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 – 2

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

Phylogenetic Inference (PI) 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. The first part[1] of this tutorial introduced a number of fundamental concepts including phenetics, cladistics, homology, homoplasy, synapomorphy, symplesiomorphy, orthology and paralogy. In this part, we will learn about models of molecular evolution, choosing the best model, overview of various genetic loci used in PI, methods of PI (including distance matrix method, NJ, and discrete data methods ML, MP and BI), issue of lineage sorting and conclude with a worked-out example.

Molecular Evolution

Evolution can be defined as a change in the allele frequency of a population over time. As we know, alleles are variants of the same gene. Allele frequency changes due to a number of processes, important among which are mutation, selection – either natural or artificial, gene flow and genetic drift.

Mutations of nucleotide bases can be of two types: synonymous (silent) or non-synonymous. Synonymous mutations show no

Keywords
Phylogeny, cladistics, evolution, phenetics, synapomorphy.

[1] *Resonance*, Vol.20, No.4, 2015.

Box 1. The Degeneracy of Codons

A DNA sequence, (more precisely, the mRNA sequence after its transcription), is ‘read’ by tRNA molecules in ribosomes in groups of three nucleotides, the triplet genetic code, known as a ‘codon.’ As there are four nucleotide bases (A, T, G and C), the total number of possible codons (permutations with repetition) are $4^3 = 64$. Most of the codons (except the start and the stop codons) code for an amino acid; sometimes more than one codon code for the same amino acid – a phenomenon called ‘degeneracy’. For example, CCA, CCT, CCG and CCC all code for the amino acid proline; here the 3rd position of the codon is said to be ‘four-fold degenerate’. Any point mutation at this site will be synonymous, as it conserves the amino acid. Point mutations in the 3rd codon rarely cause AA change (30%), while those in the 2nd position always and in the 1st position mostly (96%) changes the AA. The Indian origin biologist, Har Gobind Khorana (*Resonance*, Vol.17, No.12, 2012) received the Nobel Prize in 1968 for the elucidation of the genetic code using synthetic templates.

change in coded amino acids (AA) due to the degeneracy of the genetic code, whereas, non-synonymous mutations are those in which the coded amino acid changes, and that can have great ramifications on the functionality of the expressed protein (*Box* 1).

The majority of mutations occurring in nature are synonymous. Although rare, non-synonymous mutations are very important driving forces for creating variants in population for the natural selection to work on. The ratio of non-synonymous to synonymous mutation is a signature for the selection; if the value is more than 1 (i.e., if non-synonymous mutations > synonymous mutations), we call it positive selection. For neutral selection, the value is 1 and for purifying selection (the vast majority of cases) the ratio is less than 1.

Nucleotide bases can be either pyrimidines or purines (*Figure* 1). Pyrimidines are heterocyclic organic compounds with one ring structure. Thymine (T) and Cytosine (C) are pyrimidines. Purines are heterocyclic organic compounds with a pyrimidine ring fused to an imidazole ring. Adenine (A) and Guanine (G) are purines. Depending on the substituted base, mutations can be grouped as either transitions or transversions. In transitions, purines are replaced with purines, or, pyrimidines with pyrimidines. There are four possible transition errors: A to G or vice versa and C to T or vice versa. In transversions, purines are replaced with

The ratio of non-synonymous to synonymous mutation is a signature for the selection.

