

BARCODING ARTICLE

FINGERPRINT: visual depiction of variation in multiple sequence alignments

MELANIE LOU and G. BRIAN GOLDING

Department of Biology, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4L8

Abstract

There is a lack of programs available that focus on providing an overview of an aligned set of sequences such that the comparison of homologous sites becomes comprehensible and intuitive. Being able to identify similarities, differences, and patterns within a multiple sequence alignment is biologically valuable because it permits visualization of the distribution of a particular feature and inferences about the structure, function, and evolution of the sequences in question. We have therefore created a web server, FINGERPRINT, which combines the characteristics of existing programs that represent identity, variability, charge, hydrophobicity, solvent accessibility, and structure along with new visualizations based on composition, heterogeneity, heterozygosity, d_N/d_S and nucleotide diversity. FINGERPRINT is easy to use and globally accessible through any computer using any major browser. FINGERPRINT is available at <http://evol.mcmaster.ca/fingerprint/>.

Keywords: cytochrome *c* oxidase I, DNA barcoding, fingerprint, multiple sequence alignment, polymorphism, sequence diversity

Received 12 April 2007; revision accepted 11 June 2007

Introduction

The mitochondrial gene, cytochrome *c* oxidase subunit I (COI), is the terminal enzyme in the electron transfer chain that transfers electrons to molecular oxygen without forming reactive oxygen species (Ludwig *et al.* 2001). It helps form the electrochemical gradient across the inner mitochondrial membrane by pumping positively charged particles across it (Ludwig *et al.* 2001). COI is a vital player in generating energy and is found broadly across many taxonomic categories. The Barcode of Life Initiative has employed COI as the standard gene because it is able to discriminate between many closely related animal species (Hebert *et al.* 2003) and there is evidence to suggest that it also works well in algae (Saunders 2005), arthropods (Smith *et al.* 2005), fish (Ward *et al.* 2005) and some plants (Kress *et al.* 2005). Identifying sequence changes in homologous sites provides insights about the structure, functional genomics, and evolution of a protein. Although some tools are currently available through the Barcode of

Life Database for COI analysis, it is a continuing goal of the project to develop tools that can analyse and display data effectively.

There are various graphical multiple alignment editors, such as CLUSTAL_X (Thompson *et al.* 1997), SEAVIEW (Galtier *et al.* 1996), and JALVIEW (Clamp *et al.* 2004), that display an alignment in its entirety. The problem is that it becomes difficult to summarize the characteristics or diversity of a site relative to other sites within a multiple sequence alignment. To qualitatively analyse up to 1000 sequences or more at lengths of over 1000 residues is very tedious, time-consuming and difficult. To aid in such a task, there are a variety of multiple alignment shading programs available: ALSCRIPT (Barton 1993), ESPRIPT (Gouet *et al.* 1999), BOXSHADE, AMAS, WEBLOGO (<http://weblogo.berkeley.edu/>) (Crooks *et al.* 2004), SEQUENCE SIMILARITY PRESENTER (Fröhlich 1994) and T_EXSHADE (Beitz 2000). Unfortunately, most of these programs require download and installation of software, support complicated documentation, impose a fee or limit the number of sequences allowed in the input file. Furthermore, most of these programs focus on providing sequence-by-sequence representations and not alignment overviews. With the continued advancement in technology, increasing amounts of sequence data are

Correspondence: G. Brian Golding, Fax: 905-522-6066; E-mail: Golding@McMaster.ca

becoming readily available which spurs the need for more visualization software.

