for raising broader issues and sharing personal experiences and viewpoints on matters related to teaching and learning science.

Tutorial on Phylogenetic Inference – 1

Felix Bast
Centre for Biosciences
Central University of Punjab
Bathinda 151 001, India
Email: felix.bast@gmail.com

Phylogenetic Inference is a statistical technique to trace the evolutionary legacy of a wide range of subjects including biological taxa (species), biomolecules, languages, ancient texts and so on. In Part 1 of the tutorial, we begin with an introduction to this field and discuss *phenetics* and *cladistics* — two major techniques used for phylogenetic inference. A number of fundamental concepts for understanding phylogenetic inference are introduced. In Part 2, we will learn models of molecular evolution and methods of phylogenetic inference, concluding with a worked-out example.

Phylogenetic Inference: A Primer

Theodosius Dobzhansky[1], one of the greatest twentieth century biologists, once famously said, "Nothing in biology makes sense except in the light of evolution". The theory of evolution lights up fields as vast as taxonomy and systematics (molecular systematics), biochemistry and biophysics (evolution of biomolecules), epidemiology and medicine (evolutionary medicine), behavior, bioinformatics and so on. For example, by phylogenetic reconstruction of HIV strains that evolve with each new infection leaving certain relics of the past, a dentist in Florida, USA, was

[1] See *Resonance*, Vol.5, No. 10, 2000.

Keywords
Phylogeny, cladistics, evolution, phenetics, synapomorphy.

found guilty of deliberately infecting his patients with the virus [1]. Phylogenetic analyses using 16S rRNA also helped to trace the geographical origin of unknown cadavers by making use of genotypes of the gut bacterium, *Helicobacter pylori*, isolated from intestinal mucosa of the dead bodies [2]. Research from my own group has revealed in 2015 that sporadic episodes of the 'blood rain' phenomenon, reported since time immemorial and which can even be found in Homer's *Iliad*, were due to the spores of the subaerial green microalgae, *Trentepohlia annulata*. Phylogenetic analyses using nuclear DNA internal transcribed spacer region revealed evolutionary affinity of this algal strain from South India to that from Central Europe with very low sequence divergence, which is suggestive of recent biological introduction. Given its unique aerial spore dispersal mechanism through rain, authors hypothesized that the spores might have got dispersed through clouds at an intercontinental scale [3].

The process of evolution, organic or not, can be computationally modelled as a tree-like illustration using PI.

Organic evolution also shares a number of attributes with other forms of evolution, such as the evolution of languages (linguistic evolution, which Darwin famously termed, "curious parallels"), computer viruses, personality, and so on. The study of the process of evolution essentially means to compare observable attributes of the subject that is being analyzed, such as species, languages, etc. Let us consider species. To study how species evolved from their last common ancestor, we start by comparing their observable attributes, such as anatomical features, DNA/amino acid (AA) sequences, etc. Phylogenetic Inference (PI), also called phylogenetics or phylogeny, deals with statistical modelling of the process of evolution.

The German entomologist, Willi Hennig, is often credited with laying the foundations of this discipline in 1950. Oftentimes, PI is carried out computationally, using appropriate algorithms that implement statistical tests for reconstructing the 'true' phylogeny (see 'species tree vs. gene tree' subsection in Part 2 of this article, to be published in a subsequent issue of *Resonance*), and this comes under the realm of 'computational phylogenetics'. Conclusion of such a phylogenetic reconstruction is often

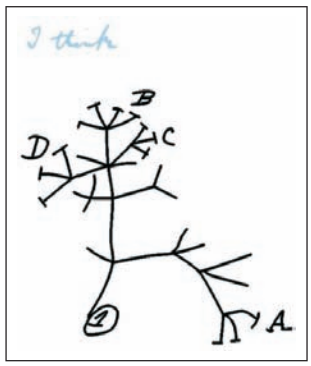

Figure 1. One of the earliest phylogenetic tree illustrations from Charles Darwin's *Notebook*. According to his description, the common ancestor for the entire tree is 1, and leaves *B* and *C* are more closely related to *D* rather than *A*.

