Cryptogamie, Algologie, 2010, 31 (4): 451-465
© 2010 Adac. Tous droits réservés

A comparative analysis of COI, LSU and UPA marker data for the Hawaiian florideophyte Rhodophyta: implications for DNA barcoding of red algae

Alison R. SHERWOOD [a,e*], *Thomas SAUVAGE* [a], *Akira KURIHARA* [a,b], *Kimberly Y. CONKLIN* [a,c] & *Gernot G. PRESTING* [d]

[a] *Botany Department, 3190 Maile Way, University of Hawaii, Honolulu, Hawaii 96822 USA*

[b] *Current address: Kobe University Research Center for Inland Seas, 1-1 Rokkodai, Kobe 657-8501 Japan*

[c] *Current address: Hawaii Institute of Marine Biology, P.O. Box 1346, Kaneohe, HI USA 96744*

[e] *Molecular Biosciences and Bioengineering, 1955 East-West Rd., University of Hawaii, Honolulu, Hawaii 96822 USA*

Abstract – A total of 290 florideophyte samples from the Hawaiian Rhodophyta Biodiversity Survey, spanning 17 orders of red algae, were sequenced for the mitochondrial COI DNA barcode, partial nuclear LSU rRNA gene, and plastid UPA rRNA marker. COI was A-T rich (> 60% overall) in comparison to the two rRNA markers, and was also the least conserved, with the third codon position showing the most substitution. Saturation was reached at F84 distances of approximately 0.11 for COI and 0.30 for LSU, but not at all for UPA, and some differences were found when orders or groups of families were individually examined. Rates of sequencing success for the three markers ranged from 46.8% for COI to 64.7% for UPA and 79.6% for LSU, indicating substantial differences in ease of data acquisition. Concatenation of marker sequences starting with the least saturated marker (UPA), and adding in order of degree of saturation (UPA+LSU, followed by UPA+LSU+COI) resulted in a strong increase in bootstrap support with neighbor-joining analysis, indicating that some phylogenetic utility can be gained from barcode-like sequences that are obtained for biodiversity surveys.

COI / diversity / DNA barcode / Hawaii / LSU / Rhodophyta / UPA

Résumé – Le code-barres mitochondrial ADN COI, un segment partiel du gène rARN LSU et le marqueur plastidial rARN UPA, ont été séquencés sur un total de 290 échantillons de Florideophytes (représentant 17 ordres) issus du projet "Hawaiian Rhodophyta Biodiversity Survey". Le marqueur COI est globalement plus riche en A-T (> 60 %) comparé aux deux autres marqueurs rARN, mais aussi le moins conservé, avec la troisième position des codons présentant une forte substitution. Sur la base d'un modèle F84, la saturation est atteinte à une distance d'environ 0,11 pour COI and 0,30 pour LSU, mais absente pour UPA, et certaines différences existent quand les ordres ou groupes de familles sont examinés individuellement. Le taux de réussite du séquençage pour les trois marqueurs varient de 46,8 % pour COI, 64,7 % pour UPA et 79,6 % pour LSU, indiquant des différences substantielles pour l'obtention de données. La concaténation successive des marqueurs, en démarrant du moins saturé (UPA), et en ajoutant LSU et COI par ordre de saturation (UPA + LSU, puis UPA + LSU + COI) conduit à une augmentation forte du soutien bootstrap par la méthode d'analyse du neighbor-joining, indiquant l'utilité phylogénétique de séquences courtes d'ADN ("barcode-like") obtenues dans le cadre de projets d'exploration de la biodiversité.

