Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory

Title

Beyond Linear Sequence Comparisons: The use of genome-level characters for phylogenetic reconstruction

Permalink

https://escholarship.org/uc/item/1m6577xx

Author

Boore, Jeffrey L.

Publication Date

2004-11-27

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Beyond linear sequence comparisons: The use of genome-level characters for phylogenetic reconstruction

Jeffrey L. Boore[1,2,3]

[1] DOE Joint Genome Institute and Lawrence Berkeley National Laboratory. 2800 Mitchell Drive, Walnut Creek, CA 94598

[2] Department of Integrative Biology, 3060 Valley Life Science Building, University of California, Berkeley, CA 94720

[3] Genome Project Solutions, 1024 Promenade Street, Hercules, CA 94547

Corresponding author: Boore, J. L. (JLBoore@berkeley.edu)

Although the phylogenetic relationships of many organisms have been convincingly resolved by the comparisons of nucleotide or amino acid sequences, others have remained equivocal despite great effort. Now that large-scale genome sequencing projects are sampling many lineages, it is becoming feasible to compare large data sets not only of DNA and protein sequences, but also of genome-level features such as gene arrangements and the positions of mobile genetic elements. Although it is unlikely that the comparisons of genome-level features will address a large number of evolutionary branch points across the broad tree of life owing to the infeasibility of such sampling, they have great potential for resolving many critical, contested relationships for which no other data seem promising. However, it is important that we recognize potential pitfalls, establish reasonable standards for acceptance and use rigorous methodology to guard against any tendency to accept a plausible narrative as a substitute for a careful analysis. Here I discuss the advancements, advantages, methods and problems of the use of genome-level characters for reconstructing evolutionary relationships.

Introduction

Although molecular sequence comparisons have revolutionized systematics, solving some evolutionary relationships seems to require another approach

Over the past 25 years, molecular sequence comparisons have overturned many long-standing views on evolutionary relationships. For example, molecular phylogenetics provided the first strong evidence that humans are most closely related to chimpanzees [1,2]. Despite many similarities to fungi, oomycetes are now known to be more closely related to diatoms and brown algae [3], whereas microsporidians, a group once believed to be protists basal within the

Eukaryota, are actually fungi [4,5]. Extensive revision of phylogenetic relationships among animals has overturned the previous reliance on features of the coelom and segmentation [6]. However, some of the relationships that were equivocal in early molecular studies have remained highly recalcitrant, even as much more DNA sequence data has become available. Box 1 has a set of potential explanations for why sequence comparisons sometimes fail.

It may be that molecular sequence comparisons have created an expectation that many true evolutionary relationships will never be clearly demonstrated by a strong signal. Rather, there will be much 'noise' in the data, caused by superimposed, parallel and convergent molecular substitutions that confuse reconstruction and generate false signals. We have come to expect that many relationships will only be solved by ever-more sophisticated methods of data analysis to tease the weak signal from the noisy data. This causes concern that differing preferences for various methods will ensure that no consensus on many evolutionary relationships will ever be reached.

Yet, despite a focus on unsolved evolutionary relationships, there are many clades that are so well supported by synapomorphies that they are universally accepted. For example, Mammalia, Tetrapoda and Echinodermata are each solidly supported by shared features that are very unlikely to have reverted to an earlier condition or to have occurred multiple times in parallel, so they could only have arisen once in the common ancestor of the group. Are there other, unstudied characters that could be similarly powerful and that could provide confident resolution of some problematic evolutionary relationships? Box 1 includes a "wish-list" for the types of characters that would be ideal. Although these might be found otherwise, perhaps in the study of tissue ultrastructure, for example, the most promising area in which to look might be within the burgeoning field of comparative genomics.

The genomics revolution provides many new phylogenetic characters

Much more DNA sequence is being determined than ever before. The six largest genome sequencing centers can collectively produce well over 150 billion nucleotides of DNA sequence per year, and complete genome sequences of at least draft quality will soon be available for more than 200 eukaryotes and more than 1,000 prokaryotes (see http://www.genomesonline.org/). This provides systematists not only more sequences to compare, but also many higher order features that can be called "genome-level characters", examples of which are listed in Box 2.

In general, features of this type would be expected to change in a saltatory, non-clocklike manner. At first, this might seem to be a disadvantage, given that evolutionary biologists typically search for genes with clocklike behavior so that estimates can be made for times of lineage splitting. However, it is this non-clocklike behavior that makes genome-level characters especially useful for addressing the most difficult sets of relationships, that is, where the internodes are short (so that there has been little opportunity for signal to have accumulated) and the terminal branches are long (so that there has been much opportunity for the scant signal to have eroded). Clocklike characters perform poorly in this case, because the signal-to-noise ratio in the data closely matches the ratio of the time periods of the internode to the terminal branches. It is the least clocklike of characters that are expected to prevail, where an occasional and abrupt change might have occurred and then remained.

Of course, the use of genome-level characters will generally only be definitive if they are heavily weighted relative to other characters from morphology or DNA sequences, and it will take a thorough examination to see how reliable, consistent and informative their use will be. Genome-level characters have addressed only a few evolutionary branch points, but have

provided strong resolution of some relationships that had previously been unyielding despite extensive investigation, and have shown a remarkable lack of homoplasy, bolstering confidence in the reliability of these conclusions. (See [7] for a set of early examples.) The two most fruitful examples to date are the use of mitochondrial gene arrangements and retroelement positions, so I will focus on these, although I touch on other genome-level characters with potential for further clarification of evolutionary relationships.

Comparisons of mitochondrial genomes have pioneered the use of genome-level characters for phylogenetic inference.

GenBank (http://www3.ncbi.nlm.nih.gov/) has complete mtDNA sequences for > 600 species and entries large enough to include three or more contiguous genes for > 5000 species. The arrangements of the genes they contain provide a complex set of characters, appear to be selectively neutral, and change in a saltatory, non-clocklike manner [8,9], and so constitute the first well-developed data set of genome-level characters. See Box 3 for a list of cases where these have outlined evolutionary relationships.