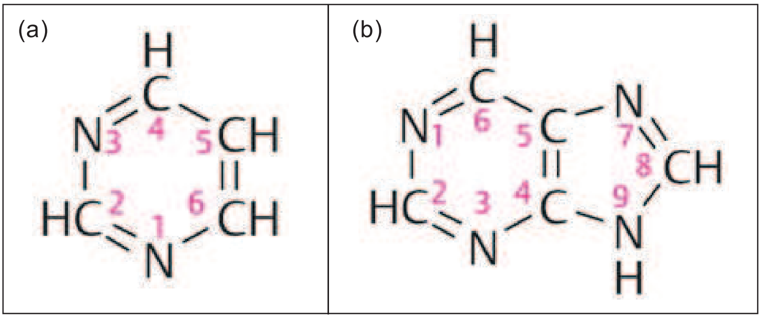
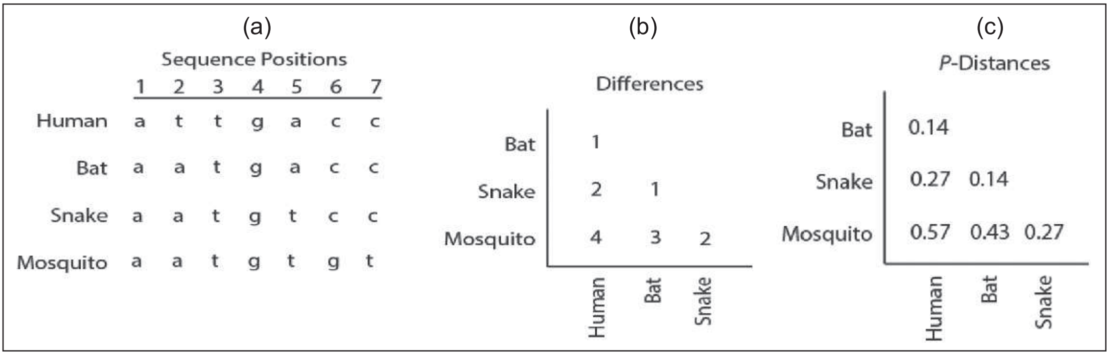


Figure 1. (a) Pyrimidine. **(b)** Purine.

pyrimidines or vice versa and there are 8 possible transversion errors: A↔C / A↔T / G↔C / G↔T. It would appear that transversions are two times (8/4) more likely to occur. However, transversions are seen very rarely in nature. This is because the molecular mechanisms by which mutations are generated, tautomeric shifts (amino form to imino form and vice versa), induce transitions far more spontaneously than transversions. In addition, transversions are likely to result in non-synonymous mutations. Those non-synonymous transversions are quickly removed from the population by the natural selection in the so-called 'purifying selection'. Therefore, transversions occur at lesser rates than transitions in nature. The majority of observed transversions render a new AA with similar chemical properties (for example, hydrophobic) so that the tertiary structure of proteins is not altered much.

PI, in essence, is an attempt to portray molecular evolution. Consider four orthologous sequences of seven nucleotides in length. *Figure* 2a shows a sequence alignment, in which

Figure 2. Calculating nucleotide *p*-distance from a sequence alignment.



(a)

Sequence Positions

	1	2	3	4	5	6	7
Human	a	t	t	g	a	c	c
Bat	a	a	t	g	a	c	c
Snake	a	a	t	g	t	c	c
Mosquito	a	a	t	g	t	g	t

(b)

Differences

	Human	Bat	Snake
Bat	1		
Snake	2	1	
Mosquito	4	3	2

(c)

P-Distances

	Human	Bat	Snake
Bat	0.14		
Snake	0.27	0.14	
Mosquito	0.57	0.43	0.27

nucleotides are arranged such that each position is aligned vertically with the other sequences. We now calculate the total number of nucleotide differences between each pair of taxa. Consider human and bat: these sequences differ by only one nucleotide at position 2, that is A↔T transversion, and therefore, the difference is 1, which we write in the matrix (*Figure* 2b). Consider human and mosquito: there are four differences, at positions 2, 5, 6 and 7. Do it for the rest of the pairs to complete the matrix. From this, we construct a new matrix (*Figure* 2c) in which each number is divided by the total number of nucleotide, i.e., 7. For example, difference between human and bat is 1, which is divided by 7, to get the corresponding value, which is 0.14. These values are now known as the **nucleotide *p*-distance**, the simplest representation of pair-wise genetic distances.

Nucleotide *p*-distance suffers from a severe shortcoming that it does not take into consideration the multiple substitutions at a single site, which is common in distant relationships and for rapidly evolving sites.

As we have just seen, in *p*-distance calculation, there are no differential treatments for transitions and transversions. In other words, all base changes are considered to have equal probability. Unfortunately, this method of calculation suffers from a severe shortcoming; it does not consider multiple substitutions at a single site, which is common in distant relationships and rapidly evolving sites. Also, we know that different base substitutions occur at different rates, transitions occuring more frequently than transversions. For example, consider two DNA sequences, X (AAGTCC) and Y (TTGTTT). For an evolution from X to Y, two transitions (last two positions) and two transversions (first two positions) are required. If we assume constant rate for transitions and transversions, our calculated time is for sure going to be shorter than the actual time, as transversions are rare events, and they occur much slowly. While calculating evolutionary distances, transversions need to be given more weightage due to their rarity. Therefore, *p*-distances need to be corrected to get a more accurate numerical approximation of molecular evolution.