In this study, we introduce FINGERPRINT, a web server application that produces diagrams called fingerprints. A fingerprint is a horizontal bar made up of coloured or grey-scale vertical lines representing an overview of a desired feature in a sequence or in a set of aligned sequences. The concept of the alignment fingerprint was first introduced by Fröhlich (1994) in his SEQUENCE SIMILARITY PRESENTER and was subsequently adopted and updated by Beitz (2000) in his T_EX-based alignment shading package. Although these programs do produce fingerprints, only one feature is available for representation or the user is required to learn how to format documents in L_AT_EX to use the shading package, respectively. With new developments on five features, our tool provides options for a total of 11 distinct types of fingerprints, each depicting a different feature or 'flavour' of variation, and requires little to no overhead in learning how to use the program. The fingerprint concept has been incorporated into an online web interface, thus making it globally accessible via any major web browser. By default, information regarding the number of sequences and the average branch length of the aligned sequence set is given to provide a crude estimate of the significance of the fingerprint and a confidence level in the data presented.

Although the development of this tool was geared towards identifying diversity in COI barcodes, FINGERPRINT can be applied to a wide variety of data sets from any sequence data. Overall, FINGERPRINT is an effective tool to quickly and intuitively view the similarities, differences, and patterns in a multiple sequence alignment. The human eye can quickly assimilate these patterns, making data exploration much easier.

System and methods

FINGERPRINT was written using PHP, Perl, PostScript and the PHYLIP (Felsenstein 1989) suite of programs. It was tested with Internet Explorer, Konqueror, and Mozilla.

Algorithm and implementation

FINGERPRINT is available online freely; no registration or download is necessary. As input, the user can choose to upload a single file or multiple files containing a sequence or a set of aligned sequences in FASTA format. Depending on the preferences of the user, the output can be placed in a single PDF file or multiple PDF files which can be viewed in Acrobat Reader (free downloadable software) or any other PDF viewer.

The tool is currently capable of producing 11 different types of fingerprints, each depicting a particular feature or 'flavour' of variation. The fingerprint is a consensus overview of the desired feature within the aligned sequences.

Composition and heterogeneity

In a *composition* fingerprint, each residue is represented by its own colour. This fingerprint depicts the unique composition of elements encoded by a sequence or a unique consensus of a set of sequences which can be used to differentiate species based on the colouring and pattern of the residues (Fig. 1A).

With regard to a consensus composition fingerprint, there is a loss of information since the tool represents the residues with the highest frequency of occurrence. To prevent this loss of information, an alternate presentation is encoded. Each possible residue at a given site corresponds to a distinct coloured percentage of the vertical line drawn to represent a site. The heterogeneous composition of an alignment is viewed using a *heterogeneity* fingerprint. For example, invariable sites (represented by only one residue) are represented by one colour that extends for the entire length of the vertical line representing that particular site; the colour is determined by the residue. If, at a particular site, one residue occurs with a frequency of 0.25 and the second occurs with a frequency of 0.75, then the former colour will represent 25% of the height of the drawn line, and the latter will represent the remaining 75% (Fig. 1B).

Identity, variability, heterozygosity, and nucleotide diversity

The diversity at sites possessing more than one residue is quantified and graphically depicted in different types of fingerprints. An *identity* fingerprint differentiates between invariant (identical residues) and variable (more than one residue possible) sites (Fig. 1C). More information about the variable sites is obtained in a *variability* fingerprint. The variability of a site is quantified by considering the number of possible residues occurring at a site and is coloured accordingly. Thus, sites with the highest variability, are coloured black; in contrast, invariant sites are coloured white (Fig. 1D1). Depending on user preference, the opposite colour scheme can be selected as a preference (Figs 1D2). Sites existing between these two extremes are shaded/coloured accordingly.

Measures of diversity are calculated and graphically depicted in a *heterozygosity* fingerprint; this calculates the expected heterozygosity measure according to the equation:

$$1 - \sum_{i=1}^m x_i^2$$

(Li & Graur 1991), where x_i is the frequency of the i^{th} residue at a particular site. The value can also be interpreted as the probability that two residues chosen at random are different from each other.