One example is in addressing the relationships among major groups of arthropods and related phyla. Until a decade ago there was a consensus that Hexapods (insects and similar animals) are the sister group to Myriapods (e.g. millipedes and centipedes) within Arthropoda. Some even attempted to explain the embryological and genetic details of how the unspecialized, serially repeated segments of myriapods gradually evolved into the specialized segments of insects [10]. Many scientists also accepted the inclusion of the Onychophora in this group [11] to form the Uniramia, named for their sharing of single-branched (i.e. uniramous) appendages in contrast to the double-branched appendages of crustaceans. This hypothesis grew less tenable

with the finding of a gene transposition that appears to reliably mark a clade of hexapods and crustaceans, and to include the Pentastomida (previously considered to be a separate phylum). This clade excludes the myriapods and onychophorans, which retain the ancestral gene arrangement of the more distantly related chelicerate arthropods and of many less related phyla (Figure 2) [12-14]. Although this transposition is a single evolutionary character, it is the only change in the arrangement of all 37 mitochondrial genes for many of the taxa shown in Figure 2, even though they have been separated for hundreds of millions of years. It seems unlikely that this transposition would have occurred identically in separate lineages or that the gene would have transposed back to its ancestral location, as would have had to happen if the myriapods or onychophorans were closer to hexapods than are these crustaceans.

How should gene arrangement data be analyzed?

The most useful approach, first presented as a 'gene-adjacency matrix' [12] and elaborated with the name 'maximum parsimony of multiple encodings' [15], views gene boundaries as the characters. Each character takes the form 'downstream (or upstream) of gene A' and the possible character states take the form 'the 5' (or 3') end of gene B'. The main shortcoming of this approach is that a gene moving 'out' of a location independently in two lineages will create a synapomorphy as the two flanking genes become adjacent, whereas the different locations of the transposed gene could provide the information to resolve this. For this reason, and because single gene boundaries are generally weak characters, little reliance should be placed on uniting taxa simply because they share the juxtaposition of two genes, whereas sharing longer co-linear blocks can be very reliable. As illustrated in Figure 2, cladistic reasoning [16] should be used to

reconstruct evolutionary relationships from gene arrangements rather than simply uniting taxa bases on shared similarity.

Computational biologists working to improve methods are largely focused on improving computational tractability by using phenetic methods, trying to better calculate "distances" among genomes. This is futile, however, because accuracy depends on the assumption of a gene rearrangement "clock", which is not met by the data. Phenetic methods do not differentiate between traits that are shared owing to a slow rate of change from an ancestral condition and those that are uniquely derived for a clade, thus generating errors in phylogenetic reconstruction, a shortcoming that has been long recognized for molecular sequence comparisons.

Retroelement positions as genome-level characters

The process of retrotransposition inserts a copy of a DNA sequence element at another genomic location by using an RNA intermediate (see [17] for details). Because these elements insert into random genomic locations and remain there until normal degradation processes (e.g. base substitution and deletion) erase them, it seems unlikely that they would integrate more than once into the same location or be abruptly removed. Therefore, cases of evolutionary convergence or reversion are expected to be rare.

Box 3 lists examples where phylogenetic analysis of retroelement positions has resolved several relationships, even where all other data had been equivocal. One example is in the relationships among several groups of primates (Figure 3). The evolutionary derivation of tarsiers has been long unresolved, with some inferring them to be the sister group to the strepsirhines (e.g. lemurs and lorises) but others placing them as the sister group to the Anthropoidea (new world monkeys, old world monkeys and apes). The strongest evidence to

date is found in the integration of retroelements at shared locations in the genes for zonadhesion, α-1-microglobulin-bikunin and ATP synthase β subunit for *Tarsius bancanus*, human, and representatives of both groups of monkeys [18]. Other studies have identified retroelement insertions that define numerous branch points within the Strepsirhini [19] and the apes [20]. In all of the analyses outlined in Figure 3, there is only one case of homoplasy identified, that apparently being the result of the pattern of segregation of polymorphism (see below).

It is not sufficient to screen for insertions by PCR alone

Generally, one first identifies retroelement positions in one genome, then searches for them in other taxa of interest by PCR using primers matching flanking sequence. However, variation in the size of the products alone is insufficient for determining character states, because similar variation in size could be caused by independent events. The sequence must be determined for those fragments hypothesized to be diagnostic to ensure that the size variation is caused by the same insertion event.

In the example above regarding the evolutionary placement of the tarsiers [18], a total of 14 loci were identified where the PCR gave a larger fragment for the tarsier, humans, and monkeys, suggesting that these were all shared integrations. However, only three of these were confirmed as shared integrations, with the other 11 being independent transpositions into the same intron, but at different locations. Furthermore, even the three that are inferred to each stem from a single integration have variations in amplified fragment size caused by subsequent events, but determining the sequence still allowed inference of each integration being a single event.

New methods for identifying genome rearrangements

New techniques are available for high throughput screening to identify genome rearrangements. One possibility is to determine the end-sequences of a large number of clones from a genomic library (usually in BAC or fosmid vectors) then to align these with the whole-genome sequence of a related organism to identify inconsistencies that signal genome rearrangements. The clones that contain rearrangements are then sequenced to determine the differences between the two genomes in finer detail. Gene rearrangements, losses and duplications can also be identified using comparative genomic hybridization (CGH) chips with tiled large-insert clones, as has been done for a sampling of diverse human populations [21] and more broadly across the great apes [22] or by using arrays of oligonucleotides (representational oligonucleotide microarray analysis, ROMA) [23].

