MedCrave
Step into the World of Research

MOJ Proteomics & Bioinformatics

Review Article

Open Access

CrossMark

Selecting molecular markers for a specific phylogenetic problem

Abstract

In a molecular phylogenetic analysis, different markers may yield contradictory topologies for the same diversity group. Therefore, it is important to select suitable markers for a reliable topological estimate. Issues such as length and rate of evolution will play a role in the suitability of a particular molecular marker to unfold the phylogenetic relationships for a given set of taxa. In this review, we provide guidelines that will be useful to newcomers to the field of molecular phylogenetics weighing the suitability of molecular markers for a given phylogenetic problem.

Keywords: phylogenetic trees, guideline, suitable genes, phylogenetics

Volume 6 Issue 3 - 2017

Claudia AM Russo, Bárbara Aguiar, Alexandre P Selvatti
Department of Genetics, Federal University of Rio de Janeiro, Brazil

Correspondence: Claudia AM Russo, Molecular Biodiversity Laboratory, Department of Genetics, Institute of Biology, Block A, CCS, Federal University of Rio de Janeiro, Fundão Island, Rio de Janeiro, RJ, 21941-590, Brazil, Tel 21 991042148, Fax 21 39386397, Email claudia@biologia.ufrj.br

Received: March 14, 2017 | **Published:** November 03, 2017

Introduction

Over the last three decades, the scientific field of molecular biology has experienced remarkable advancements in data gathering and extensive phylogenetic analyses. The development of new technologies and the subsequent accessibility of refined methods due to cost reduction contributed to an immeasurable expansion of molecular facilities worldwide.[1] The rate of sequence submission has recently intensified for three primary reasons: the numerous and successful DNA barcoding projects,[2,3] the advent of Next Generation Sequencing[4] and the subsequent decrease in prices for molecular sequencing services.[5]

As a consequence, genetic data repositories such as GenBank have been doubling in size every 18 months,[6] rising from 606 sequences in the first 1982 release to close to 200 million sequences in the 218th release in February 2017. Molecular data from more than 260 thousand nominal species is now widely accepted as a paramount source of biological information in all life sciences.[7]

This is an exciting time. Many long existing controversies are in the process of being resolved by an unparalleled amount of data.[6,8,9] Phylogenomics, a dream just a few decades ago, is now changing the face of molecular phylogenetics. It is a revolution second only to the introduction of molecules in the field of phylogenetics in the 1960's.

However, the availability of large number of sequences is not necessarily associated with an accurate estimation of phylogenies, due to analytical errors associated with very large sets of sequence data.[10–12] Hence, the contentious matter of molecular marker sampling is inhibiting this new breakthrough. Different genes may yield strikingly contradictory topological patterns for a given diversity group.[13,14] Thus, the selection of suitable markers is critical in obtaining accurate estimates, but it is not a straightforward task. In this review, we aim to provide some guidelines for newcomers weighing the suitability of particular molecular markers for a given phylogenetic problem.

Homology in molecular markers

In phylogenetic reconstruction, as in comparative biology studies, the single most important concern lies in the matter of homology, a concept that occupies a central position in evolutionary biology.[15] Homology is a qualitative term, defined by equivalence of parts due to inherited common origin.[16–18]

Homology has been more recently defined as the relationship that binds all states of a single character and sets them apart from the states of other characters, supporting the logical equivalence of the notions of homology and synapomorphy (for review see).[19] The comparison of homologous sequences is critical in a phylogenetic analysis, because only homologous characters may reveal the actual phylogenetic pattern.

Nevertheless, a number of authors erroneously refer to homology as a synonym for similarity. Molecular biologists are particularly prone to this error, as they assert that 'two sequences share 70% homology' (for reviews on this problem see).[15,20] Two sequences might show 70% *similarity*, if 7 out of 10 aligned base pairs are identical between them.

In molecular sequences, the higher the similarity between two sequences, the more likely it is that they are homologous, because the probability of both sequences acquiring identical base pairs decreases as the sequences grow.[18] For instance, two identical 30-nucleotide-long sequences have an extremely low probability of being non-homologous, meaning that they would have attained the same sequence independently.

As defined above, homology cannot be measured, and thus is not a quantifiable concept.[18] Two characters are either homologous or they are not. Their homology indicates that their similar parts were already present in their common ancestor. This argument makes comparison between homologous parts a necessary component for phylogenetic inference about common ancestry recovery.[21]

Homology is not enough

To properly investigate phylogenetic relationships among a set of taxa, only homologous characters should be compared to reveal their common evolutionary history.[15,22] When comparing molecular data, homology of the sequences is obviously essential, but it is not sufficient. This is due to two main processes that result in homologous genes: *speciation* and *gene duplication*.[17] Two genes are homologous if they descend from a common ancestor. Nevertheless, if their divergence is due to a duplication event, these genes are *paralogous*. On the other hand, if their divergence is due to a speciation event, they are *orthologous* genes. In a phylogenetic context, the user must select only orthologous genes.