Redrawn from archives of Natural History Museum, London.

illustrated as a phylogenetic tree, also called a phylogram [4, 5], the earliest depiction of which is the now the famous "I Think" tree illustration from Darwin's *Notebook*, presented in *Figure* 1.

In summary, the process of evolution, organic or not, can be computationally modelled as a tree-like illustration using PI. There are broadly two methods for this comparison: phenetics and cladistics.

Phenetics

Phenetics, (also called taximetrics), deals with clustering of subjects based on overall observable similarity. Results are often displayed in the form of a tree called a phenogram or a dendrogram. For example, for a set of species, gross morphology, or their DNA/AA sequences may be compared, without considering their evolutionary legacy. Carl Linnaeus was a pheneticist. His hierarchical classification of life was entirely based on phenetic clustering. For him, *relatedness* of species meant the propinquity of the Creator's design, rather than evolution. For example, an increasing number of papers on molecular phylogenetics attest the validity of Linnaean taxonomy for a number of taxa, although the theory of evolution was not known in his period. A phenetic method called Neighbor Joining (NJ) is frequently employed for approximating the phylogeny of complicated and computationally expensive datasets (see 'Distance-based Methods' section in Part 2 of this article). Phenetics is widely used to analyze DNA microarray data, as clustering genes based on their expression levels involves no *a priori* evolutionary reasoning. The American biostatistician Robert R Sokal is often credited for developing this discipline.

Cladistics explicitly deals with evolution as the subjects are portrayed onto an evolutionary tree (cladogram) based on 'shared conserved characteristics', or synapomorphy.

Cladistics

Cladistics explicitly deals with evolution as the subjects are portrayed onto an evolutionary tree (cladogram) based on 'shared conserved characteristics', or synapomorphy [6, 7]. Willi Hennig is credited with developing this discipline. One of the main principles of Darwin's theory of evolution is the 'principle of

common descent' which states that entire life on earth descended from one organism that appeared only once in the origin of life. That suggests that all life forms are indeed evolutionarily related in some way or other.

Types of Evolultion

As we know, evolution can broadly be of two types, convergent or divergent.

***Convergent Evolution*:** This is the evolution of similar features in unrelated lineages. For example, wings have evolved in insects and birds, which are functionally similar (for flight), but structurally different. Another example is wings of birds and bats; again functionally similar but structurally different. These characters are called *analogous* or *homoplasic*. In other words, homoplasic characters are shared by two unrelated groups, but not their common ancestor. Therefore, if we take these characters to analyze evolution, we are going to arrive at a false conclusion for sure, as illustrated in *Figure* 2. Clades (groups comprising of a node and its descendants) that are defined by homoplasic characters are called '*polyphyletic*', and include some descendants of different clades, with the exclusion of the last common ancestor. An example is the clade of warm-blooded animals; it includes birds and mammals, but the last common ancestor for them, a