There are specific problems with high-throughput screening techniques

Using traditional, smaller scale studies, one could rely on the low frequency of cloning artifacts for accuracy. For example, if fosmid clones containing homologous genes are isolated from ten animals, it is likely that their sequences accurately represent those of the corresponding genomes. However, if one determines the sequences of the ends of 30,000 fosmid clones and aligns these to a related genome, enabling a very sensitive screen, it is likely that some will appear falsely to contain rearrangements; the large number of clones screened multiplied by the low frequency of cloning artifact gives an observable number of clones. It is important to verify that results of high-throughput screening represent the actual condition in the genome by conducting follow-on experiments such as identifying additional independent clones containing the same sequence or PCR amplifying portions of the genome to verify results.

General problems, pitfalls and reasonable standards

In addition to the caveats mentioned above that are specific to certain types of data, there are some general considerations for deriving phylogeny using genome-level characters.

Whole-genome shotgun projects yield incomplete genome sequences

The sequences of many more genomes are being drafted than finished. Consequently, the data sets have many gaps, missing genes, misassemblies and errors. The problem is exacerbated as more genomes are compared, because the problems might lie in different relative portions of each genome, so that only a small portion of the overall data set might be comprehensive. Even when the only genome-level sequences were of mtDNAs, some studies [24] made errors by assuming that whatever was in GenBank for gene annotation was correct, and this problem would be much worse when comparing whole nuclear genomes. All interpretations must consider that elements missing from the whole-genome shotgun sequence assembly might not be actually missing from the genome and that errors of gene annotation will be common. One should always examine the specific features being used to draw phylogenetic conclusions for correctness by using primary data and, whenever possible, should verify that these features are present in the genome itself.

Gene models contain errors

Gene models are created using *ab initio* methods [25], matching to EST and cDNA sequences and similarity searches to the gene models of other organisms. Errors occur, including the prediction of genes that are not real, the fusion of two genes into one, the split of one gene into

two, the erroneous prediction of the initiation site, the missing of an exon, or the modeling of an exon that does not exist. Furthermore, genomes often contain significant numbers of pseudogenes – non-functioning, broken and disintegrating genes that often arose by gene duplication. In some cases, it is difficult to accurately predict whether the sequence being considered is functional versus being a pseudogene, since indicators of the latter, e.g. frameshift mutations or in frame stop codons, may instead be sequencing errors or may be corrected by splicing out of small introns from the mRNA. Once again, it is imperative to consider the possibilities of errors in the data being used to draw phylogenetic conclusions.

False signal of relatedness can result from random segregation of ancestral polymorphisms

This is a long-recognized problem that is not unique to the use of genome-level characters. If a polymorphism in a trait persists through two successive lineage splits and the same variant is then lost convergently in two lineages that are not the most closely related pair, the result is a gene tree that differs from the real species tree. (See [26] for an explanation specific to the use of SINES.) This is described in some detail in a study of retroelement insertion pattern that robustly supports a clade of human and chimpanzee to the exclusion of gorilla [20], where there was, surprisingly, a single insertion shared uniquely between human and gorilla. The most probable explanation is that there was a polymorphism for this insertion in the ancestral population that persisted between the split of gorilla from human-chimp and the subsequent split of chimp from human, with the ‘inserted’ form becoming fixed in gorilla; thus at the point of the human-chimp split, both variants were still available for partitioning into these two lineages.

Some genome-level characters have particular problems

Phylogenies based on gene content have specific limitations. First, the shared loss of any character is generally weak support for relatedness, because there are many possible paths to loss that can occur independently. Secondly, gene losses might, in some cases, not be real, either because the gene has evolved beyond easy recognition of homology or because it is in a gap in a draft genome assembly or is poorly modeled computationally.

Some studies have been based on shared gene complements that emphasize the use of specific diagnostic amino acids, particularly for members of the *Hox* gene family [27]. This is, in large part, a phylogeny based on gene content with the same types of problem, plus it also suffers from the possibility of convergence, since only a few amino acids are used to diagnose gene homology.

Variations in the use of the genetic code, especially common for mitochondrial genomes, but also known for some nuclear genomes [28], have been used to infer phylogeny [29]. It is not yet clear how commonly lineages adopt the same variation convergently, and this concern is exacerbated by noting that such variations are almost all a shift within one codon family; for example, whereas the 'universal' code uses ATA, ATC and ATT for isoleucine and only ATG for methionine, a common variant shifts ATT to also specify methionine. Cases of apparent convergence have been noted, for example, where AAA has shifted to specify asparagine (rather than lysine) independently in the mitochondrial systems of echinoderms and some flatworms [29]. One particular problem is that few of the code variations are determined by experimentation; almost all have been made based only on comparing sequences among various mtDNAs, such that there is significant potential for errors (and for compounding of them).

The potential secondary structures of tRNAs and rRNAs might contain phylogenetic information. Aside from the obvious problem with the possibility of real convergent change, there is a concern that the models for folding will falsely generate apparently similar structures for RNAs in cases where they have simply evolved independently to be shorter or longer. The convergence then would be only in the optimization of inferring the folding, rather than in the folding itself.

It might be that insertions or deletions of individual amino acids are a class of exceptionally reliable characters [30,31]. However, here too, one problem might be that convergence to having simply a larger or smaller size, coupled with the propensity of the alignment algorithms to favor aligned gaps, might create false signal of shared insertion or deletion. Even if the alignment correctly reflects homology, there might also be some regions of any protein that can more easily accommodate insertions or deletions (perhaps owing to structural requirements), such that real convergence is not unlikely.

Conclusion

Organelle genomics gave science its first set of genome-level characters for phylogenetic reconstruction. Although many robust conclusions have come from these comparisons, the community has not yet established robust methods or standards for data analysis even for these diminutive genomes. Yet we have now crashing on our shores the first powerful waves of whole-nuclear genome sequences, vastly larger and more complex. Much of these analyses are being conducted from first principles by those whose primary training has not been in evolutionary biology and, in some cases, who are naive of its standards for drawing phylogenetic conclusions. Genome-level characters provide the best data set for reconstructing convincing relationships for

some of the most hotly contended nodes in the tree of life and establishing a framework for all organismal relationships. However, we will be successful only if we work assertively to recognize the potential pitfalls, establish reasonable standards for acceptance and use rigorous methodology to guard against any tendency to accept a plausible narrative as a substitute for a careful analysis.

