Journal of Phylogenetics & Evolutionary Biology

Editorial

Open Access

Phenotypic and Evolutionary Distances in Phylogenetic Tree Reconstruction

Luciano Brocchieri*

Department of Molecular Genetics and Microbiology and Genetics Institute, University of Florida, Gainesville, FL, USA

Advances in sequencing technology and the resulting deluge of molecular sequence data have provided vast opportunities to study the evolution of gene and protein families together with the phylogenetic relations of the species harboring them. Each family of homologous sequences can provide hundreds or thousands of characters, that is, all homologous sites making up a sequence alignment, that are a potential source of valuable information for phylogenetic tree reconstruction. Moreover, molecular sequences have other advantages over, say, morphological characters. Among these, is a natural, unambiguous definition of "evolutionary distance", which allows estimating the amount of evolutionary divergence of sequences, represented in phylograms. This precise definition of evolutionary distance stimulated the development over the last 35 years of evolutionary models that provide means to estimate evolutionary relations and to develop theories on how molecular sequences evolve, connecting phylogenetics to evolutionary biology [1-5].

The evolution of molecular sequences is most often analyzed based on a multiple sequence alignment that identifies across a set of homologous sequences all homologous positions (sites), each represented by a column in the alignment. An alignment is treated as a collection of independent "characters" (alignment positions) with four possible states in the case of nucleic acid sites and twenty states in the case of protein sites. Base mutations or amino acid substitutions are the elementary evolutionary events and evolutionary distanceis defined as the number of elementary substitution events that occurred during the time of divergence of two homologous characters, irrespective of the direction of time. The evolutionary distance between two sequences of aligned positions is simply the average of these counts over all positions, i.e., a normalized count of elementary substitution events. As long as it can be assumed that all characters followed the same evolutionary path (i.e., no differential later gene transfer and recombination among genes), it is irrelevant to the analysis whether two characters (sites) belong to the same gene (protein) or to different concatenated genes (proteins). In a phylogenetic tree, the length d of a branch separating two sequences represented at its end points represents the estimate of their evolutionary distance. If this estimate is based on the multiple alignment of n positions, nd estimates the total integer number of substitution events that occurred during the evolutionary divergence of the two sequences. Thus, by definition evolutionary distances are additive, and the evolutionary distance (number of substitution events) between sequences connected through multiple branches is the sum of the evolutionary distances (substitution events) represented by each branch, i.e. the sum of their lengths (patristic distances). The problem of inferring evolutionary trees is essentially the problem of estimating counts of substitution events.

In molecular phylogenetics evolutionary distances are not only unambiguously defined but can also be estimated given a measurable *phenotypic distance* between sequences, the sequence dissimilarity. Furthermore, we notice that phenotypic distance between sequences is also defined using elementary evolutionary events (substitutions), as the most parsimonious evolutionary distance between sequences (the p-distance), i.e., the minimum number of elementary evolutionary operations needed to transform one sequence into the other. To estimate the number of substitutions that actually occurred in evolutionary history, a model of sequence evolution is needed to predict the effect of evolutionary distance on phenotypic distance. Probabilistic methods based on transition-rate matrices have been developed to capture the effect of the randomness of the mutational process and of short-term selection on long-term evolution. Although different models of how the evolutionary process depends on site and on amino acid or nucleotide type produce different inferences on the relation between evolutionary distance and phenotypic distance (dissimilarity), all models result in similar general properties of this relation (Figure 1A): evolutionary distance is described by an increasing convex function of phenotypic distance, with slope 1.0 at phenotypic distance p=0.0, and tending to infinity as phenotypic distance approaches an asymptotic value

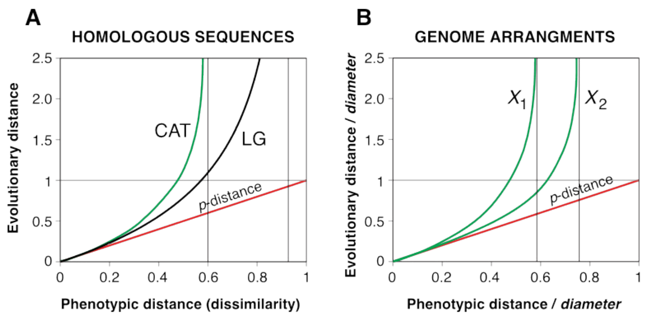

Figure 1A: In describing homologous-sequence evolution, point mutations (substitutions) accumulate generating increasing dissimilarity between sequences. The most parsimonious estimate of how many substitutions have accumulated (evolutionary distance) is called the p-distance, and coincides with the phenotypic distance (sequence dissimilarity). Probabilistic models (e.g., green and black curves) predict that the p-distance (red line) underestimates evolutionary distance. Different models generate different estimates of evolutionary distance and predict that sequence dissimilarity converges to different asymptotic values as evolutionary distance tends to infinity. LG is the relation between evolutionary distance and sequence dissimilarity predicted by the amino acid substitution model of Le and Gascuel [5]. CAT represents a possible outcome of a model with position-dependent equilibrium frequencies [4].