Figure 1 shows a single copy gene that existed at t0 in a hypothetical organism *Zalrus originalis*. In *Z. originalis*, this copy went through a duplication event that resulted in the paralogous copies α and β. Eventually, this condition (two copies) would have spread through all individuals of *Z. originalis*. In time, both copies would differentiate into distinct sequences, with or without functional differentiation. In this scenario, all *Z. originalis* individuals will present two homologous, paralogous copies of the original gene. Such paralogous copies provide no clues for phylogenetic inference, because they originated through a duplication event.

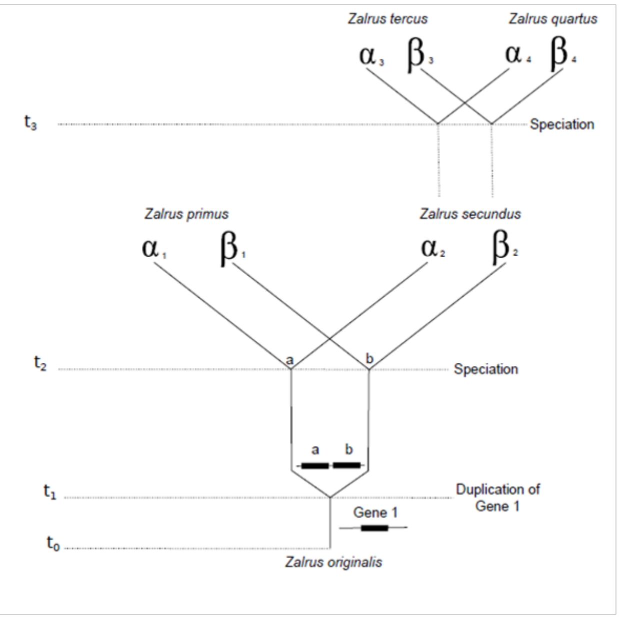


Figure 1 Homologous relationships between paralogous and orthologous copies of a gene.

If, however, *Zalrus originalis* goes through a speciation event, both descendant species will carry two copies of the gene α and β. In time, *Z. primus* and *Z. secundus* will go through diversification. Eventually, *Z. secundus* will also speciate into *Z. tercus* and *Z. cuartus*. Notice that all six copies of the original gene are homologous at t1, but copies α1, α3 and α4 (or β1, β3 and β4) are orthologous, whereas all copies α and β are paralogous.

Orthologous genes from different species should be selected for unraveling phylogenetic relationships among organism lineages,[23] because only these genes will carry speciation related information. Alternatively, paralogous copies should be used if the researcher aims to study duplication patterns in a gene family. In this case, all gene copies of that family must be used to disclose the duplication patterns in the corresponding phylogeny. It is necessary to add several species, all of which contain paralogous copies of the gene. This procedure will yield relative times of the duplication events related to the speciation events.

This approach sounds simple enough. When the researcher is concerned with phylogenetic problems, the same orthologous copy should always be chosen to build a phylogenetic hypothesis. Unfortunately, the distinction between copies α and β on a chromosome is not straightforward, because there are no labels on the chromosome.[15,22] This can become a major problem when dealing with gene families or genes with multiple copies, which are very common in most genomes.[23] In this sense, if we compare copy a of species X with copies b of species Y and Z, species Y and Z will be joined in the phylogeny, regardless of their phylogenetic relationship. In this case, the divergence time between species X and the remaining species will certainly be overestimated, and spurious phylogenetic relationships may be found.

For example, in most drosophilids there are several homologous copies of the gene that encodes for the alcohol dehydrogenase enzyme.[24] The phylogenetic tree in Figure 2 depicts the homologous relationships between orthologous and paralogous *Adh* genes in drosophilids. The relationship may be uncovered using topological analysis. Ancient duplication events such as this can be very helpful when employed as reciprocal out groups (an *Adh*1 sequence can be used as the out-group in an *Adh*2 phylogeny, and vice-versa).

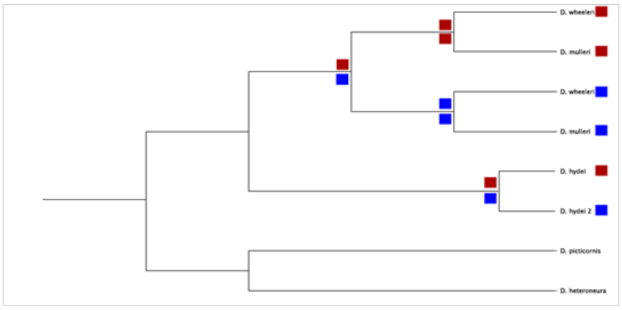


Figure 2 Homologous relationships between paralogous (different colours) and orthologous (same colour) of a gene shown in a phylogenetic analysis.