Acknowledgements

This work was performed under the auspices of the US Department of Energy, Office of Biological and Environmental Research, under contract No. DE-AC02-05CH11231 with the University of California, Lawrence Berkeley National Laboratory. Funding was provided by the National Science Foundation through grants EAR-0342392, DEB-0120709, MCB-0242131, DEB-0089624, EF-0228729, DEB-0445047 and EF-0328516.

References

1 Miyamoto, M.M. *et al.* (1987) Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science* 238, 369-373

2 Bailey, W.J. *et al.* (1992) Reexamination of the African hominoid trichotomy with additional sequences from the primate beta-globin gene cluster. *Mol. Phylogenet. Evol.* 1, 97-135

3 Sogin, M.L. *et al.* (1996) Ancestral relationships of the major eukaryotic lineages. *Microbiologia* 12, 17-28

4 Baldauf, S.L. *et al.* (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290, 972-977

5 Van de Peer, Y. *et al.* (2000) Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. *Gene* 246, 1-8

6 Adoutte, A. *et al.* (1999) Animal evolution – The end of the intermediate taxon? *Trends Genet.* 15, 104-108

7 Rokas, A. and Holland, P. W. H. (2000) Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* 15, 454-459

8 Boore, J.L. and Brown, W.M. (1998) Big trees from little genomes: Mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev*. 8, 668-674

9 Boore, J.L. (1999) Animal mitochondrial genomes. *Nucleic Acids Res*. 27(8), 1767-1780

10 Raff, R.S. and Kaufman, T.C. (1983) *Embryos, Genes, and Evolution*, pp. 251-261, Indiana University Press

11 Manton, S.M. and Anderson D.T. (1979) Polyphyly and the evolution of the arthropods. In: *The Origin of Major Invertebrate Groups* (House MR, ed), pp. 269-322, Clarendon Press.

12 Boore, J.L. *et al.* (1995) Deducing arthropod phylogeny from mitochondrial DNA rearrangements. *Nature* 376, 163-165

13 Lavrov, D. *et al.* (2004) Phylogenetic position of the Pentastomida and (pan)crustacean relationships. *Proc. R. Soc. B* 271, 537-544

14 Boore, J.L. *et al.* (1998) Gene translocation links insects and crustaceans. *Nature* 392, 667-668

15 Wang, L.S. *et al.* (2002) Fast phylogenetic methods for the analysis of genome rearrangement data: an empirical study. *Pac. Symp. Biocomput.* 524-535

16 Kitching, I.J. *et al.* (1998) *Cladistics*, Oxford University Press

17 Shedlock, A. and Okada, N. (2000) SINE insertions: powerful tools for molecular systematics. *Bioessays* 22, 148-160

18 Schmitz, J. *et al.* (2001) SINE insertions in cladistic analyses and the phylogenetic affiliations of *Tarsius bancanus* to other primates. *Genetics* 157, 777-784

19 Roos, C. *et al.* (2004) Primate jumping genes elucidate strepsirrhine phylogeny. *Proc. Natl. Acad. Sci. U. S. A.* 101, 10650-10654

20 Salem, A.-H. *et al.* (2003) Alu elements and hominid phylogenetics. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12787-12791

21 Sharp, A.J. *et al.* (2005) Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* 77, 78-88

22 Locke, D.P. *et al.* (2003) Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* 13, 347-357

23 Sebat, J. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science* 305, 525-528

24 Blanchette, M. *et al.* (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.* 49, 193-203

25 Yao, H., *et al.* (2005) Evaluation of five *ab initio* gene prediction programs for the discovery of maize genes. *Plant Mol. Biol.* 57, 445-460.

26 Shedlock, A. *et al.* (2004) SINES of speciation: tracking lineages with retroposons. *Trends Ecol. Evol.* 19, 545-553

27 deRosa R. *et al.* (1999) Hox genes in brachiopods and priapulids and protostome evolution. *Nature* 399, 772-776

28 Santos, M.A.S. *et al.* (2004) Driving change: the evolution of alternative genetic codes. *Trends Genet.* 20, 95-102

29 Telford, M.J. *et al.* (2000) Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. *Proc. Natl. Acad. Sci. U. S. A.* 97, 11359-11364

30 Wolf, Y.I. *et al.* (2004) Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res.* 14, 29-36

31 Leebens-Mack, J. *et al.* (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Mol. Biol. Evol.* 22, 1948-1963

32 Naylor, G.J.P. and Brown W.M. (1998) Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* 47, 61-76

33 Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401-410

34 Fukuda, M. *et al.* (1985) Mitochondrial DNA-like sequences in the human nuclear genome: Characterization and implications in the evolution of mitochondrial DNA. *J. Mol. Biol.* 186, 257-266

35 Richly E. and Leister D. (2004) NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* 21, 1081-1084

36 Santos, M.A.S. *et al.* (2004) Driving change: the evolution of alternative genetic codes. *Trends Genet.* 20, 95-102

37 von Nickisch-Rosenegk, M. *et al.* (2001) Sequence and structure of the mitochondrial genome of the tapeworm *Hymenolepis diminuta*: Gene arrangement indicates that platyhelminths are derived eutrochozoans. *Mol. Biol. Evol*. 18, 721-730

38 Helfenbein, K.G. and Boore, J.L. (2004) The mitochondrial genome of *Phoronis architecta*—Comparisons demonstrate that phoronids are lophotrochozoan protostomes. *Mol. Biol. Evol*. 21(1), 153-157

39 Boore, J.L. and Staton, J. (2002) The mitochondrial genome of the sipunculid *Phascolopsis gouldii* supports its association with Annelida rather than Mollusca. *Mol. Biol. Evol*. 19(2), 127-137