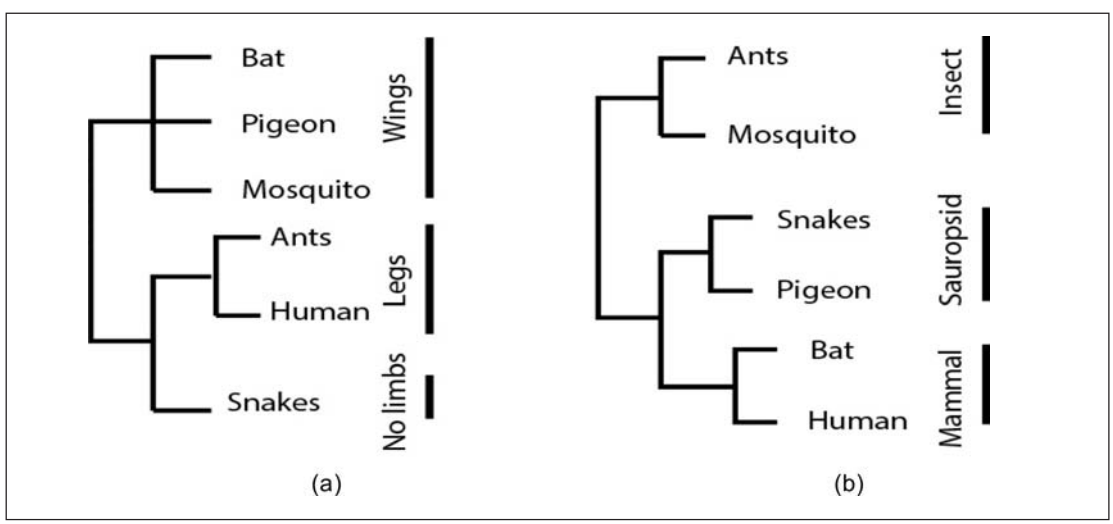

Figure 2. Fallacy of homoplasy – a dendrogram based on homoplasic characters, such as presence or absence of wings, legs, etc., can lead to polyphyletic clades and erroneous conclusions. **a.** Phenogram suffering from extreme homoplasy. **b.** Actual tree.

Divergent Evolution is the true evolution in which a species is split into two or more species and therefore any similarity is due to their shared ancestry.

reptile, was cold-blooded. Other examples are the clades of animals with wings, animals with legs and animals without legs in *Figure* 2a.

Homoplasic characters are indeed very problematic for phylogenetic inference, and should by all means be avoided.

***Divergent Evolution*:** True evolution is one in which a species is split into two or more species and therefore any similarity is due to their shared ancestry. Divergent evolution gives rise to *homologous* structures or homologous sequences. An example of a homologous character is the arrangement of bones in the arms of primates and wings of bats; anatomically they are very similar. Homologous characters are shared by related clades and their common ancestor and therefore, are essential tools for PI. Homologous characters are in turn grouped into two:

Synapomorphic characters are exclusively shared within a clade. For example, the presence of mammary glands, hair and middle ear bones are all synapomorphic characters for the clade mammalia -- these characters are shared only amongst mammals (*Figure* 3). This kind of specialized/derived/changed character state is called *apomorphic*. When this character evolves through multiple lineages and all the descendants of a common ancestor share it, it is referred to as *synapomorphy*. Synapomorphic characters are

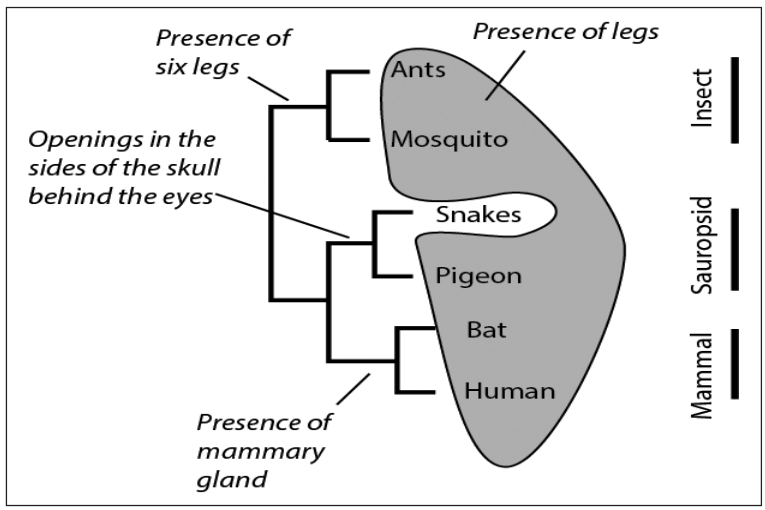

Figure 3. Synapomorphy vs. Symplesiomorphy. Shared derived characters highlighted on the nodes are synapomorphic, which correspond to true monophyletic groups indicated on the right along vertical lines. The character, 'presence of legs', is symplesiomorphic, as only some of the members (highlighted with gray background) share it. This is because the common ancestor to the two lower clades, mammal and sauropsid, a tetrapod vertebrate, had legs which were subsequently lost in some of its descendants, like snakes, as shown here. Groups like this are called paraphyletic. In strict sense, 'presence of legs' with respect to this entire group is homoplasic, as functional similarity between insect appendages and vertebrate limbs are resultant of convergent evolution.

precious tools for PI and identifying them is mostly the crucial part. Clades like mammalia that are defined by synapomorphic characters are called *monophyletic*.