One multiple copy gene that is often used is the ribosomal RNA gene. There are often hundreds of copies in vertebrate genomes, making it clear that the duplication event that resulted in these copies took place before the divergence of vertebrates. However, due to a phenomenon called concerted evolution, in many species all copies within an individual are virtually identical; thus, paralogy/orthology should not be problematic when using this gene in vertebrates. How common this phenomenon is in other taxonomic groups remains to be shown, but it is now widely accepted that gene conversion and unequal crossing over are the main causes of concerted evolution.[25] Furthermore, the shorter the divergence time intervals between lineage splits, the more perfectly we have to assume concerted evolution dictates evolutionary rates, so as not to yield errors during phylogenetic reconstruction.

Homology in cytoplasmatic markers

In regard to animal mitochondrial DNA, issues such as gene duplications and paralogy/orthology are no longer a problem, because gene content, size and function are fairly constant across all metazoans. This ensures comparisons among orthologous genes in almost all cases. Conversely, plant mitochondrial DNA have distinct gene compositions compared to those found in metazoans and fungi.[26,27] Plants exhibit mitochondrial genomes 10 to 100 times as large as most metazoans, and many gene duplications have been reported.[28] The symbiotic events that resulted in the origination of the eukaryotic mitochondria[29,30] and chloroplast[31] were unique, but due to recombinations and duplications, it would be best to use mitochondrial genomes for more restricted phylogenetic purposes.[26,27,32] This is also true for chloroplast genomes that do not exhibit gene content stability among major lineages of plants.[33]

However, caution should be taken with mitochondrial and chloroplast gene copies that are horizontally transferred into the nucleus (paralogous copies), also known as *numts*.[5,28] The largest problem with *numts* is that copies inside the nucleus are susceptible to distinct evolutionary forces driving mutations compared to the original mitochondria. Hence, the model of evolution will change, making it impossible to fit into any (single) given available model.

Furthermore, purifying selection will tend to be relaxed so, resulting in a *numt* that usually evolves faster and rapidly turns into a pseudogene. Pseudogenes may be easily spotted when inspecting the alignment (see next section) and removed. If the *numt* is not yet a pseudogene, it is most likely very similar to the original mitochondrial gene and should not disrupt phylogenetic patterns. This is a good reason to verify protein coding gene alignments by checking to ensure that all sequences translate perfectly (with no stop codons) into amino acid sequences.

Extra caution must be taken when selecting genes to warrant comparison with orthologous sequences. If the user is unsure about the orthologous nature of the sequences they should be avoided, or all homologous sequences must be used to construct a preliminary phylogeny to define in advance the homologous relationships among sequences. For instance, after inspection of phylogenetic patterns of the Adh related genes, it is clear that only the *Adh*1 or the *Adh*2 genes should be used in a phylogenetic analysis of drosophilids of the mulleri species group.

Another issue that needs attention is heteroplasmy, or the fact that the organelle genome may exist as different copies in a single individual. In the vast majority of cases, organelle genomes are inherited through maternal lineages,[28,34] but cases of paternal inheritance have been reported in a handful of species, such as plants, bivalves, and mammals.[35] However, these inheritance patterns do not seem to be widespread enough to cause concern.[28] Furthermore, in phylogenies that sample species, genera or higher taxonomic ranks, the differences between male and female mitochondrial genomes in a single species will most likely not alter the phylogenetic pattern.

The alignment

A multiple alignment makes three major assumptions. The first is that the names of all sequences represent natural groups that are clustered in the correct tree (i.e., monophyletic groups). This is an important assumption that must serve as a guide when selecting sequence names. The names will carry biological information that will lead to the unfolding of real biological meaning in the final phylogenetic tree according to the phylogenetic patterns recovered. In this sense, the name must contain the name of a species if species monophyly may be assumed; otherwise, the name must include the geographical location from which the individual was sampled.

The second assumption is that all sequences are homologous, as previously discussed. Finally, the third major assumption of any multiple alignments is that each alignment column includes homologous bases for all species sampled.[18] Sequences modify over the course of evolution due to nucleotide substitutions or insertion/deletion events, including *indels*. For a given marker, such *indels* will result in sequences of different lengths when compared to orthologous copies from different species or paralogous copies within a species. Thus, the purpose of the sequence alignment procedure is to add *indels* for comparison of not only homologous genes but also those at a given alignment position (i.e., column) that encloses homologous base pairs between sequences.