40 Boore, J.L. and Brown, W.M. (2000) Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Mol. Biol. Evol.* 17, 87-106

41 Mindell, D. *et al.* (1998) Multiple independent origins of mitochondrial gene order in birds. *Proc. Natl. Acad. Sci. U. S. A.* 95, 10693-10697

42 Boore, J.L. *et al.* (1999) Complete sequence, gene arrangement and genetic code of mitochondrial DNA of the cephalochordate *Branchiostoma floridae* ("Amphioxus"). *Mol. Biol. Evol.* 16, 410-418

43 Kumazawa, Y. and Nishida, M. (1995) Variations in mitochondrial tRNA gene organization of reptiles as phylogenetic markers. *Mol. Biol. Evol.* 12, 759-772

44 Macey, J.R. *et al.* (2004) Phylogenetic relationships among amphisbaenian reptiles based on complete mitochondrial genome sequences. *Mol. Phylogenet. Evol*. 33: 22-31

45 Pääbo, S. *et al.* (1991) Rearrangements of mitochondrial transfer RNA genes in marsupials. *J. Mol. Evol.* 33, 426-430

46 Scouras, A. and Smith, M.J. (2001) A novel mitochondrial gene order in the crinoid echinoderm *Florometra serratissima*. *Mol. Biol. Evol.* 18, 61-73

47 Smith, M.J. (1993) The phylogeny of echinoderm classes based on mitochondrial gene arrangements. *J. Mol. Evol.* 36, 545-554

48 Black, W.C. and Roehrdanz, R.L. (1998) Mitochondrial gene order is not conserved in arthropods: prostriate and metastriate tick mitochondrial genomes. *Mol. Biol. Evol.* 15, 1772-1785

49 Macey, J.R. *et al.* (2000) Evolution and phylogenetic information content of mitochondrial genomic structural features illustrated with acrodont lizards. *Syst. Biol*. 49, 257-277

50 Bridge, D. *et al.* (1992) Class-level relationships in the phylum Cnidaria: Evidence from mitochondrial genome structure. *Proc. Natl. Acad. Sci. U. S. A.* 89, 8750-8753

51 Flook, P. *et al.* (1995) Homoplastic rearrangements of insect mitochondrial tRNA genes. *Naturwissenschaften* 82, 336-337

52 Dowton, M. and Austin, A.D. (1999) Evolutionary dynamics of a mitochondrial rearrangement "hot spot" in the Hymenoptera. *Mol. Biol. Evol.* 16, 298-309

53 Lang, B.F. *et al.* (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* 387, 493-497

54 Kunisawa, T. (2001) Gene arrangements and phylogeny in the class proteobacteria. *J. Theor. Biol.* 213, 9-19

55 Qiu, Y.-L. *et al.* (1998) The gain of three mitochondrial introns identifies liverworts as the earliest land plants. *Nature* 394, 671-674

56 Nozaki, H. *et al.* (2003) Phylogeny of plastids based on cladistic analysis of gene loss inferred from complete plastid genome sequences. *J. Mol. Evol.* 57, 377-382

57 Palmer, J.D. *et al.* (2000) Dynamic evolution of plant mitochondrial genomes: Mobile genes and introns and highly variable mutation rates. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6960-6966.

58 Venkatesh, B. *et al.* (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10267–10271

59 Shimamura, M. *et al.* (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* 388, 666–670

60 Nikaido, M. *et al.* (1999) Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10261–10266

61 Nikaido, M. *et al.* (2001) Retroposon analysis of major cetacean lineages: the monophyly of toothed whales and the paraphyly of river dolphins. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7384–7389.

62 Takahashi, K. *et al.* (2001) Retroposition of the AFC family of SINEs (short interspersed repetitive elements) before and during the adaptive radiation of cichlid fishes in Lake Malawi and related inferences about phylogeny. *J. Mol. Evol.* 53, 496-507

Glossary

BAC vector: Bacterial Artificial Chromosomes, a type of vector for cloning large inserts of DNA, typically 100-200 kb.

Character: any trait being considered for phylogenetic analysis, for example “type of limb”.

Character state: the condition of the character for a particular taxon, for example “branched limbs”.

Clade: a group of organisms that shares a common ancestor to the exclusion of the other considered taxa.

Cladistics: a school of thought that emphasizes reconstructing evolutionary relationships solely through recognizing clades and by a set of specific criteria for inference.

Evolutionary convergence: identical changes occurring independently in separate lineages.

Evolutionary reversion: the change of a character state back into an earlier occurring state.

Fosmid vector: a type of vector for cloning DNA, typically in fragments of about 40 kb.

Genome-level character: features of a genome or its products that can be assessed for phylogenetic analysis other than the linear sequences of nucleotides or amino acids.

Homoplasy: a pattern of character states that supports an alternative to the true, accepted or most parsimonious evolutionary tree, generally caused by evolutionary convergence or reversion.

Internode: the parts of an evolutionary tree that connect one internal node to the next, i.e. not terminal branches.

Intron: portions of a gene internal to exons that are transcribed, but are enzymatically removed from the mRNA prior to its translation into protein.

Numt: mitochondrial DNA that is incorporated into the nuclear genome as a pseudogene.

Orthologous: homologous genes in two of more organisms that are related only by lineage splitting, i.e. not by gene duplication.

Outgroup: one or more taxa assumed for the purpose of analysis to be less closely related to the other organisms being considered than they are to one another.

Phenetics: phylogenetic reconstruction based on measures of similarity.

Retroelements: mobile genetic elements that use an RNA intermediate in their transposition.

Synapomorphy: a character state that can be inferred by outgroup comparison or other methods to be derived specifically for a particular clade.

Terminal branch: the parts of an evolutionary tree that lead only to the taxon considered, i.e. not internode branches.

Universal genetic code: a misnomer based on an earlier, incorrect belief that all genomes share the same code for specifying amino acids from triplets of nucleotides (although it is nearly universal).