Models of Molecular Evolution

There are a number of probabilistic models developed for converting *p*-distances to evolutionary distances. An in-depth discussion on various models is beyond the scope of this tutorial, so an overview is provided about some well-known models:

Jukes–Cantor 69: The Jukes–Cantor model is the simplest attempt to model mutations in DNA. It assumes that there is a single mutation rate for all bases (i.e., transitions and transversions are treated equally). It also assumes that all bases are present as 25% each, and that all sites mutate at equal rates. That is, neither base frequencies, nor base substitutions are allowed to change and therefore, the degree of freedom (*df*, number of parameters of the system that may vary independently) is 0. This model incorporates correction of *p*-distance for multiple hits using a Poisson distribution model.

Kimura-2-Parameter: Like the Jukes–Cantor model, this model also treats frequencies of nucleotides as constant; it assumes that all bases are present as 25% each. However, this model assumes that the rate of transitions per site (α) differs from the rate of transversions per site (β). Degree of freedom, $df = \alpha/\beta$ ratio, 1.

Felsenstein 81: In this model, the base substitution probability is constant, but frequency of each base is allowed to change. As there are four bases, and if the total is known, *df* is 3.

Hasegawa, Kishino and Yano 85: This model is a combination of K2P and F81; it allows different α/β ratio as well as different base frequencies. *df* is 4.

General Time Reversible (GTR): This model is similar to HKY 85, but base substitutions are not merely limited to transitions or transversion. Each of the six possible base substitutions can have its own probabilities. *df* is 8.

Each of these basic models can be modified further to have rate heterogeneity or rate invariability or both.

Rate Heterogeneity (G): This assumes that the rate of molecular evolution is different across sites. That is, some sites evolve at a faster rate, while others are slower. Rate heterogeneity is modeled by Gamma distribution.

Rate Heterogeneity assumes that the rate of molecular evolution is different across the sites; i.e., some sites evolve at a faster rate while other sites evolve slower. Rate heterogeneity is modeled by Gamma distribution.

Rate Invariability (I): This assumes that a certain fraction of the sites is evolutionarily invariable.

Choosing the Best Model

When confronted with model choice, complicated models like GTR+G+I might not be the most appropriate. When a simple model (e.g., K2P/F84) fits the data not significantly worse than a more complex model, the former should be preferred, as per the philosophical concept of parsimony[2]. Models are not chosen at random, but appropriate statistical 'Goodness of Fit' tests need to be performed for choosing the best fitting model. The Hierarchical Likelihood Ratio Test (hLRT) is a common statistical test for testing the goodness of fit of different models to the input data. This can be performed using the computational software MEGA, in which models with the lowest Bayesian Information Criterion (BIC)[3] scores should be chosen as the best model.

Genetic Loci Used in Phylogenetic Inference

In phylogenetic literature, the terms 'locus' and 'DNA barcode' are used interchangeably, and what we are referring to is the short stretch of genomic region that is amplified and sequenced to infer phylogeny of the target taxa. Of the many advantages of using DNA sequence data for constructing phylogenetic trees, one is the scope to select among conserved and more rapidly evolving sequence regions, the so called 'tortoise and hare' approach. Regions of the DNA have been differentially used to construct phylogeny at different hierarchical levels accordingly, i.e., more slowly evolving loci for analyzing higher taxonomic levels and more rapidly evolving loci for analyzing relationships between closely related species. Commonly used genetic loci for inferring phylogeny are presented in *Table* 1.

Methods of Phylogenetic Inference

Properties of molecular sequence data such as the availability of a large amount of characters and the recognizability of independent characters have encouraged utilization of statistical models to

[2] The principle of parsimony (also known as *lex parsimoniae* or Occam's Razor) states, "*Pluralitas non est ponenda sine necessitate*", i.e., "Plurality is not to be posited without necessity." According to this principle, among competing hypotheses that predict a phenomenon equally well, the one with the fewest assumptions should be selected. The analogy with a shaving razor is that it is better to have one simple but sharp linear edge, rather than making the edge complicated with multiple 'teeth', as in a saw.

[3] As per the principle of Occam's razor, Bayesian information criterion (BIC) selects the best-possible simple model by introducing a penalty for the number of parameters in the model. It is based, in part, on the likelihood function and it is closely related to the Akaike information Criterion (AIC).