Symplesiomorphic characters are those shared by the clade and some of its ancestors. A famous example is the presence of five toes in the hind limbs of apes and rats. This character is very primitive, originated in tetrapod vertebrates, and is shared by a number of other tetrapods. Therefore, if we take this character and group rats and apes together in one clade, the tree is going to be fallacious. Another example is the 'presence of legs' for tetrapods, as shown in *Figure* 3. This ancestral/original/primitive character state is referred as *plesiomorphic*, and when this character evolves through multiple lineages and some descendants share it, it is referred to as symplesiomorphy. Clades that are defined by symplesiomorphic characters are called '*paraphyletic*' and include descendants of a common ancestor with the exclusion of some. An example is the phenetic group of reptiles, as descendants of the last common ancestor of all reptiles also include birds and mammals.

In the context of DNA/AA sequences, homology is broadly classified into orthology and paralogy (*Figure* 4).

In the context of DNA/AA sequences, homology is broadly classified into orthology and paralogy.

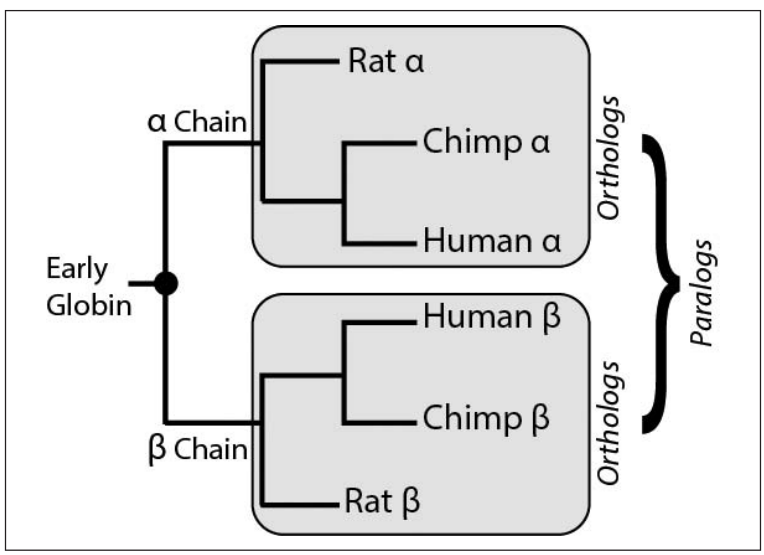

Figure 4. Homology of the globin protein family. Solid dot in the node near the root indicates gene duplication event, while other nodes represent speciation events.

Cladistics differentiates amongst the character states synapomorphy, symplesiomorphy and homoplasy, considering only synapomorphies while phenetics considers all, oftentimes arriving at fallacious conclusions.

Orthologous sequences are sequences (proteins or genes) separated by speciation events. Sequences need to be orthologous in order to be of use in molecular phylogenetics, as these may map speciation events. An example of orthologous sequences is the gene coding for Cytochrome C Oxidase (COX, see 'Loci used in Molecular Phylogenetics' in Part 2), or the gene coding for the α-chain of the protein globin (*Figure* 4). Identifying orthologs is crucial for developing an accurate phylogeny.

Paralogous sequences are sequences separated by gene duplication events. Example, genes coding for hemoglobin and myoglobin, or the α- and β- chains of the protein globin (Figure 4). Paralogous genes may be involved with 'gene birth', the origin of new genes.

In summary, the major difference between phenetics and cladistics is that the latter differentiates amongst the character states synapomorphy, symplesiomorphy and homoplasy, so that only synapomorphy is taken in to consideration while the former considers all, oftentimes arriving at fallacious conclusions. Conceptual relationships between various terms used in phylogenetic inference are illustrated in *Figure* 5 [8].