Figure Legends

Figure 1. A clade of arthropods is demonstrated by a shared transposition of a tRNA gene. The circular mitochondrial gene map of the chelicerate *Limulus polyphemus* is shown as graphically linearized at an arbitrarily chosen point. That of the insect *Drosophila yakuba* differs by only a single gene transposition, of *trnL(yaa)*, as indicated by the arrow. The genes for tRNAs are indicated here by the single letter code for the corresponding amino acid, plus the two genes for tRNAs that specify leucine are differentiated by their anticodons in parentheses. Underlining indicates reverse transcriptional orientation. The transposition of *trnL(yaa)* appears to have occurred a single time in the evolutionary history of animals, at the base of a clade that includes many major groups of hexapods and crustaceans and of the (traditionally recognized) phylum Pentastomida. This clade excludes two major arthropod groups, Chelicerata and Myriapoda, as well as the Onychophora, since these groups retain the ancestral gene arrangement seen in outgroup phyla such as Phoronida and Mollusca. The ancestral position of *trnL(yaa)* is shown at the bottom of the figure and the derived position at the base of the clade it defines. This has led to radical revisions of the way we interpret the evolution of the arthropod body plan [12-14]. Numbers in superscript indicate the taxa that are represented by these higher level categories according to the following key. [1]*Armillifer armillatus* , [2]*Argulus americanus*, [3]seven genera, [4]*Hutchinsoniella macracantha*, [5]*Daphnia pulex*, *Artemia franciscana*, *Triops cancriformis*, *T. longicaudatus*, [6]17 genera, [7]*Speleonectes tulumensis*, [8]223 genera, [9]*Gomphiocephalus hodgsoni*, *Onychiurus orientalis*, *Tetrodontophora bielanensis*, *Podura aquatica*, [10]*Japyx solifugus*, [11]*Lithobius forficatus*, *Narceus annularus*, *Thyropygus* sp., [12]six genera, [13]*Euperipatoides leuckartii*, [14]*Phoronis architecta*, [15]10 genera. More information, including a complete set of citations for this data, can be found at <http://evogen.jgi.doe.gov/top_level/organelles.html>.

Figure 2. Identical gene arrangements do not necessarily signal close evolutionary relationships.

Trees (a), (b) and (c) describe all potential relationships among one outgroup and the three ingroup species 1, 2 and 3; trees (d), (e) and (f) do likewise for another outgroup and species 4, 5 and 6. Upper-case letters represent the relative order of sets of genes. Valid conclusions about evolutionary relationships can be made only by using an outgroup to show that shared arrangements are derived for a subset of taxa, not by simply uniting those that are most similar. This is shown by comparing the conclusions possible for species 1-3 versus those for species 4-6. None of trees (a) through (c) can be preferred, even though (c) unites those most similar, because each requires exactly two changes (marked by arrows). However, in the second set of species, species 4 matches the arrangement of the outgroup taxon, thus tree (f) requires only one change, versus two required for either tree (d) or (e). There are two alternative reconstructions for trees (d) and (e), in each case requiring two changes, either having one taxon revert to the previous condition of the outgroup as marked by the arrows, or convergent rearrangements in two taxa as marked by the asterisks. All are less parsimonious than the reconstruction in tree (f).

Figure 3. Relationships among primates as outlined by patterns of retrotransposition.

This summarizes the results of three studies that used patterns of shared retroelement integration to place the tarsiers with monkeys and hominoids (rather than with strepsirhines) [18] and to clarify the relationships among the Strepsirhini [19] and the apes [20]. Each relationship is signaled by one or more unique retroelement integrations as indicated by numerals.

Box 1. Problems with molecular sequence comparisons and the search for an ideal set of characters

Although comparisons of the sequences of nucleotides or amino acids have solved many phylogenetic relationships, others have remained equivocal even with the accumulation of large amounts of data. Here are some potential explanations for these failures: (i) Multiple nucleotide or amino acid substitutions might have occurred at a single site, obscuring any accumulated signal; (ii) Convergent or parallel substitutions might have occurred among different lineages as there are only four (for nucleotides) or 20 (for amino acids) possible character states, exacerbated by convergent biases in base composition [32]; (iii) The analysis might show artifactual association of the more rapidly changing lineages, the so-called "long branch attraction" artifact [33]; (iv) In some cases, gene copies that are not orthologous might be inadvertently compared among various lineages owing to ancestral gene duplications followed by differential losses, or owing to incomplete sampling; (v) Differing views of scientists on alignments, exclusion sets and weighting schemes are difficult to arbitrate based on objective criteria and can lead to radically different phylogenetic reconstructions; and (vi) Perhaps the most difficult problems are when the time of shared ancestry is short relative to the subsequent time of divergence, where there has been little opportunity to accumulate signal and ample time for it to have been erased.

In the search for a set of characters to supplement these studies, we imagine several ideal traits. These characters would be ubiquitously distributed and unambiguously homologous among the taxa of interest. They would exist in a large number of complex states so that we can recognize similar, but non-identical changes, and so that identical changes are unlikely to occur independently. They would be very unlikely to revert to an earlier state, be neutral with regard to

natural selection so as to minimize the chance of convergent change, and change at a rate appropriate for the relationships to be addressed.

Box 2. Features of genomes that could potentially be used as phylogenetic characters

Complete genome sequences are being determined for a great number of diverse organisms. This data set contains many features that can be compared for phylogenetic analysis, potentially including: (i) gene content; (ii) presence versus absence of particular biochemical pathways; (iii) the relative arrangements of genes; (iv) movements of genes among intracellular compartments (i. e., plastid, mitochondrion, nucleus); (v) insertions of segments of DNA, including transposons and numts [34,35]; (vi) variation in intron positions; (vii) subunit structures of various proteins; (viii) components of multi-unit complexes, such as the ribosome, splicosome, DNA replication machinery, or oxidative phosphorylation enzymes; (ix) secondary structures of rRNAs or tRNAs; (x) details of genome level processes, such as the DNA rearrangements that generate antibody diversity; (xi) deviations from the 'universal' genetic code [36]. Some, as referenced here, have been used already; others may be in the future.