Locus	Genome	Taxa used	Phylogenetic Resolution
Cytochrome C Oxidase Subunit 1 (*COX*1) gene	Mitochondria (Maxicircle)	Eukaryotes	At or above species level
Internal Transcribed Spacer region (ITS1-5.8S-ITS2)	Nuclear	Fungi, Algae and Plants	Below species level
18S rDNA	Nuclear	Eukaryotes	At or above species level
16S rDNA	Nucleoid	Prokaryotes	At or above species level
Transfer RNA L-F (TrnL-TrnF) Spacer	Choroplast	Plants	Below species level
Rubisco Large Subunit gene (rbcL)	Chloroplast	Plants	At or above species level

Table 1. Commonly used genetic loci in phylogenetic inference.

infer phylogenies. Inference methods will ideally extract maximum amount of information available in the dataset, combine this information with prior knowledge of patterns of sequence evolution and will deal with model parameters whose values are not known a priori. The representation of molecular hypotheses about the evolutionary ancestry of the sequences is achieved by phylogenetic trees (phylograms). Several statistical approaches exist to infer phylogeny from molecular sequences. In this section, we will cover the two broad classes of statistical approaches that have dominated in the literature: distance matrix methods and discrete data methods, although neither of them reproduces the evolutionary tree absolutely [1, 2].

Representation of molecular hypotheses about the evolutionary ancestry of the sequences is achieved by phylogenetic trees (phylograms).

Distance Matrix Methods

Distance matrix methods measure the genetic distances between the sequences from a set of sequences as a Multiple Sequence Alignment (MSA). Pair-wise evolutionary distances of the sequences in MSA are calculated and represented in a rooted or unrooted phylogram such that closely related sequences appear in the same interior node.

Advantages of the distance matrix methods lie in the analyses being fast and from their ability to model a substitution bias to correct multiple mutations. They produce only one tree – seemingly the best bet. However, this is done at a great cost to the phylogenetic accuracy.

The Neighbor-Joining (NJ) algorithm is one of the widely implemented distance matrix methods. Unlike other methods, NJ does not assume that the lineages evolve concurrently (molecular clock hypothesis) and therefore produces an unrooted tree. By including known taxa as outgroups, it is possible to root the NJ phylogram, and when done that way, it always produces an ultrametric tree (equal distance from root to the branch tips). An implication of such an equidistant ultrametric tree is that it employs a strict molecular clock assumption; i.e., mutations in DNA or protein molecules happen at constant rates across all branches, and therefore, the number of mutations (changes) is proportional to the time. However, real data do not strictly follow molecular clock; each branch can have different rates of evolution, and, therefore, representing them in an ultrametric tree is erroneous.

Discrete Data Methods

Two of the discrete data methods commonly used are Maximum Parsimony (MP) and Maximum Likelihood (ML). The MP, a relatively simple, non-parametric, statistical method, infers that the best representation of evolutionary relationships is the one that requires minimum number of steps (i.e., nucleic acid substitutions). The input data for MP analysis, known as ‘characters’, are phenotypical or genealogical attributes that are heritable and observed to vary between the taxa. MP produces a number of phylograms with considerable topological variations and therefore, an evaluation of all such phylograms is very complicated. A strict consensus phylogram is usually constructed by heuristic approaches that usually involve the steepest-descent

Maximum Parsimony infers that the best representation of evolutionary relationships is the one that requires minimum number of steps (i.e., nucleic acid substitutions).

style minimization mechanisms. The major disadvantage of this method is that the character states are generally noisy to an extent that the overly simplistic approach of MP results in erroneous conclusions. Its inability to apply nucleotide substitution models and the common notion of evolution being non-parsimonious, further limits MP's usefulness in PI.

Most of the modern phylogenetic analyses are based on ML, a parametric statistical method, which is reported to be more accurate (i.e., more likely to predict the evolution) and robust (less sensitive to faulty assumptions and models) than other PIs. ML criterion assesses probability of particular mutations by a substitution model and allows varying rates of evolution across both lineages and sites. Highly probable ML phylograms tend to have interior branches that require minimum number of mutations to construct, and vice versa. ML is a preferable method for phylogenetic analysis of distantly related sequences, although it demands greater computational capabilities.