COI / diversité / code-barres ADN / Hawaii / LSU / Rhodophyta / UPA

* Correspondence and reprints: asherwoo@hawaii.edu

INTRODUCTION

With the official advent of DNA barcoding in 2003 (Hebert *et al.*, 2003) and the first assessment of the mitochondrial COI barcode for red algae in 2005 (Saunders, 2005), a number of studies have appeared using this method of species delimitation for algae (e.g. Robba *et al.*, 2006; Saunders, 2009; Clarkston & Saunders, 2010; Le Gall & Saunders, 2010; Sherwood *et al.*, in press). DNA barcoding has been demonstrated to be a powerful tool for detection of cryptic species, assignment of unknown samples to well-characterized taxa and delimitation of species complexes (Hebert *et al.*, 2003; Hebert *et al.*, 2004a; Newmaster & Ragupathy, 2009). The potential to revolutionize the way in which algal taxonomic and systematic research is performed is being realized with the population of the Barcode of Life Data Systems (BOLD) with algal COI sequences (11,922 for red algae alone as of 02 November 2010; *http://www.barcodinglife.com/*).

The Hawaiian Rhodophyta Biodiversity Survey was a four-year effort (years 2006-2010) to collect and characterize the red algae of the Hawaiian Archipelago, using a combination of morphological and molecular (i.e. DNA barcode) analyses. The principal aims of the study were to generate baseline molecular data for as many representatives of the flora as possible, and to provide an updated assessment of the Hawaiian red algae through DNA sequences linked to vouchered specimens and DNA extracts, with all data electronically available *via* a project database, the Hawaiian Algal Database (Wang *et al.*, 2009). The molecular focus of the project was on easily-amplifiable regions with corresponding universal primers (universal at least for red algae). Three short, barcode-length markers (i.e. short enough to be sequenced as a single read on an automated sequencer) were selected: the mitochondrial COI barcode region (Saunders, 2005), the "Y" fragment of the nuclear LSU rRNA gene (Harper & Saunders, 2001) and the rRNA plastid UPA marker (Presting, 2006; Sherwood & Presting, 2007). DNA barcoding of algae was in its infancy at the time of the survey's design (2004), and we aimed to include COI barcode data alongside other marker sequences, based on the earliest available protocols (Saunders, 2005). The intent of the marker selection was to represent the content of the three genomes of red algae and to establish well-populated molecular data sets that could be compared across a common set of identified and vouchered specimens.

With the recent conclusion of the Hawaiian Rhodophyta Biodiversity Survey, we now have the opportunity to critically examine and compare the sequence data from the three molecular markers. Over 2,400 DNA sequences were generated through the survey, and almost 300 florideophyte samples were successfully sequenced for all three markers. This data set represents a wide swath of the recognized red algal diversity in the Hawaiian Islands and is also broadly representative of tropical red algal diversity, given that approximately three-quarters of the world's tropical red algal genera are present in Hawaii (Abbott, 1999). Thus, the aims of the present study were (1) to evaluate the COI, LSU and UPA markers for ease of sequence acquisition, (2) to analyze sequence variation and substitutional saturation of the three markers at a variety of taxonomic levels, and (3) to compare a simple exploratory analysis using neighbor-joining (NJ) based on a concatenated alignment to trees built from smaller data sets in order to determine whether further inference can be made from the data. Observed clustering patterns on the UPA+LSU+COI concatenated tree are discussed in light of a recent assessment of the red algal tree of life (Verbruggen *et al.*, 2010) which was based on a heterogeneous data set obtained by data mining.

MATERIALS AND METHODS

The Hawaiian Algal Database (HADB) was queried for red algal accessions with sequence data for all three markers used in the Hawaiian Rhodophyta Biodiversity Survey: the mitochondrial barcode (COI), nuclear (LSU), and plastid (UPA) markers. A total of 290 accessions were returned, representing 17 florideophyte orders, 35 families, 79 genera and 102 named species. Downloaded sequences were aligned separately for each marker using Clustal X (Thompson *et al.*, 1997), and alignments were trimmed using MEGA (Tamura *et al.*, 2007). Fast evolving loops of LSU (indel rich regions) were removed because sequence homology could not be ascertained with confidence, leaving only stems for this marker. Similarly, a very short segment of UPA, which aligned ambiguously on the scale of the entire data set, had to be removed. Data files are available from the corresponding author.