Box 3. Summary of some evolutionary relationships that have been supported by genome-level characters

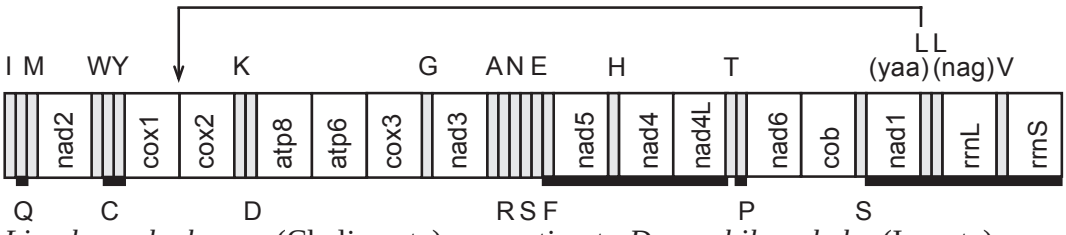
Comparisons of mitochondrial gene arrangements have provided strong support toward the resolution of several long-controversial phylogenetic relationships, including: (i) The superphylum Eutrochozoa includes cestode platyhelminths [37] and the phylum Phoronida [38]; (ii) Sipuncula is closely related to Annelida rather than to Mollusca [39]; (iii) Annelida is more closely related to Mollusca than to Arthropoda [40]; (iv) Arthropoda is monophyletic and, within this phylum, Crustacea is united with Hexapoda to the exclusion of Myriapoda and Onychophora [12-14]; (v) Pentastomida is not a phylum, but rather a type of crustacean, and joins with Cephalocarida and Maxillopoda to the exclusion of other major crustacean groups [13]. Furthermore, there are several groups of animals whose monophyly is generally accepted, but where this is reinforced by their sharing derived sets of mitochondrial gene arrangements, including Aves [41], Vertebrata [42], Crocodylidae [43], biped amphisbaenians [44], Marsupialia [45], Echinodermata (wherein interclass relationships are also addressed) [46,47], metastriate ticks [48], and acrodont lizards [49]. Also, the groups Hydrozoa, Scyphozoa and Cubozoa are inferred to group within Cnidaria to the exclusion of Anthozoa based on their having a linear mtDNA structure [50].

Can homoplasy occur in mitochondrial gene arrangements? A 1998 study by Mindell *et al.* [41] is frequently cited (e.g. [7]) as an example of homoplasious gene rearrangement in birds. In fact, ref. [41] reports no gene rearrangements whatsoever, and all birds examined to date have an identical rearrangement of all 37 mitochondrial genes. See [8] for a discussion of this. However, there are some examples of homoplasious rearrangements in mitochondrial genes – in amphisbaenian reptiles [44], orthopteran insects [51] and hymenopteran insects [52]. As

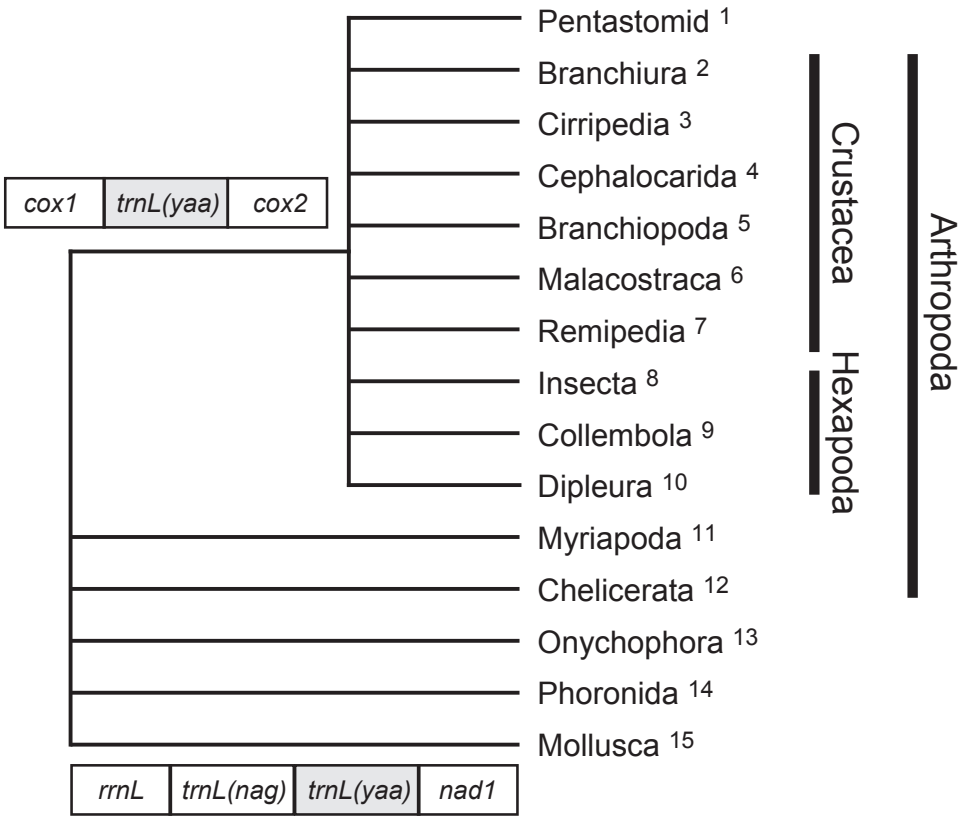
explained in [8], special caution is in order when the shared rearrangements are an exchange of nearest neighbor tRNA genes or of genes immediately downstream of an origin of replication; each of these examples of homoplasy are in one of these two categories.

