**LDDist**: A Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences

Mikael Thollesson

Dept. molecular Evolution, Evolutionary Biology Centre and Linnaeus Centre for Bioinformatics, Uppsala University, Norbyvägen 18C, SE-752 36 UPPSALA, Sweden
ABSTRACT

Summary: LDDist is a Perl module implemented in C++ that allows the user to calculate LogDet pair-wise genetic distances for amino acid as well as nucleotide sequence data. It can handle site-to-site rate variation by treating a proportion of the sites as invariant and/or by assigning sites to different, presumably homogenous, rate categories. The rate-class assignments and invariant proportion can be set explicitly, or estimated by the program; the latter using either of two different capture-recapture methods. The assignment to rate categories in lieu of a phylogeny can be done using Shannon-Wiener index as a crude token for relative rate.

Availability: LDDist and its companion Perl script PLD are freely available at http://artedi.ebc.uu.se/molev/software/LDDist.html

Contact: mikael.thollesson@ebc.uu.se

Background

Most substitution models used in phylogenetic inference, either in maximum likelihood applications or for transforming sequence data into pair-wise distances, assume stationarity. For stationarity in base frequencies of DNA sequences, it is often evident that this assumption is violated, as various compositional biases are common in DNA sequences. In many cases, these violations lead to inconsistency in the phylogenetic inference (e.g., Saccone et al., 1989; Penny et al., 1990; Hasegawa & Hashimoto, 1993; Steel et al., 1993), where sequences with similar compositional bias tend to form clades. To overcome the problem with non-stationarity in base composition, Steel and coworkers (Steel, 1994; Lockhart et al., 1994; as LogDet distances) and Lake (Lake, 1994; as Paralinear distances) independently proposed a distance measure based on the determinant of the divergence matrix (a matrix
comprising the relative frequencies of all nucleotide, or amino acid, pairs) between two sequences.

The general notion has been that protein sequences are relatively free of compositional bias (e.g., Loomis & Smith, 1990; Lockhart et al., 1992), and LogDet distances have mainly been used as a tool in analyses of DNA sequences; there are several implementations of LogDet distances for DNA sequences. Foster et al. (1997), however, showed that protein sequences are sometimes biased as well, and a subsequent study (Foster & Hickey, 1999) showed that such a bias also can affect phylogenetic inference and lead to misleading results. Foster and Hickey (1999), as well as Waddell and co-workers (Waddell et al., 1999) applied LogDet to amino acids, albeit using unpublished software. The aim of the present work is to provide a tool that allows LogDet distances to be calculated for amino acid as well as DNA sequences, while handling site-to-site rate variation.

Implementation

Recent large-scale genome projects have created a need for phylogenetic inferences that are insensitive to compositional bias (or to test for its effects; e.g., Lockhart et al., 1999; Lockhart & Cameron, 2001), and software that can efficiently process a large number of alignments (data sets). Perl is commonly the preferred language to handle genome projects, and the thus the calculations are implemented (in C++) as a module accessible from Perl, LDDist. The application script PLD written in Perl provides a front-end, and serves as an example, for utilizing LDDist. PLD takes an alignment in one of a number of popular formats (e.g., clustal, fasta, NEXUS) from standard input and produces a NEXUS file with pair-wise distances and commands on standard output. No phylogenetic analysis is done by PLD (nor by LDDist), but the NEXUS file is subsequently used as input for PAUP* (Swofford, 2002), which provide the
phylogenetic analyses (e.g., by minimum evolution or neighbor-joining). *LDDist* can do bootstrap resampling of the original alignment to generate pair-wise distance matrices to assess the sampling variation in the sequences. Options to *PLD* (exclusion of sites, rate classes, bootstrap, input format) are provided as command line switches.