Downloaded sequences (GenBank: HQ421677-HQ422584 (LSU), HQ420906-HQ421676 (UPA) and HQ422585-HQ423135 (COI)) and the complete list of survey accessions for which at least one sequence was obtained ($n = 1{,}168$) were used to calculate the percent success of sequencing for each marker at the familial level. This analysis consisted of scored (present or absent) accessions with a successful sequence from at least one of the three markers. Sequences for the UPA marker were generated with a single primer pair (Sherwood & Presting, 2007), as were those for the LSU marker (Conklin *et al.*, 2009), while those for the COI marker were generated either with the first published primer pair for red algae (Saunders, 2005) or with a newly designed reverse primer (R686 5'-CCACCWGMAGGATCAA-3'). Accessions with undetermined taxonomy at the ordinal and/or familial level based on morphology, but still included in the analyses, were assigned provisional taxonomy based on DNA sequence analysis where possible (data available upon request to the corresponding author). These accessions were not included in the percent success analysis described above.

Base pair frequencies and pairwise distances (as number of nucleotide differences) for the three markers and for the three codon positions of the COI marker were computed with the software package APE (Paradis *et al.*, 2004) and displayed using other packages in the R library (R Development Core Team, 2010). Violin plots were used to display pairwise distances, and these plots show data distribution in a way similar to boxplots, but illustrate a probability density of the data estimated by the kernel density method. Substitution saturation analyses of the entire data set as well as for individual orders (or groups of families, in the case of the Ceramiales) for each marker (and separately for each codon position for the COI marker) were performed using the program DAMBE v.5.2.6 (Xia, 2010), following the method of Phillipe *et al.* (1994) as described in Salemi (2003). Base frequencies for each marker (and codon position for COI) were calculated based on an average of the entire data set and were displayed using R.

Three separate exploratory NJ analyses were performed to examine the effect of marker concatenation on bootstrap values (each analysis with 2,000 replicates and based on maximum composite likelihood distances in MEGA) starting from UPA (the least saturated) to UPA+LSU and finally UPA + LSU + COI. Bootstrap support for each node was assessed and represented on the trees if $\geq 70\%$, with white indicating 70-79% support, grey indicating 80-89% support, and black indicating 90-100% support. Trees were visually compared to

assess differences in support based on the three data sets, and the number of nodes supported at the various levels was counted to compare support. Each bootstrap value was also evaluated for the taxonomic level(s) at which it provided support, and these data were also compared across the three analyses.

RESULTS

Data acquisition analysis – The percent success in obtaining target sequence data for the three markers is presented in Table 1. Overall percent success values ranged from 44.8% for the COI marker, to 67.4% for UPA, to 76.9% for LSU. In some taxonomic groups, however, success values may be artificial (e.g. for the Bonnemaisoniaceae, a focus on COI for phylogeographic study resulted in LSU and UPA not being attempted for all extracts), and some families were represented by low numbers of accessions (fewer than five), which may have skewed the results. Thus, adjusted means and standard deviations of success were calculated, removing these accessions from consideration (those indicated by * in Table 1), resulting in slightly adjusted success values of 46.8% for COI, 64.7% for UPA and 79.6% for LSU.

Nucleotide distances by marker – The pairwise nucleotide distances by marker, and by codon position for the COI marker, are illustrated in Fig. 1. Comparing the overall pairwise distances for all three markers (first three data displays from left to right) shows that the LSU marker is the most conserved (although ambiguously aligned regions were removed from the data set), followed by UPA, and that the COI marker has substantially higher pairwise distances than either of the first two. Examining the COI pairwise distances by codon position (final three data displays) shows that the third codon position accounts for the vast majority of divergence, followed by the first and second positions. Overall mean pairwise nucleotide difference was 3-4 times higher for COI than UPA or LSU, and the third codon position of COI had three times higher nucleotide differences than the first and second position.