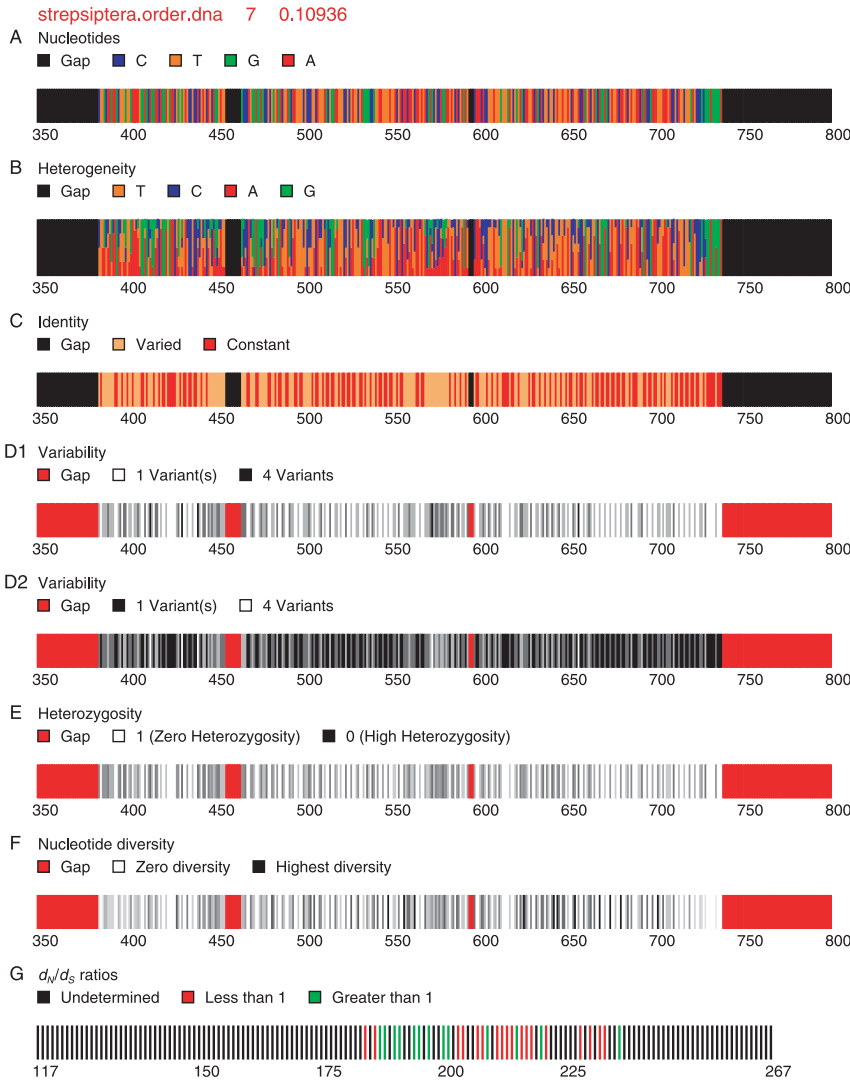


Fig. 1 Nucleotide fingerprints based on the cytochrome *c* oxidase I (COI) gene from the order Strepsiptera (twisted-wing parasites). A. Composition; B. Heterogeneity; C. Identity; D1. Variability (black); D2. Variability (white); E. Heterozygosity; F. Nucleotide diversity; G. d_N/d_S Ratio. All fingerprints were constructed using the same input file.

Highly variable sites possess high heterozygosity measures; those with the highest heterozygosity measures are coloured black. In contrast, invariant or invariable sites (one to a few residues) possess low heterozygosity measures; these sites are coloured white or close to it (Fig. 1E). The heterozygosity measure, however, may not be accurate for nucleotide sequences due to the more extensive variation at the DNA level over large sequence lengths (Li & Graur 1991). For nucleotide sequence data, the nucleotide diversity measure is calculated for each site using the equation:

$$\sum_{ij} x_i x_j \pi_{ij}$$

(Li & Graur 1991), where x_i and x_j are the frequencies of the i^{th} and j^{th} residues at a particular site, respectively, and π_{ij} is either 1 or 0 if there is or there is no difference between the i^{th} and j^{th} residues, respectively. Like the heterozygosity

fingerprint, the *nucleotide diversity* fingerprint lies on a grey scale where sites possessing high nucleotide diversity measures are coloured black or close to it, and sites with low nucleotide diversity measures are coloured white or close to it (Fig. 1F).