The original expression of the LogDet distance is

\[
d_{xy} = \frac{1}{r} \ln \frac{\det F_{xy}}{\det P_x \det P_y}
\]

where \( r \) is the number of character states (\( r=20 \) for protein, \( r=4 \) for DNA/RNA), \( F_{xy} \) is an \( r \times r \) divergence matrix for sequences \( X \) and \( Y \), and \( P_x \) and \( P_y \) are diagonal matrices of the character-state frequencies in sequences \( X \) and \( Y \) respectively. This expression will only give a distance proportional to the number of changes when amino acid residue frequencies are equal (i.e., 0.05 and 0.25 respectively). A modification (e.g., Tamura & Kumar, 2002) is used in *LDDist*, where the distance is

\[
d_{xy} = \frac{1}{r} \ln \frac{\det F_{xy}}{\det P_x \det P_y}
\]

and \( \square \) are the frequencies of the different states (amino acids/nucleotides).

A problem is that a state may be absent in one or more sequences, in which case the determinants will be zero and consequently the distance undefined. The best way to deal with this is yet to be established, but *LDDist* will set the corresponding diagonal element to a small value (1/2 before normalizing). The behavior of LogDet distances calculated in this way when the number of missing states increases also remains to be explored.

Another difficulty with LogDet distances is to account for site-to-site rate heterogeneity (Swofford et al., 1996; Waddell et al., 1999). Waddell (1995) showed
that by subtracting an appropriate proportion of invariant sites from the diagonal elements of $F_{xy}$, LogDet distances can become nearly additive even if the distribution of rates across sites follows a continuous distribution (e.g., a gamma distribution).

When using $LDDist$ a fraction of the constant sites (sites with the same amino acid/nucleotide in all sequences) can be excluded from the calculation as invariant (sites not free to vary, e.g., due to biological constraints). $LDDist$ provides two methods to estimate the proportion of invariant sites using capture-recapture methods. One is the method proposed by Sidow et al. (1992) based on capture-recapture within the codon. It is only available for amino acid sequences and the universal genetic code is assumed. The other is the method proposed by Steel et al. (2000) based on capture-recapture of quartets among the sequences, which is applicable to DNA as well as amino acid sequences.

Another approach to accommodate rate variation is to classify sites into a few, presumably homogenous, rate classes, apply the LogDet transformation to each class separately, and finally sum the contribution from each class to obtain the final pairwise distance (Swofford et al., 1996). This option is implemented in $LDDist$, where each site can be assigned to one of any number of rate classes, and may be used in conjunction with the invariant sites exclusion. It is worth pointing out, however, that LogDet is a transformation that needs quite long sequences to give good results and that dividing the sequence among several rate classes will increase the required sequence length. The number of rate classes should thus be kept small, and the number of sites in each as large as possible.

How, then, to assign sites to rate classes? There are some well-known methods, for example by maximum likelihood or maximum posterior probability, although they are not easily calculated and need an a priori phylogeny. The preferred way is to use other available software and provide the rate classes explicitly to $LDDist$; PLD can
read a character vector of the same size as the alignment, representing rate classes for each of the sites.

To assign states independent of a particular phylogeny (thus confounding rate variation and phylogenetic signal), the Shannon-Wiener information index (Shannon & Weaver, 1949; Wiener, 1949) is available as a simple token of relative rate in LDDist (for examples of application of SW index to sequences, see Thollesson, 1999; Xia et al., 2003). This index is

\[ H_n = \sum_{i=1}^{N} (p_i) \log_2(p_i) \]

where \( p_i \) is the relative frequency of state \( i \) (\( N=4 \) or \( N=20 \)) at site \( n \). The range of \( H \) values for the alignment is divided in \( c \) equally wide classes, and each site is assigned to one of them based on its \( H_n \) value.

Finally, I would like to encourage readers to explore the behavior and shortcomings of LogDet distances on real, large scale, data sets showing different amino acid and nucleotide frequency biases.

ACKNOWLEDGEMENTS

Financial support for this work was received from the Linnaeus Centre for Bioinformatics. The input from Björn Canbäck and the comments from two anonymous referees are gratefully acknowledged.

REFERENCES


Urbana: The University of Illinois Press.


Waddell, P.J., Cao, Y., Hauf, J. and Hasegawa, M. (1999) Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid invariant sites LogDet plus site stripping, to detect internal conflicts in the data, with special