Substitution saturation analyses – Transitional saturation of the COI (Fig. 2a), LSU (Fig. 2b) and UPA (Fig. 2c) data sets was visually examined by plotting the estimated number of transitions and transversions for each pairwise comparison against F84 genetic distance. Transitional saturation was reached (the point where transversions outnumber transitions) at F84 distances of approximately 0.11 for the complete COI marker data set, which represents approximately one fifth of the maximum divergence in the data set. Saturation for the LSU marker occurred very close to the maximum divergence for the data set (at approximately 0.30), and is thus negligible compared with COI. The UPA data set did not reach saturation. Analyses for each order with substantial representation (or families/groups of families, in the case of the Ceramiales), were also performed for each marker to detect deviations from the overall analysis (Table 2). Although the overall data sets for COI and LSU both indicated saturation, no LSU marker saturation was detected for the Bonnemaisoniales, Corallinales, Gelidiales, Gigartinales or Nemastomatales. All taxonomic groups appeared to reach saturation for COI at some point, although that value varied depending upon the group. For example, the Dasyaceae + Delesseriaceae *s.l.* (Ceramiales) were estimated to reach saturation at distances of only 0.06, while the Gelidiales and Gigartinales were estimated to reach saturation at or above distances of 0.20. Estimated saturated

Table 1. Percent success of obtaining target sequence data for the COI, LSU and UPA markers, by order and family. Families for which fewer than five accessions were obtained are indicated with *. Final row includes means and standard deviations

Order	*Family*	*COI*	*LSU*	*UPA*	*Total accessions*
Acrosymphytales	Acrosymphytaceae*	50.0	50.0	100.0	2
Batrachospermales	Batrachospermaceae	80.0	40.0	100.0	5
Bonnemaisoniales	Bonnemaisoniaceae	94.9	28.2	20.5	39
Bonnemaisoniales	Naccariaceae*	66.7	66.7	66.7	3
Ceramiales	Callithamniaceae	70.8	58.3	62.5	24
Ceramiales	Ceramiaceae	61.2	86.6	71.6	67
Ceramiales	Dasyaceae	56.3	81.3	70.3	64
Ceramiales	Delesseriaceae	23.8	52.4	66.7	21
Ceramiales	Rhodomelaceae	58.8	73.8	72.9	328
Ceramiales	Sarcomeniaceae*	0.0	100.0	0.0	1
Ceramiales	Spyridiaceae	23.4	71.9	56.3	64
Ceramiales	Wrangeliaceae	35.1	56.8	70.3	37
Colaconematales	Colaconemataceae	44.4	100.0	77.8	9
Corallinales	Corallinaceae	88.1	61.2	70.2	67
Corallinales	Hapalidiaceae*	0.0	100.0	100.0	1
Gelidiales	Gelidiaceae	21.4	100.0	32.1	28
Gelidiales	Gelidiellaceae	42.9	85.7	14.3	7
Gigartinales	Caulacanthaceae*	100.0	100.0	100.0	2
Gigartinales	Cystocloniaceae	69.0	79.3	75.9	29
Gigartinales	Dumontiaceae	50.0	100.0	90.0	10
Gigartinales	Gigartinaceae	14.3	100.0	71.4	7
Gigartinales	Gloiosiphoniaceae*	0.0	100.0	100.0	1
Gigartinales	Kallymeniaceae	0.0	85.7	71.4	7
Gigartinales	Phyllophoraceae	19.2	92.3	61.5	26
Gigartinales	Rhizophyllidaceae	50.0	75.0	87.5	8
Gigartinales	Solieriaceae	40.0	80.0	86.7	15
Gracilariales	Gracilariaceae	74.5	72.7	67.3	55
Halymeniales	Halymeniaceae	26.5	85.3	76.5	34
Hildenbrandiales	Hildenbrandiaceae	50.0	83.3	50.0	6
Nemaliales	Galaxauraceae	64.6	54.2	18.8	48
Nemaliales	Liagoraceae	41.9	71.6	73.0	74
Nemaliales	Scinaiaceae*	75.0	100.0	75.0	4
Nemastomatales	Nemastomataceae	28.6	100.0	42.9	7
Nemastomatales	Schizymeniaceae	83.3	83.3	83.3	6
Peyssonneliales	Peyssonneliaceae	5.9	94.1	64.7	17