d_N/d_S ratio

To gain some insights about the type of selective forces in operation, FINGERPRINT calculates the d_N/d_S ratio for each codon (triplet of nucleotides) within a sequence or set of sequences. The d_N/d_S fingerprint maps possible sites of purifying, neutral, and adaptive evolution (Yang 1997; Fig. 1G).

Note that this algorithm is computationally extensive and may take time to complete. Also beware that this algorithm makes use of a simple neighbour-joining (NJ) tree that could be easily be improved; hence, these

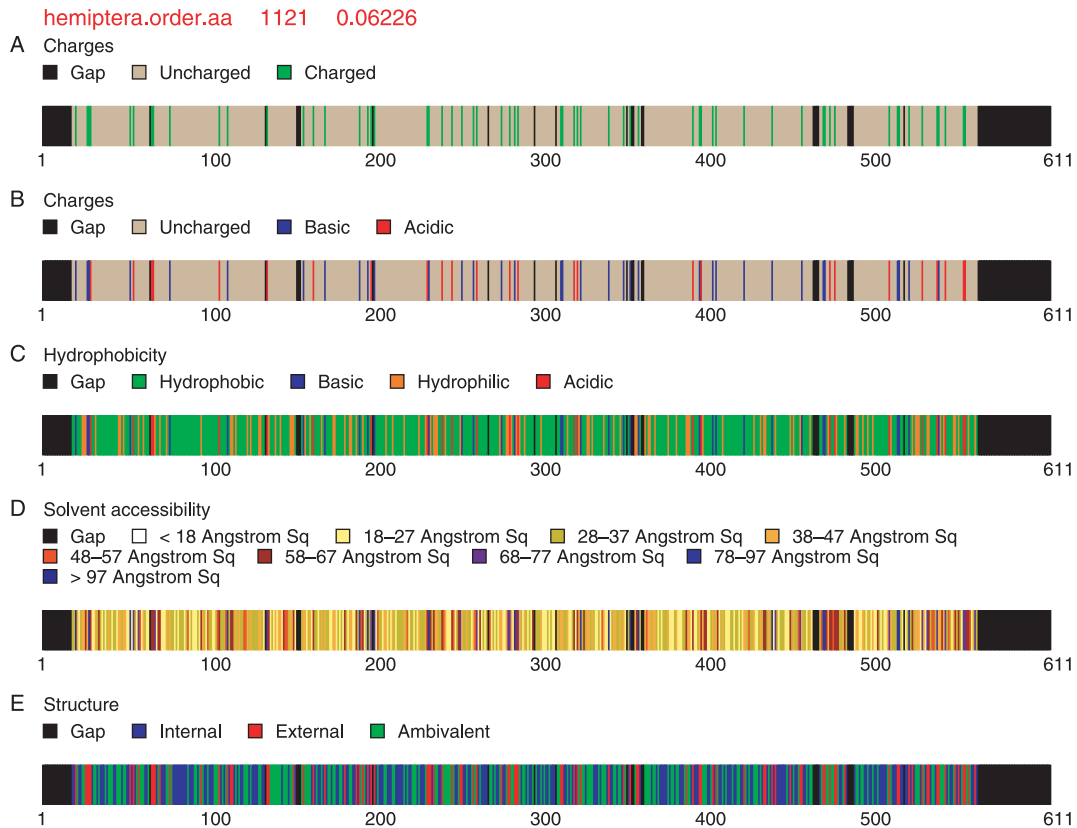


Fig. 2 Amino acid fingerprints based on the COI gene from the order Hemiptera (true bugs). A. Charges; B. Charges (Acidic and Basic); C. Hydrophobicity; D. Solvent accessibility and E. Structure. All fingerprints were constructed using the same input file.

results should be used only in a data exploration framework.