A perfect alignment is the assurance that homologous positions are compared throughout the sequences, despite *indels* that have occurred during their evolution.[21] As previously mentioned, in a phylogenetic analysis homology is a critical asset. As variations accumulate between sequences, homology inference becomes more difficult due to homoplasies masking the (synapomorphic) evidence of homology. In fact, the amount of variation among analyzed sequences must be sufficient to unfold their actual evolutionary history, but not so extensive that substitutions are saturated.

Figure 3a shows an example alignment that is full of *indels*, indicating that the gene examined evolved quickly in the given diversity set. In this case, the sequences contain so many substitutions that the alignment likely has more substitutions than those that can be directly observed. Hence, when encountering an alignment that resembles this example, it is highly recommended to find more conservative genes to resolve the phylogenetic question.

There are many computer programs used to align sequences,[36–38] but some authors argue that if a computer is needed to perform an alignment, you should reconsider using those sequences to build a phylogenetic tree. Although this is often an exaggeration, it does bring attention to how crucial the alignment is when constructing phylogenies. Regardless of the alignment procedure, the computer-generated alignment should always be manually checked. However, if the number of possible alignments for a given sequence set is enormous, then the chances of generating incorrect or biased results are considerably high when over-manipulating the alignment.[39]

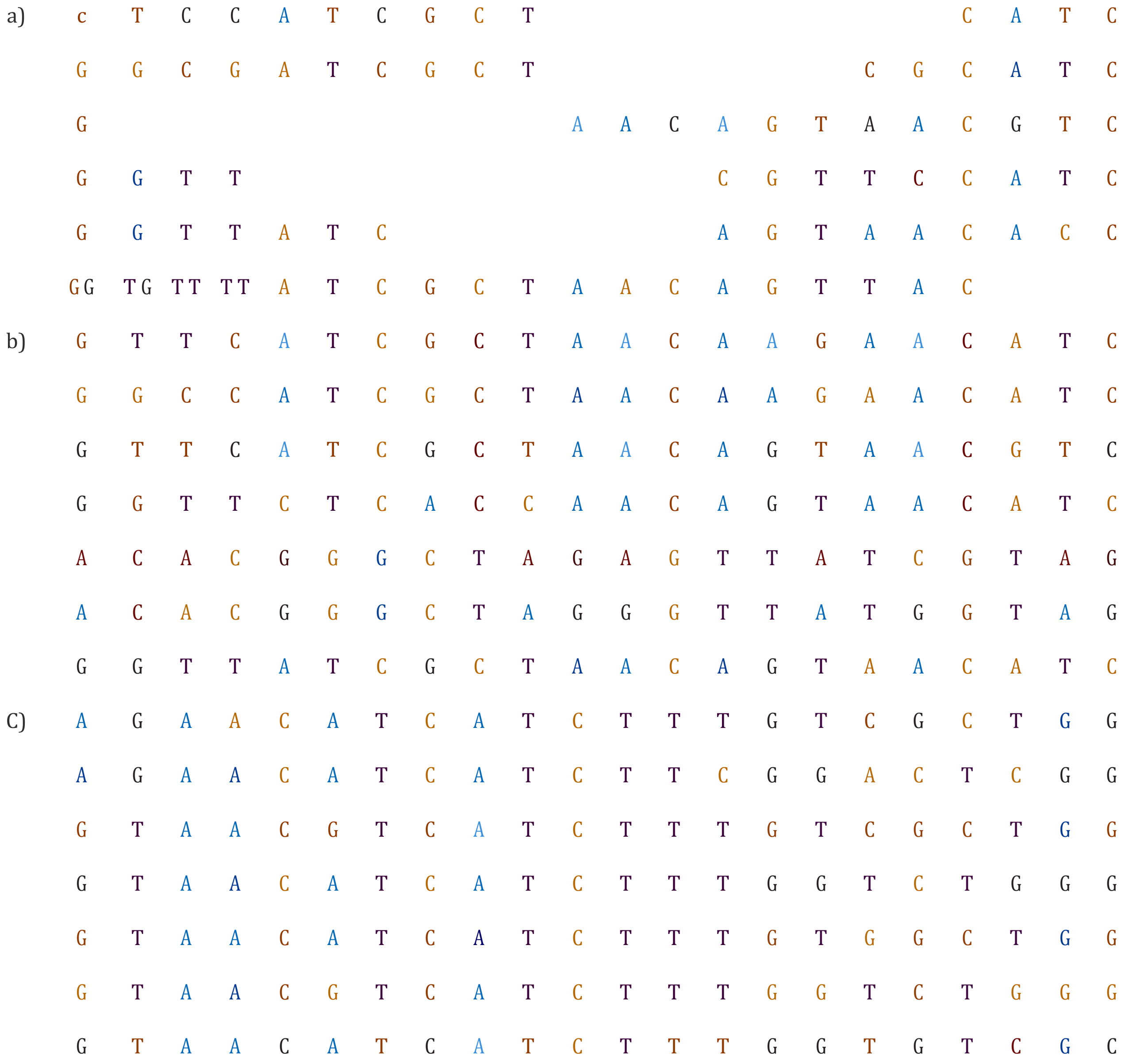
It is also possible that one or a few sequences do not fit well in the alignment (Figure 3b). In most cases, such sequences are reversed (or complementary reversed) or misidentified. In order to detect these cases, alignment must be inspected in detail. It is crucial that before proceeding with the phylogenetic analysis, they are perfectly aligned or removed.