Figure 1B: In genome evolution phenotypic distance can also be defined by the most parsimonious (minimum) distance between genome arrangements, here normalized by the *diameter* of the set of distances, i.e., the maximum distance between two elements of that set (see text). As evolutionary distance tends to infinity the expected phenotypic distance between genome arrangements converges asymptotically to values corresponding to the average of the phenotypic distances between random permutations. Different choices of possible elementary permutation operations X (e.g., X_1, X_2) correspond to different asymptotic behavior of phenotypic distance (see text).

***Corresponding author:** Luciano Brocchieri, Cancer and Genetics Research Complex 2033 Mowry Rd, Gainesville, FL 21610, USA, Tel: +1 352 273 8131; E-mail: lucianob@ufl.edu

Received November 28, 2013; **Accepted** November 28, 2013; **Published** December 07, 2013

Citation: Brocchieri L (2013) Phenotypic and Evolutionary Distances in Phylogenetic Tree Reconstruction. J Phylogen Evolution Biol 1: e106. doi:10.4172/2329-9002.1000e106

corresponding to the expected dissimilarity of unrelated sequences. Three regions in the domain of phenotypic distance can be described (Figure 2). The first region, the "parsimony zone", corresponds to incipient evolutionary differentiation, when evolutionary distance can be predicted with sufficient accuracy by phenotypic distance, hence using a parsimonious estimate. A second region, the "probabilistic zone", corresponds to stochastic accumulation of multiple substitutions and result in significant increasing under-estimation of evolutionary distance by phenotypic (parsimonious) distance. The third region, the "mutational saturation zone", includes the interval of phenotypic distances that do not differ significantly from asymptotic expectations. When the observed phenotypic distance does not significantly differ from this upper limit, evolutionary distance can only be inferred to be above a minimum value and there is*mutational saturation* between the two sequences.

The same conceptual framework can be applied to describe the evolution of other types of characters, for example to describe the evolution of the arrangement of genes in genomes. "Global mutations" such as transpositions, reversals, duplications, deletions, and combinations of these, have been used to describe genome evolution, dating back to the work of Dobzhansky and Sturtevant [6]. See also the work of Palmer and Herbon [7] on the importance of such genome scrambling in the mitochondrial genomes in cabbage and turnip. The principle of parsimony is often advocated in the analysis of genome evolution [8,9]. However, as in the case of sequence evolution, parsimonious evolution of genome structure is difficult to justify beyond the case of incipient evolution. Indeed, the assumption of parsimony can be shown to lead to contradicting results even in simple cases of genome rearrangement as, for example, illustrated by the evolution of herpes virus genomes (Figure 3). As in molecular-sequence analysis, probabilistic approaches describing the evolution of genome rearrangements have also been proposed [10-13]. Following the framework outlined for molecular sequences, "evolutionary distance" between genome rearrangements can also be defined based on some natural set of elementary evolutionary events (e.g., transpositions, reversals, interchanges, and their variations), defining how a gene or a block of genes can be rearranged in a genome. Furthermore, based on the same set of operations, the "phenotypic distance" between genome arrangements can be defined as the minimum number of elementary evolutionary events required for transforming one genome arrangement into another. We can then ask what is the expected relation between phenotypic distance and evolutionary distance in terms of genome rearrangements, what is the expected phenotypic distance between genomes when their evolutionary distance tends to infinity (unrelated arrangements), and what is the range of phenotypic distances of genome arrangements for which we can expect evolutionary information to be preserved. The stochastic accumulation of elementary events will generally result in evolutionary distances between two genomes that differ from their phenotypic distance (Figure 1B), similarly to what occurs in sequence evolution. In the case of genome rearrangements and depending on the particular set of elementary operations by which permutations can be obtained, phenotypic distances are bounded by an upper value, called the *diameter*, which is defined as the maximum of the phenotypic distances between all possible pairs of genome arrangements. The expected phenotypic distance between genome arrangements when their evolutionary distance approaches infinity, corresponding to the expected phenotypic distance between unrelated genomes, is the average distance between random pairs of arrangements. Evolutionary distance and phenotypic distance of a pair of genome arrangements, as well as their maximum(diameter) and average (asymptotic) values, can be quite difficult to calculate and their analytical solutions are not available for all types of rearrangements. For example, a solution for the minimum distance between rearrangements is available for inversions [14] and for block rearrangement [15] distances. For the latter, estimates of average phenotypic distance and of the diameter are also available [16]. Estimates for all distances can however be obtained by computer simulations [17]. The usefulness of comparing genome rearrangements to identify the evolutionary relations of genomes will also depend on how large is the variance of the estimate of the phenotypic distance of two random genome arrangements. This variance will affect how much evolutionary distance can be accumulated before the phenotypic distance of diverging genomes reaches the "mutational saturation zone" of genome rearrangements (Figure 1B) when phylogenetic information is lost.