Charge, hydrophobicity, solvent accessibility, structure

The definitions for residue groupings, charge, structure, hydrophobicity, and solvent accessibility, were taken from Beitz (2000). The *charge* fingerprint identifies sites that are charged and uncharged (Fig. 2A). The user may choose to differentiate the charged sites as either acidic or basic (Fig. 2B). A *hydrophobicity* fingerprint categorizes sites as being acidic, basic, hydrophobic, or hydrophilic (Fig. 2C). In a *solvent accessibility* fingerprint, each residue is categorized according to experimentally determined solvent accessibilities based on the position that such a residue is usually found in a folded protein (Fig. 2D). A *structure* fingerprint identifies sites that are usually localized in the core (internal), on the surface (external) or neither of a globular protein (Fig. 2E). Similar to the *composition* fingerprint, these features work best for a data set consisting of one sequence. Given a multiple sequence data set, the residue with the highest frequency of occurrence is used to represent that site.

Managing fingerprint appearance

For publication purposes, the user has the option of manipulating several features associated with the appearance of the fingerprint. The *FINGERPRINT* assumes that the first residue in the sequence is indexed as the first position. Alternatively, the user has the option of specifying the first residue position, if it does not start at 1, and the last residue position, if not all the sites are to be represented. The *FINGERPRINT* program gives the user the option of selecting the range of sequence to be shown; the result is a 'zoomed-in' view of the desired portion of the sequence. All the fingerprints in Fig. 1 depict the nucleotide sequence in the range of 350–800 nucleotides. The height of the fingerprint is adjustable but must be larger than 0.1 inch. If no height is given, it is set to 1 inch by default. With the exception of the *heterogeneity* fingerprint, whose minimum height is intrinsically set to 0.5 inch, all other fingerprints shown in Fig. 1 are shown at a height of 0.3 inch. Each label can be either hidden or displayed in the final output. While the label serves as a means of identification, labels identifying the number of sequences and the average branch length also serve as a measure of

the meaningfulness of the output; these measures are located next to the input file name within the output file(s) in red (Fig. 1). The output of the FINGERPRINT is, by default, written to a single PDF file. Within the output file, output from each input file is identified by the input file name. Alternatively, the user can select multiple file output in which case, output from each distinct input file is placed into its own PDF file.

Average branch length

Trees are constructed using the NJ algorithm (Saitou & Nei 1987) based on Kimura 2-parameter distances (Kimura 1980). Average branch lengths are calculated as the total tree length of the NJ tree divided by the number of branches.

Results

For illustrative purposes, FINGERPRINT was applied to 9195 lepidopteran sequences that were annotated by their genus and species designations. For each sequence, the appropriate family name was determined; subsequently, the sequences were partitioned by family. Composition fingerprints revealed very similar fingerprints made distinct by subtle changes throughout the length of the sequences (e.g. Fig. 3A). Despite the compositional similarity, the

variability, heterozygosity and nucleotide diversity (e.g. Fig. 3B) fingerprints revealed distinct patterns of variation between families. Within each family, these three types of fingerprints were similar with respect to the location of the sites exhibiting diversity; however, the extent of diversity differed between corresponding sites among the three representations. Generally, sites showing diversity had lower values in the heterozygosity fingerprints relative to the corresponding nucleotide diversity representations which showed higher values for the same sites. The variability fingerprints possessed sites with values between the two extremes; this is expected since the colouring scheme is based on how variable a site is relative to sites of minimum and maximum variability.