Also, in some cases, the alignment is good enough but a portion has an unreliable alignment (Figure 3c), one must consider if there is a definite cutoff for removing that portion of the alignment. For instance, if introns show unreliable alignment, the removal criteria are straightforward and those parts should be removed before the analysis. However, if the region of bad alignment is defined by the user, difficulty arises as to how to avoid subjective analysis. If the region represents less than 10% of the alignment, maintaining the segment should not be necessary and over-manipulation of sequences can be avoided. If the region is more than 10%, the user must eliminate sequence segments in which the alignment homology cannot be guaranteed. In these cases a computer program such as T-Coffee or Gblocks are recommended to avoid subjective manipulation.[40]

It is important to align each marker individually to verify the reading frame of protein coding sequences or the secondary structure of ribosomal sequences before analysis. In many computer programs, this verification is coded into their algorithm.[41] Additionally, flanking positions that have not been sequenced for most individuals should be removed before proceeding with the phylogenetic analysis.

After the alignment

An indicator of the strength of the final alignment for a particular marker is given by calculating the proportion of different sites between sequences. Ideally, the average proportion of different sites should be between 0.1 and 0.3 for all sequence pairs, providing additional confidence that the comparisons are made through homologous positions. If the proportion of difference between sequence pairs is much lower than 0.1, the alignment most likely does not have enough variability to reveal the evolutionary relationships among lineages. In this case, tree topology will tend towards a polythomy, or a tree with no resolution.



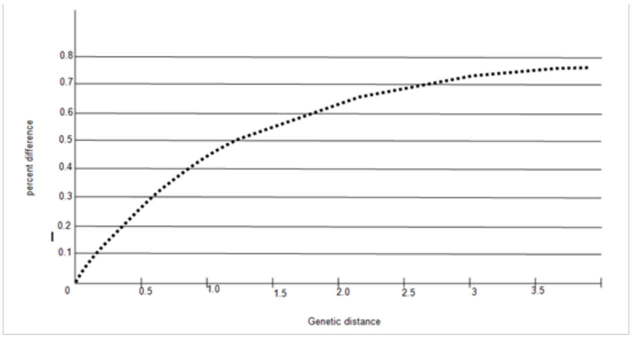
a)	c	T	C	C	A	T	C	G	C	T									C	A	T	C
	G	G	C	G	A	T	C	G	C	T							C	G	C	A	T	C
	G										A	A	C	A	G	T	A	A	C	G	T	C
	G	G	T	T										C	G	T	T	C	C	A	T	C
	G	G	T	T	A	T	C							A	G	T	A	A	C	A	C	C
	GG	TG	TT	TT	A	T	C	G	C	T	A	A	C	A	G	T	T	A	C			
b)	G	T	T	C	A	T	C	G	C	T	A	A	C	A	A	G	A	A	C	A	T	C
	G	G	C	C	A	T	C	G	C	T	A	A	C	A	A	G	A	A	C	A	T	C
	G	T	T	C	A	T	C	G	C	T	A	A	C	A	G	T	A	A	C	G	T	C
	G	G	T	T	C	T	C	A	C	C	A	A	C	A	G	T	A	A	C	A	T	C
	A	C	A	C	G	G	G	C	T	A	G	A	G	T	T	A	T	C	G	T	A	G
	A	C	A	C	G	G	G	C	T	A	G	G	G	T	T	A	T	G	G	T	A	G
	G	G	T	T	A	T	C	G	C	T	A	A	C	A	G	T	A	A	C	A	T	C
C)	A	G	A	A	C	A	T	C	A	T	C	T	T	T	G	T	C	G	C	T	G	G
	A	G	A	A	C	A	T	C	A	T	C	T	T	C	G	G	A	C	T	C	G	G
	G	T	A	A	C	G	T	C	A	T	C	T	T	T	G	T	C	G	C	T	G	G
	G	T	A	A	C	A	T	C	A	T	C	T	T	T	G	G	T	C	T	G	G	G
	G	T	A	A	C	A	T	C	A	T	C	T	T	T	G	T	G	G	C	T	G	G
	G	T	A	A	C	G	T	C	A	T	C	T	T	T	G	G	T	C	T	G	G	G
	G	T	A	A	C	A	T	C	A	T	C	T	T	T	G	G	T	G	T	C	G	C

Figure 3 Three alignments that indicate problems (shaded) for a phylogenetic analysis. a) Too many indels will decrease confidence on homology detection. b) Two sequences are aligned with each other but not with the remaining sequences. c) A portion of the alignment is not aligned properly.