Maximum Likelihood criterion assesses the probability of particular mutations by a substitution model and allows varying rates of evolution across both lineages and sites.

A much faster alternative that is often simultaneously performed with ML in the same data set is the Bayesian Inference (BI) – the name of which was derived from an 18th century statistician, Thomas Bayes[4]. The BI combines prior probabilities of a phylogeny with the likelihood of trees to produce posterior probability distribution on phylograms. Because tree topologies and branch lengths are not treated as parameters as in ML, but as random variables, it is impossible to obtain BI probabilities analytically. Therefore, BI probabilities are approximated by numerical simulations like the Markov Chain Monte Carlo (MCMC)[5] or Metropolis Coupled MCMC (MCMCMC). These chains explore the posterior probability grids in an integrative manner with model parameters. Trees are then sampled at fixed intervals and a consensus tree is constructed. The proportion of time that the chain visited sampled trees having a particular interior branch of the consensus tree is expressed as Bayesian Posterior Probabilities (PP). The computer program MrBayes is often used to estimate BI.

[4] *Resonance*, Vol.8, No.4, 2003.

[5] *Resonance*, Vol.8, Nos.4, 7, 10 and 12, 2003.

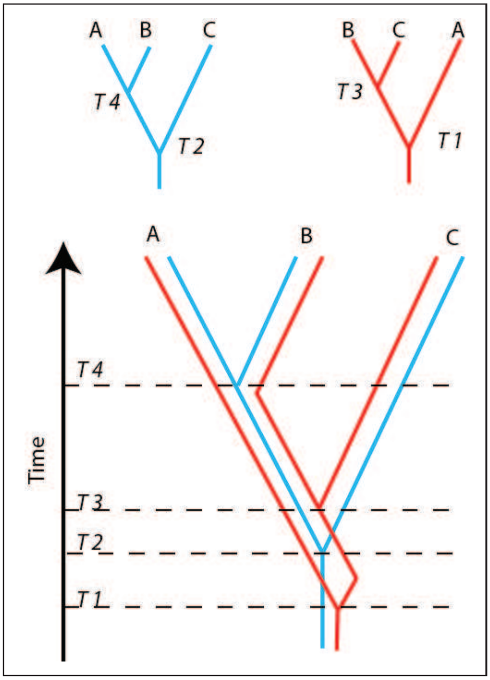


Figure 3. Lineage sorting. In the bottom image, the gene tree portrayed by blue gene (which is identical to the species tree) is super-imposed with the gene tree portrayed by red gene. Note time difference in splitting points (dashed lines). Topology of the blue and red gene trees are given above.

Gene Tree vs. Species Tree

In PI, we often begin with a set of synapomorphic morphological characters or orthologous sequences with an aim of reconstructing the evolutionary legacy through a phylogram. In practice, these phylograms portray the evolution of only the input character state, e.g., a set of genes. Trees generated by PI are often called 'gene trees' as they portray the evolution of a particular gene in question. Of course, gene trees are an attempt to reconstruct the actual or 'natural' tree, which is referred to as species tree. We do not know the species tree in most of the situations, except in computational simulation studies, and the best we can do is to take gene trees as a proxy for species tree. Gene trees and species trees are the same only if separation of the orthologs of the selected gene coincides with the separation of species (speciation events). For example, the gene tree portrayed by the blue gene in *Figure* 3 is identical to its species tree. At times, gene trees and species trees have topological incongruence called 'lineage sorting'. For example, the gene tree portrayed by the red gene in *Figure* 3 is topologically different from the species tree. In such cases, taking gene trees as a proxy for species trees can be erroneous [2, 3].

Worked-out Example

For this example, we use the freely available software MEGA[6]. At the time of writing this tutorial, the most recent version was 6. As versions are updated, there may be slight changes in software buttons and navigations from what is described here [4, 5].

1. After downloading and installing, open MEGA.

2. Click 'Examples' folder icon from the bottom navigation ribbon and open the file 'Crab_rRNA'. The file will be loaded in MEGA. Have a look at the alignment by clicking 'TA' icon, or F4 keyboard shortcut. We will see that there are 13 sequences, with

[6] http://www.megasoftware. net/ .

their names on the left column. At the bottom-left corner, we see that there are 421 positions in each of these sequences. You may close this window and return to the main menu.