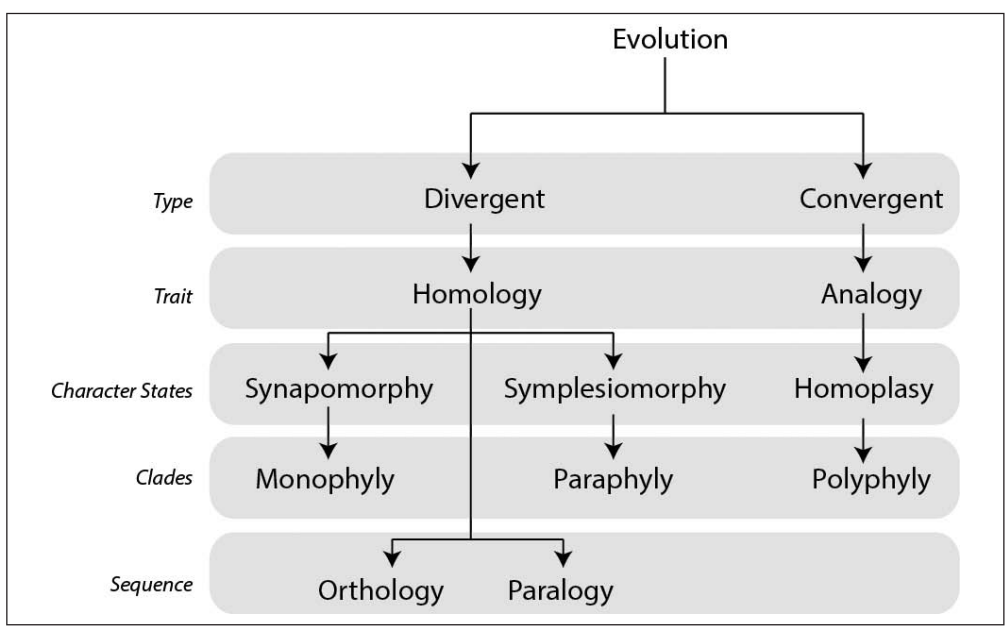

Figure 5. Conceptual relationships between various terms used in phylogenetic inference.

Acknowledgements

This work is supported by grant-in-aid from DST-INSPIRE Faculty Award (IFA11-LSPA02) and Major Research Grant from Indian Council for Social Science Research ('tracing linguistic evolutionary legacy of Indian languages using computational phylogenetics').

Readers are encouraged to take MOOCs on Molecular Evolution (MGE.512), Evo-Devo (BSS.513) and Computational Biology (BSS.512) offered through my website http://sg.sg/bastfelix. Suggestions are most welcome.

Suggested Reading

[1] **Ou C Y, Ciesielski C A, Myers G, Bandea C I, Luo C C, Korber B T, Mullins J I, Schochetman G, Berkelman R L, Economou AN, Molecular epidemiology of HIV transmission in a dental practice, *Science*, Vol.256, No.5060, pp.1165–1171, 1992.**

[2] **Nagasawa S, Motani–Saitoh H, Inoue H, Iwase H, Geographic diversity of *Helicobacter pylori* in cadavers: forensic estimation of geographical origin, *Forensic science international*, Vol.229, No.1, pp.7–12, 2013.**

[3] **F Bast, S Bhushan, A A John, J Achankunju, N M V Panikkar, C Hemetner and E Stocker–Wörgötte, European species of subaerial green alga *Trentepohlia annulata* (*Trentepohliales*, *Ulvophyceae*) caused blood rain in Kerala, India, *J. Phylogen. Evolution Biol.*, Vol. 3, No.1, 2015.**

[4] **S Pathak, A Akolkar and B S Mahajan, Onion plant as an educational tool for phylogenetic studies: molecular analysis and a new phylogeny?, *Resonance*, Vol.7, No.3, pp. 66–79, 2002.**

[5] **S L Baldauf, Phylogeny for the faint of heart: a tutorial, *Trends in Genetics*, Vol.19, No.6, pp. 345–351, 2003.**

[6] **B G Hall, *Phylogenetic Trees Made Easy: A How-to Manual*, Sunderland: Sinauer Associates, 2004.**

[7] **L Pietro and N Goldman, Models of molecular evolution and phylogeny, *Genome Research*, Vol.8, No.12, pp.1233–1244, 1998.**

[8] **F Bast. Online resources accompanying BSS.512 Bioinformatics and Computational Biology Course of Central University of Punjab. Accessible at http://bit.ly/BSS512**