Alternatively, if the proportion is much higher than 0.3, a higher degree of saturation and a very complex model of evolution will be required to accurately estimate phylogenetic relationships.[42] A complex model has many parameters (G+C content, transition/transversion ratio, etc.), and because each parameter has to be estimated, an error is associated with each particular estimate. Simpler evolutionary models include fewer assumptions and are more robust. If the proportion is higher than 0.7, homology may not be able to be inferred, due to the amount of noise. Because there are only four nucleotide types, we expect that two random sequences by chance will have 25% identical nucleotides.

It should be noted that mitochondrial genomes evolve at a different rate than the nuclear genome. In animals, the mitochondrial genome evolves much faster than the nuclear genome. Thus, it would be best to unfold evolutionary relationships among closely related taxa. However, mitochondrial genomes in plants evolve much slower than the nuclear genome (or any other genome).[26,28] Substitutions rates in the chloroplast appear to be slow, although they may vary among groups of plants.[43]

If too much variation is present, third codon positions may be eliminated to lower the noise in the alignment. Nevertheless, eliminating such positions will certainly remove most of the variability of the alignment and may result in the "not enough variation" problem. Moreover, homoplasies at one level may help resolve the tree at a different level.[44] More specifically, homoplasies are also homologies at a different level (Figure 4). Hence, removing third codon positions or fast evolving sites is bound to remove positions with phylogenetic signals, diminishing the overall stability of the tree.



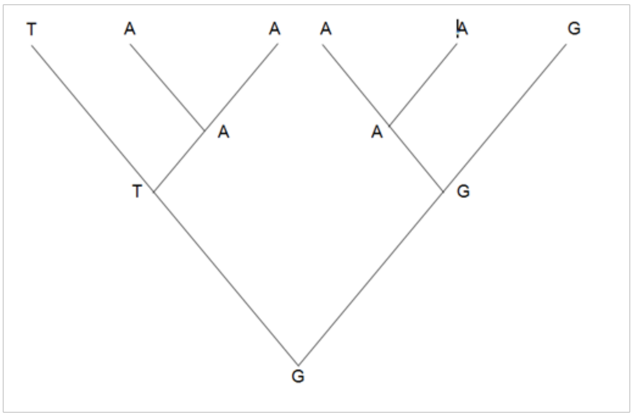


Figure 4 The A is both a homoplasy (as it appeared twice in the phylogenetic tree) but it is also a synapomorphy (as it appeared only once in each of the two clades). By removing the A position, both homoplasy and synapomorphies are removed.

It is important to ensure that the out group is perfectly aligned with the in group sequences. If it is too far removed and the alignment is made unreliable by the presence of numerous indels, a more closely related out group must be found. In cases where such out groups are not available, the mid-point technique should be used for rooting the tree without an out group. This method has been shown to perform quite well on empirical data.[45]

Number of sites

As genetic data banks grow larger, another common practice is the concatenation of genes into a metasequence of many thousand base pairs.[46,47] Although statistically reasonable[48] and widely employed to minimize sampling error, this technique is not without its shortcomings. In this section, we will go over some important points that should be addressed before selecting the fragments.

Sequence length has a great impact on phylogenetic inferences.[49] Random sampling error, or stochastic error, is a statistical definition for a class of errors or uncertainties that might be present in parameter estimates from one measurement to another. These types of errors are particularly sensitive to limited data. Thus, the combined analysis of several base pairs theoretically increases the phylogenetic signal.

However, there is a second class of errors that have a grave effect on phylogenetic inference, the systematic errors. They deserve greater attention in the present-day genomic era.[50] Systematic errors are generally defined as errors due to incorrect model assumptions and often result in inconsistent phylogenetic trees.[10] The major concern with systematic errors is that they are very difficult to detect, because they are errors associated with the measurement itself.[50] For instance, different genes may generate conflicting phylogenetic trees and still show high bootstrap values in combined analyses.[10,11] Thus, it is virtually impossible to detect conflicts among individual genes by relying on a single combined analysis.[51,52]