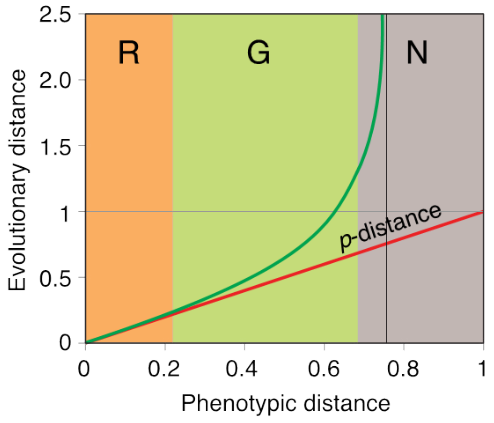

Figure 2: Regions of phenotypic distance corresponding to different estimates of evolutionary distance. R: "Parsimony zone", region where evolutionary distance is accurately approximated by phenotypic distance *d*. G: "Probabilistic zone", region where the *p*-distance under-estimates evolutionary distance. N: "Mutational saturation zone", region where the evolutionary distance cannot be estimated because of loss of phylogenetic information.

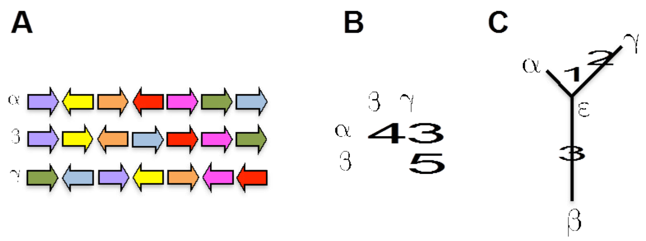

Figure 3: A. Herpes virus genomes of the Alpha, Beta and Gamma subfamilies are characterized by subfamily-specific arrangements of seven universally conserved blocks of genes, represented here in different color.

B. The minimum number of signed transpositions needed to transform any two arrangements (parsimonious pair-wise evolutionary distances) can be determined.

C. There is a unique unrooted evolutionary tree consistent with these parsimonious pair-wise distances. Branch lengths indicate the number of elementary operations required along each branch to generate the calculated pair-wise distances assuming parsimony. The parsimonious tree exists only if at least one signed permutation ε exists with evolutionary distance from the observed permutations α, β and γ corresponding to the respective branch lengths. It can be shown that such permutation does not exist, contradicting the hypothesis of parsimonious evolution.

The same framework cannot be applied to characters for which elementary evolutionary events, and hence a definition of phenotypic and evolutionary distance, is problematic. For example, it may be difficult to define what the elementary steps in the evolution of multi-dimensional morphological characters are, and thus what the evolutionary and phenotypic distances between these characters should be. Molecular features that have been proposed as markers of phylogenetic relations, such as genomic signatures [18], may also be difficult to interpret in terms of evolutionary events. For these types of characters, it is unclear whether dendrograms based on some definition of similarity can be interpreted as phylograms of evolutionary relations.

Acknowledgements

This work is supported by NIH Grant 5R01GM87485-2.

References

1. Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press. Oxford and New York.
2. Felsenstein J (2004) Inferring Phylogenies. Sinauer Associates.
3. Gascuel O (2005) Mathematics of Evolution and Phylogeny. Oxford University Press. Oxford and New York.
4. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21: 1095-1109.
5. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. MolBiolEvol 25: 1307-1320.
6. Dobzhansky T, Sturtevant AH (1938) Inversions in the chromosomes of Drosophila pseudoobscura, Genetics 23: 28-64.
7. Palmer JD, Herbon LA (1988) Plant mitochrondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol 28: 87-97.
8. Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3: 2.
9. Pevzner P, Tesler G (2003) Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes. Genome Res 13: 37-45.
10. Durrett R, Nielsen R, York TL (2004) Bayesian estimation of genomic distance. Genetics 166: 621-629.
11. Larget B, Kadane JB, Simon DL (2005) A Bayesian approach to the estimation of ancestral genome rearrangements. Mol Phylogenet Evol 36: 214-223.
12. Larget B, Simon DL, Kadane JB, Sweet D (2005) A Bayesian analysis of metazoan mitochondrial genome arrangements. Mol Biol Evol 22: 486-495.
13. York TL, Durrett R, Nielsen R (2002) Bayesian estimation of the number of inversions in the history of two chromosomes. J Comp Biol 9: 805-818.
14. Bader DA, Moret BME, Yan M (2001) A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. J Comp Biol 8: 483-491.
15. Christie DA (1996) Sorting Permutations by Block Interchanges. Information Processing Letters 60: 165–169.
16. Bóna MR, Flynn R (2008) The average number of block interchanges needed to sort a permutation and a recent result of Stanley. Information Processing Letters 109: 927–931.
17. Bourque G, Pevzner PA (2002) Genome-scale evolution: Reconstructing gene orders in ancestral species. Genome Res 12: 26–36.
18. Campbell A, Mrázek J, Karlin S (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc Natl Acad Sci USA 96: 9184-9189.