Displaying the number of sequences and average branch length for each fingerprint proved to be worthwhile as these values helped measure a fingerprint's robustness. In the lepidoptera data, fingerprints depicting little to no diversity could mean that the family of sequences are highly conserved or it could mean nothing at all depending on the number of sequences used or their level of sequence divergence. In the cases presented, the large number of sequences would support the former interpretation. Furthermore, taking into account the average branch length helps yield further insights as to the credibility of the input data. An average branch of length of 0 or -1

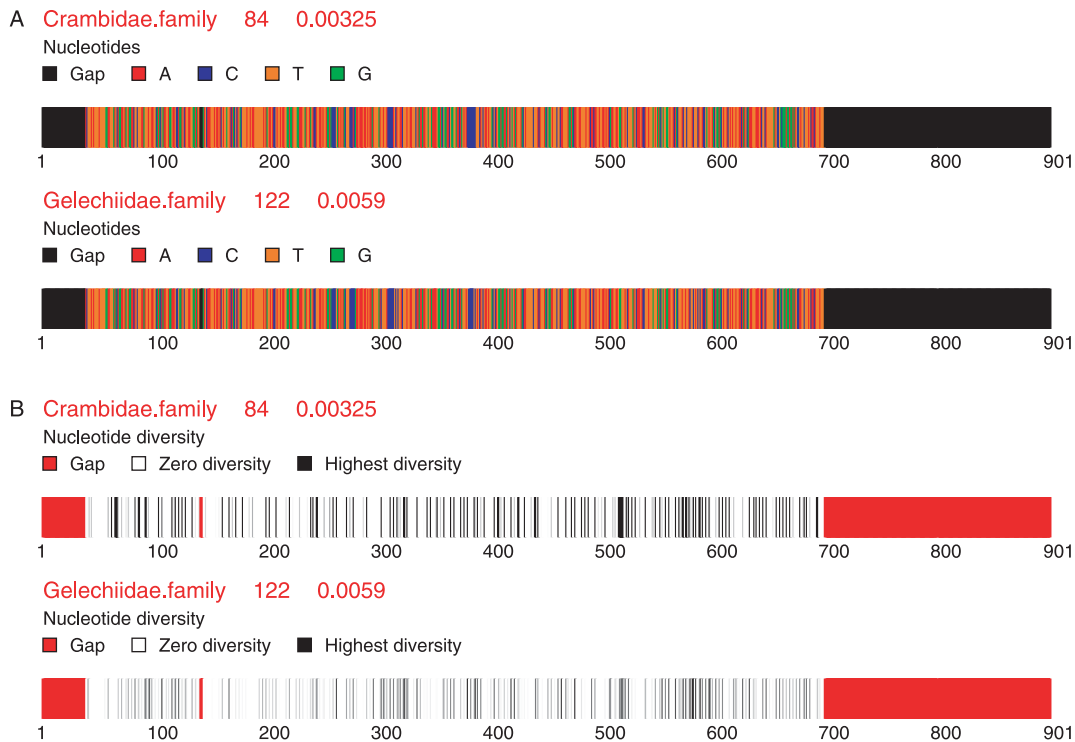


Fig. 3 Composition (A) and nucleotide diversity (B) fingerprints of two arbitrary lepidopteran (butterfly) families: Crambidae and Gelechiidae.

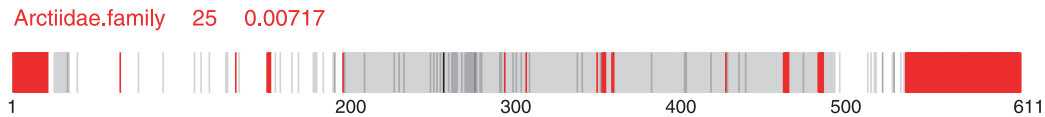


Fig. 4 An unexpected application of FINGERPRINT: it is able to catch alignment errors.

would indicate that the sequences are identical copies of each other. On the other hand, if the average branch length is of reasonable value, this would suggest a family of sequences worthy of further analysis.