3. Click 'Models' icon in the top navigation ribbon. A new window will pop up. Change 'Gaps/Missing Data Treatment' value to 'complete deletion', in case this has not been set as default.

4. Analysis will take some time and a table of results will be presented. Left column describes various models as explained in this tutorial, arranged ascending order based on their Bayesian Information Criterion (BIC) scores. The corresponding BIC scores can be noted from third column. Best model is the first model in the table, which in this case is T92+G (Tamura 3 Parameter). Note it down, and close the window to go back to the main MEGA interface.

5. Now, let us construct some phylograms, first using the NJ method. In the main window, click 'Phylogeny' icon in the top navigation ribbon and choose the second option 'Construct/Test Neighbour-Joining Tree'.

6. A preferences window will be popped-up in which make sure that Substitution Model is 'Tamura 3 Parameter' and rate among the sites is 'Gamma Distributed (G)'. Change 'Gaps/Missing Data Treatment' value to 'complete deletion', in case this has not been set default. Finally, click 'Compute'[7].

7. Analysis will take some time and result will be presented as a phylogram. Click 'Caption' on the top right of menu to display the caption of this phylogram, which reveals more information about the parameters we employed. You can save this as an image, by clicking 'Image > Save as PNG file' from the top menu bar. Close the window and return to the main window.

8. This time, choose 'Construct/Test Maximum-Likelihood Tree' with similar options explained earlier in force. Compare this tree with the earlier NJ tree. Is the topology similar?

[7] A more detailed step-by-step protocol starting from DNA sequence assembly in CodonCode Aligner® and phylogenetic analysis in Geneious® is available at Nature Protocol Exchange: BAST, F 2013. Sequence Similarity Search, Multiple Sequence Alignment, Model Selection, Distance Matrix and Phylogeny Reconstruction. *Nature Protocol Exchange*. Nature Publishing Group. doi: 10.1038/protex.2013.065 Accessible at: http://www.nature.com/protocolexchange/protocols/2740

The phylogenetic inference is a powerful tool for inferring the evolution of a wide variety of subjects, especially biological taxa from a set of orthologous sequences by taking synapomorphy in consideration.

9. Return to MEGA main menu and click ‘Close Data’.

Try redoing the same analysis with ‘Pairwise deletion’ option for treating gaps/missing data in preferences window of model test and PI. Does this change the topology of the tree? You may now repeat the same workflow with other example alignments included with MEGA.

You may also like to download sequences of the locus of your choice (for example, another locus from *Table* 1), for taxa of your choice. For example, *COX*1 for taxa in *Figure* 3 in Part 1 of this article. To download sequences, you may click ‘Align’ icon from top navigation ribbon, and click ‘Query Databanks’. An NCBI–Nucleotide window will open in the built-in browser. Type ‘human *COX*1’ in the search bar, click ‘Search’, open the first result link, and click ‘Add to Alignment’ icon on the top menu. Alignment explorer will pop up with the first sequence loaded in. Repeat this for the rest of taxa. Once completed, align these sequences in the alignment explorer by clicking ‘Alignment>Align by Clustal-W’ from top menu bar. Click ‘K’ in the preferences pop-up window, by accepting default settings. Once aligning is done, an MSA will be presented in alignment explorer. You may clip (remove) ends of alignment to have same length for all sequences, by selecting columns and deleting those. Finally, click ‘Data>Phylogenetic Analysis’ in the alignment explorer, and choose ‘Yes’ for protein coding DNA pop-up (as *COX*1 is a gene). You may now perform analyses from Step 3 as described above.

Conclusion

Phylogenetic inference is a powerful tool for inferring the evolution of a wide variety of subjects, especially biological taxa from a set of orthologous sequences by taking synapomorphy in consideration. Since the advent of DNA sequencing technologies as well as ever-increasing computational power in agreement with Moore’s law, the field has tremendously expanded to supplement massive evidence to the theory of evolution through natural selection – first proposed by Darwin[8] and Wallace[9] in 19th CE. Today, ingenious students sitting at home, can reconstruct molecular phylogeny of taxa of their interest merely by downloading DNA sequences and analyzing them, using freely available phylogenetic packages, to make reasonably robust inferences about the evolutionary legacy of those taxa, and to appreciate the beauty of the theory of evolution.

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] 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.

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

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

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

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

[8] *Resonance*, Vol.14, No.2, 2009.

[9] *Resonance*, Vol.13, No.3, 2008.