The issue of missing data versus (taxon and gene) sampling has been a focus of heated debate for over a decade.[7,53–55] A gene may be eliminated that has not been sampled for many species, such as genes only sampled in species with full genome sequences available. Alternatively, the entire species may be eliminated if sequences for only a few genes are available. This is the same principle as eliminating fossils from a morphology-based phylogenetic analysis that includes extant taxa. The debate arises on account of the fact that the exclusion of missing data will necessarily eliminate non-missing data as well. Most authors tend to support the inclusion of missing data for phylogenetic purposes.[7,41,56]

Assembling the data matrix

a. The selection of genes for specific phylogenetic problems is not a simple task. Although the statistical comparison of biological sequences is quite developed, researchers must be aware of the complexity of the evolutionary process itself. Here, we provide a quick guide for assembling a consistent matrix.

b. Select a given set of orthologous (phylogenetically homologous) markers to be used. In order to improve robustness, many markers should be chosen.

c. Align each set separately for individual markers. The alignment should be performed using protein reading frame or secondary structure information to better guide the alignment.

d. Carefully inspect alignments for each marker individually. In this step, the unaligned segments are easily detected and must be removed from the alignment.

e. Estimate the proportion of distance matrix for each of the markers after inspection. Ideally, p-distances should vary up to 0.3; that is, 30% of the sites vary when two sequences are compared so that saturation is not too high. If many markers are available, select those that fit this limit.

f. Before concatenating the individual gene alignments, it is useful to perform incongruence tests. If individual gene analyses point in different directions, they should not be assembled, because incongruence will be masked.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

References

1. Linnarsson E. Recent advances in DNA sequencing methods –general principles of sample preparation. *Exper Cell Res*. 2010;316(8):1339–1343.

2. Galimerti A, de Mattia F, Losa A, et al. DNA Barcoding as a new tool for food traceability. *Food Res Int*. 2013;50(1):55–63.

3. Weitschek E, Fiscon G, Felici G. Supervised DNA barcodes species classification: analysis comparisons and results. *Biodata Mining*. 2014;7(1):4.

4. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genetics*. 2010;11(1):31–46.

5. Li M, Schroeder R, Ko A, et al. Fidelity of capture–enrichment for mtDNA genome sequencing: influence of NUMTs. *Nuc Acid Res*. 2012;40(18):e137.

6. Clark K, Karsch–Mizrachi I, Lipman DJ, et al. Genbank. *Nucleic Acids Res*. 2016;44:D67–D72.

7. Wiens JJ, Tiu J. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. *PLoS One*. 2012;7(8):e42925.

8. Bayzid MS, Warnow T. Naïve binning improves phylogenomic analyses. *Bioinformatics*. 2013;29(18):2277–2284.

9. Johnson BR, Borowiec ML, Chiu JC, et al. Phylogenomics resolves evolutionary relationships among ants bees and wasps. *Curr Biol*. 2013;23(20):2058–2062.

10. Jeffroy O, Brinkmann H, Delsuc F, et al. Phylgenomics: the beginning of incongruence? *Trends Genet*. 2006;22(4):225–231.

11. Kumar S, Filipski AJ, Battistuzzi FU, et al. Statistics and truth in phylogenomics. *Mol Biol Evol*. 2012;29(2):457–472.

12. Seixas VC, Paiva P, Russo CAM. Complete mitochondrial genomes are not necessarily more informative than individual mitochondrial genes to recover a well–established annelid phylogeny. *Gene Reports*. 2016;5:10–17.

13. Lambret–Frotte J Perini FA, Russo CAM. Efficiency of nuclear and mitochondrial markers recovering and supporting known amniote groups. *Evolutionary Bioinformatics*. 2012;8:463–473.

14. Betancur RR, Naylor GJ, Ortí G. Conserved genes sampling error and phylogenomic inference. *Syst Biol*. 2013;63(2):257–262.

15. Haggerty LS, Jachiet PA, Hanage WP, et al. A pluralist account of homology: adapting the models to the data. *Mol Biol Evol*. 2014;31(3):501–516.

16. Lewin R. *Patterns in evolution: the new molecular view*. USA: Scientific American Library; 1997.

17. Fitch WM. Homology a personal view on some of the problems. *Trends Genet*. 2000;16(5):227–231.

18. Phillips AJ. Homology assessment and molecular sequence alignment. *J Biomed Inform*. 2006;39(1):18–33.

19. Brower AVZ, de Pinna MCC. Homology and errors. *Cladistics*. 2012;28(5):529–538.

20. Patterson C. Homology in classical and molecular biology. *Mol Biol Evol*. 1988;5(6):603–625.

21. Morgan MJ, Kelchner SA. Inference of molecular homology and sequence alignment by direct optimization. *Mol Phylogenet Evol*. 2010;56(1):305–311.

22. Dufayard JF, Duret L, Penel S, et al. Tree pattern matching in phylogenetic trees: automatic search for orthologs and paralogs in homologous gene sequence databases. *Bioinformatics*. 2005;21(11):2596–2603.