Comparisons of genomic DNA arrangements have also provided powerful evidence of relatedness. For example, the most convincing evidence to date of the endosymbiotic origin of mitochondria comes from the observation of shared gene arrangements among some bacteria and the mitochondrial genome of the jakobid protist *Reclinomonas americana* [53]. Other comparisons of the relative arrangements of genes has clarified the evolutionary history of those proteobacteria that have been completely sequenced [54].

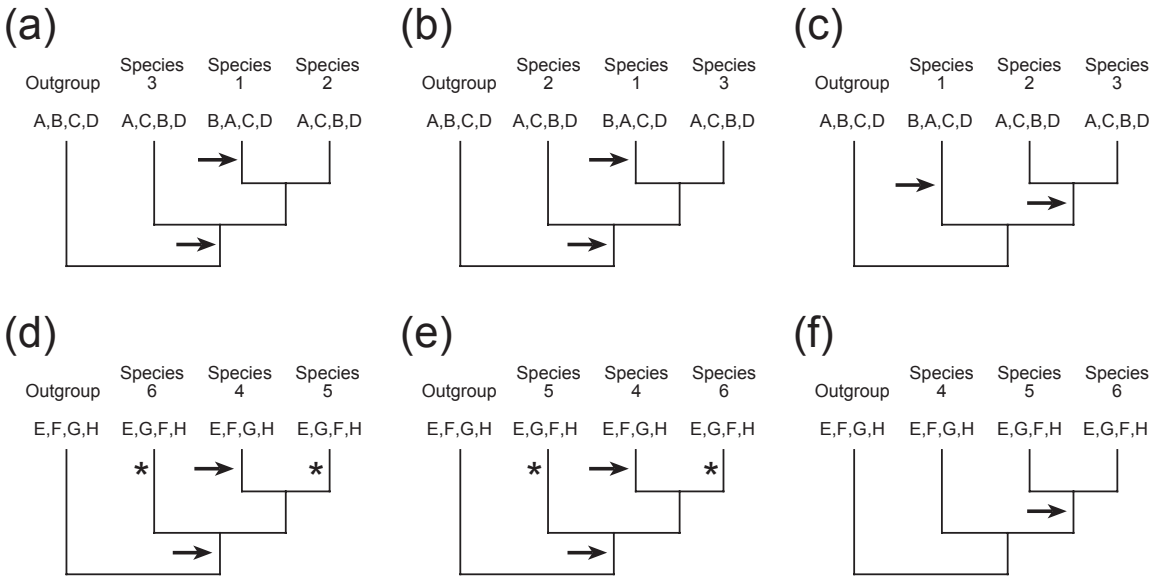
Other genome-level characters have been used for phylogenetic inference. Patterns of intron gain identified liverworts as the earliest branching group of land plants [55] and other gene losses [56] and translocations [57] have further defined plant relationships. Comparing gene membership within *Hox* clusters has addressed the relationships among several animal phyla [27]. A study of the presence of spliceosomal introns supports the monophyly of Actinopterygia and clarifies several relationships within the group, including the basal position of bichirs [58]. The positions of retroelements have been used to show that cetaceans (e.g. whales, dolphins and porpoises) are monophyletic [59], that toothed and baleen whales unite into a clade, that hippopotamuses are the sister group to cetaceans and that camels are the most basal group of cetartiodactyls [60]. The relationships among cetacean lineages have been clarified, including the conclusion that river dolphins are polyphyletic [61] and the relationships among groups of cichlid fish have been determined [62]. It remains to be seen how reliable these various types of characters will be and how common their usage will become.



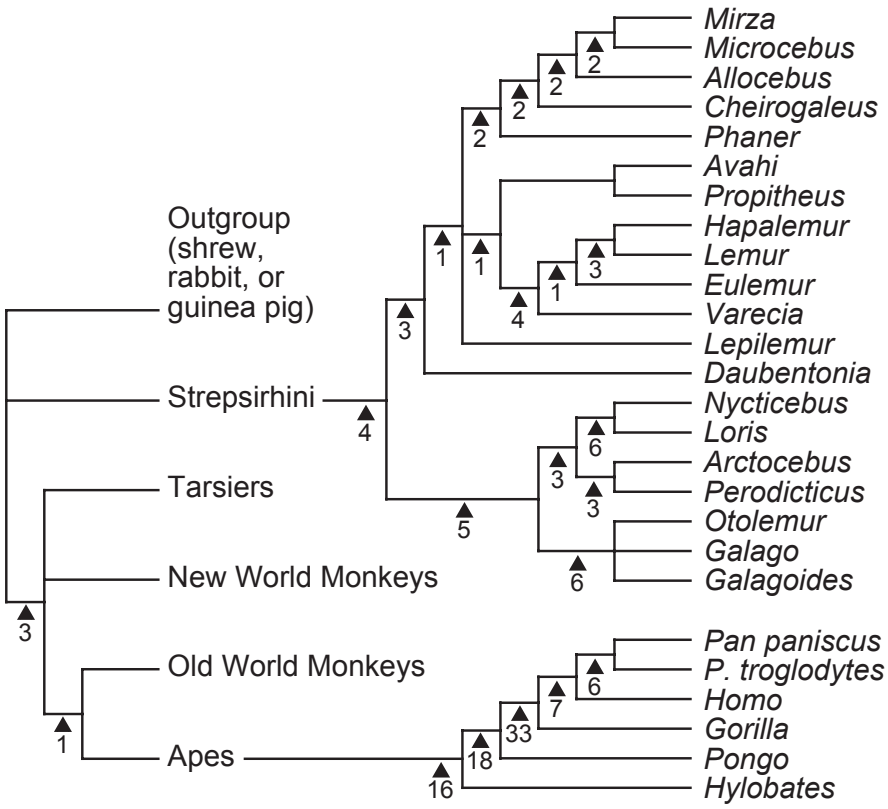
Limulus polyphemus (Chelicerata) converting to *Drosophila yakuba* (Insecta)



Boore Figure 1



Boore Figure 2



Boore Figure 3