Discussion

Beitz's fingerprint inspiration stemmed from Fröhlich (1994). The output of Fröhlich's SEQUENCE SIMILARITY PRESENTER resembles that of our *variation* fingerprint, except that our representation provides the option of representing sites of high variability (termed by Fröhlich as sites that lack identity) as either white or black. The FINGERPRINT web server combines the standard characteristics of fingerprinting with new technological developments to produce a tool that is better equipped to accommodate the needs of the biological community. In addition to similarity, functional and variability shading, fingerprints based on composition, heterogeneity, diversity (heterozygosity and nucleotide diversity) measures and d_N/d_S ratios are now available. FINGERPRINT is computer- and browser-independent and easy to use. The output is compact, intuitively understandable, and is well suited for providing a quick overview of alignments consisting of one or more sequences.

The output is written in PostScript which is used to create high-quality vector-based text and graphics. Vector-based graphics do not possess unnecessary detail in visual representations of information, thus reducing file sizes, yet superior resolution is maintained because the full resolution of the display device (printer or monitor) is exploited. Since many fingerprints can be created from single and multiple file input, the output maintains a consistent appearance that is easy to reproduce.

In addition to using the FINGERPRINT as a tool for identifying different types of variation, it may also be used to catch alignment errors. With reference to a fingerprint constructed using amino acid sequence data for the Lepidopteran family, Arctiidae (Fig. 4), it can be seen that the colour is uniform across a portion of the sequence, thus indicating the possibility of an alignment error spanning the length of that region.

However, there are some caveats to be aware of. Although, each fingerprint is accompanied by values for the number of sequences and average branch length as an indication of robustness, these values are merely two measures of fingerprint reliability. It is the responsibility of

the user to follow up on the results depicted. The rate-limiting step of FINGERPRINT for most of the algorithms is the calculation of the average branch length. The number of pairwise distances that must be determined for the NJ tree increase rapidly with larger data sets.

In summary, FINGERPRINT is effective for identifying sequence variation and for preparing high resolution, intuitive graphics for presentation.

Acknowledgements

This work was supported through funding to the Canadian Barcode of Life Network from Genome Canada through the Ontario Genomics Institute, NSERC, and other sponsors listed at www.BOLNET.ca.

References

- Barton GJ (1993) ALSRIPT: a tool to format multiple sequence alignments. *Protein Engineering Design Protein and Selection*, **6**, 37–40.
- Beitz E (2000) T_XSHADE: shading and labeling of multiple sequence alignments using L^AT_EX₂ epsilon. *Bioinformatics*, **16**, 135–139.
- Clamp M, Cuff J, Searle SM, Barton GJ (2004) The JALVIEW Java alignment editor. *Bioinformatics*, **20**, 426–427.
- Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WEBLOGO: a sequence logo generator. *Genome Research*, **14**, 1188–1190.
- Felsenstein J (1989) PHYLIP — *Phylogeny Inference Package* (Version 3.2). *Cladistics*, **5**, 164–166.
- Fröhlich KU (1994) SEQUENCE SIMILARITY PRESENTER: a tool for the graphic display of similarities of long sequences for use in presentations. *Computer Applications in the Biosciences*, **10**, 179–183.
- Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Computer Applications in the Biosciences*, **12**, 543–548.
- Gouet P, Courcelle E, Stuart DI, Metz F (1999) ESPRIT: analysis of multiple sequence alignments in PostScript. *Bioinformatics*, **15**, 305–308.
- Hebert PD, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **270**, S96–S99.
- Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.
- Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences, USA*, **102**, 8369–8374.
- Li W, Graur D (1991) *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, Massachusetts.

- Ludwig B, Bender E, Arnold S, Huttemann M, Lee I, Kadenbach B (2001) Cytochrome *c* oxidase and the regulation of oxidative phosphorylation. *Chembiochem*, **2**, 392–403.
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.
- Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1879–1888.
- Smith MA, Fisher BL, Hebert PD (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1825–1834.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **25**, 4876–4882.
- Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1847–1857.
- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, **13**, 555–556.