23. Curtis DS, Phillips AR, Callister SJ, et al. SPOCS: software for predicting and visualizing orthology/paralogy relationships among genomes. *Bioinformatics*. 2013;29(20):2641–2642.

24. Russo CA, Takezaki N, Nei M. Molecular phylogeny and divergence times of drosophilid species. *Mol Biol Evol*. 1995;12(3):391–404.

25. Li WH. *Molecular Evolution*. USA: Sinauer Associates; 2000.

26. Bullerwell CE, Gray MW. Evolution of the mitochondrial genome: protest connections to animals fungi and plants. *Curr Opin Microbiol*. 2004;7(5):528–534.

27. Zhang T, Fang Y, Wang X, et al. The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. *PLoS One*. 2012;7(1):e30531.

28. Xiong AS, Peng RH, Zhuang J, et al. Gene duplication and transfer events in plant mitochondria genome. *Bioch Biophys Res Comm*. 2008;376(1):1–4.

29. Herrmann JS. Converting bacteria to organelles: evolution of mitochondrial protein sorting. *Trends Microbiol*. 2003;11(2):74–79.

30. Zimorski V, Ku C, Martin WF, et al. Endosymbiotic theory for organelle origins. *Curr Op Microbiol*. 2014;22:38–48.

31. Xiong AS, Peng RH, Zhuang J, et al. Gene duplication transfer and evolution in the chloroplast genome. *Biotechnol Adv*. 2009;27(4):340–347.

32. Gray M. The incredible shrinking organelle. *EMBO Rep*. 2011;12(9):873.

33. Palmer JD. Comparative organization of chloroplast genomes. *Annu Rev Genet*. 1985;19:325–354.

34. Sato M, Sato K. Maternal inheritance of mitochondrial DNA. *Autophagy*. 2012;8(3):424–425.

35. McCauley DE. Paternal leakage heteroplasmy and the evolution of plant mitochondrial genomes. *New Phytolog*. 2013;200(4):966–977.

36. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 20. *Bioinformatics*. 2007;23(21):2947–2948.

37. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acid Res*. 2004;32(5):1792–1797.

38. Katoh K, Standley DM. MAFFT multiple sequence alignment software versions 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772–780.

39. Slowinski JB. The number of multiple alignments. *Mol Phylogenet Evol*. 1998;10(2):264–266.

40. Notredame C, Higgins DG, Heringa J. T–Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 2000;302(1):205–217.

41. Filipski A, Murillo O, Freydenzon E, et al. Prospects for Building Large Timetrees Using Molecular Data with Incomplete Gene Coverage among Species. *Mol Biol Evol*. 2014;31(9):2542–2550.

42. Nei M, Kumar S. *Molecular Evolution and Phylogenetics*. India: Oxford University Press; 2000.

43. Erixon P, Oxelman B. Whole–gene positive selection elevated synonymous substitution rates duplication and indel evolution of the chloroplast clpP1 gene. *PLoS One*. 2008;3(1):e1386.

44. Kallersjo M, Albert VA, Farris JS. Homoplasy increases phylogenetic structure. *Cladistics*. 1999;15(1):91–93.

45. Hess PN, Russo CAM. An empirical test of the mid–point rooting method. *Biol J Linn Soc*. 2007;92:669–674.

46. Murphy WJ, Eizirik E, Johnson WE, et al. Molecular phylogenetics and the origins of placental mammals. *Nature*. 2001;409(6820):614–618.

47. Murphy WJ, Eizirik E, O'Brien SJ, et al. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*. 2001;294(5550):2348–2351.

48. Nei M, Xu P, Glazko G. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc Natl Acad Sci USA*. 2001;98(5):2497–2502.

49. Townsend JP, Lopez–Giraldez F. Optimal selection of gene and in group taxon sampling for resolving phylogenetic relationships. *Syst Biol*. 2010;59(4):446–457.

50. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. *Nat Rev Genet*. 2012;13(5):303–314.

51. Kubatko LS, Degnan JH. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol*. 2007;56(1):17–24.

52. Galtier N, Daubin V. Dealing with incongruence in phylogenomic analyses. *Philos Trans R Soc Lond B Biol Sci*. 2008;363(1512):4023–4029.

53. Rokas A, Carrol SB. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol*. 2005;22(5):1337–1344.

54. Wiens JJ. Missing data and the design of phylogenetic analyses. *J Biomed Inf*. 2006;39(1):34–42.

55. Nabhan AR, Sarkar IN. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief Bioinform*. 2011;13(1):122–134.

56. Heath TA, Hedtke SM, Hillis DM. Taxon sampling and the accuracy of phylogenetic analyses. *J Syst Evol*. 2008;46(3):239–257.