Table 1. Percent success of obtaining target sequence data for the COI, LSU and UPA markers, by order and family. Families for which fewer than five accessions were obtained are indicated with *. Final row includes means and standard deviations *(cont'd)*

Order	*Family*	*COI*	*LSU*	*UPA*	*Total accessions*
Pihiellales	Pihiellaceae*	0.0	0.0	100.0	1
Plocamiales	Plocamiaceae*	50.0	75.0	25.0	4
Rhodymeniales	Champiaceae	58.3	75.0	41.7	12
Rhodymeniales	Faucheaceae*	50.0	100.0	100.0	2
Rhodymeniales	Lomentariaceae	100.0	80.0	40.0	5
Rhodymeniales	Rhodymeniaceae	50.0	81.3	75.0	16
Sporolithales	Sporolithaceae*	33.3	100.0	66.7	3
Thoreales	Thoreaceae*	0.0	0.0	100.0	1
	Total mean and standard deviation	44.8 ± 28	76.9 ± 24.8	67.4 ± 25.8	27.1 ± 51.8
	Adjusted mean and standard deviation	46.8 ± 23.9	79.6 ± 15.4	64.7 ± 20.2	36.8 ± 59.4

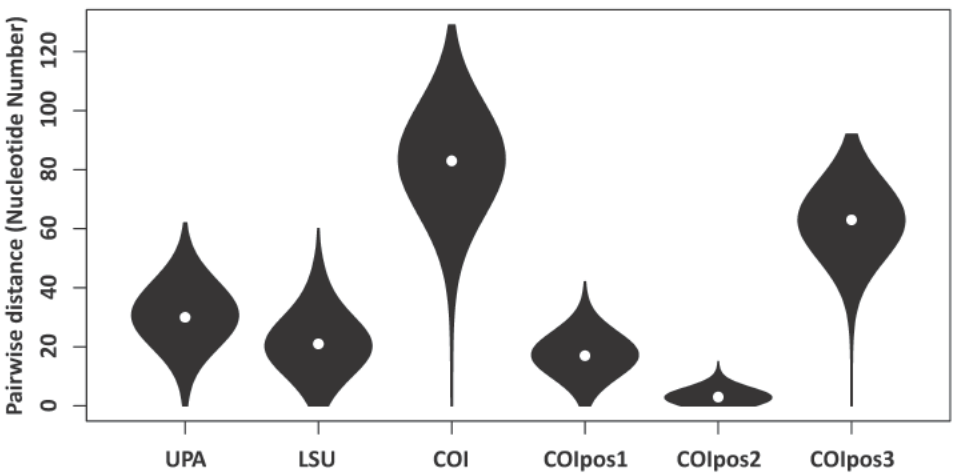

Fig. 1. Violin plot of pairwise distances (number of nucleotide differences) for the UPA, LSU and COI markers for 290 Hawaiian florideophyte red algae, and for each codon position of the COI marker. White circles indicate medians.

divergences for the LSU marker also ranged widely, from distances of 0.03-0.07 in the Gracilariales, Halymeniales and Nemaliales, to not at all for those mentioned previously.

Base frequencies per marker – The COI nucleotide base frequency analysis indicated that this protein-coding marker was A-T rich (> 60% overall, and > 80% for the third position) (Fig. 3). In contrast to this, the two rDNA markers (UPA and LSU) were more balanced in nucleotide composition, with both A-T and G-C percentages close to 50% (Fig. 3).

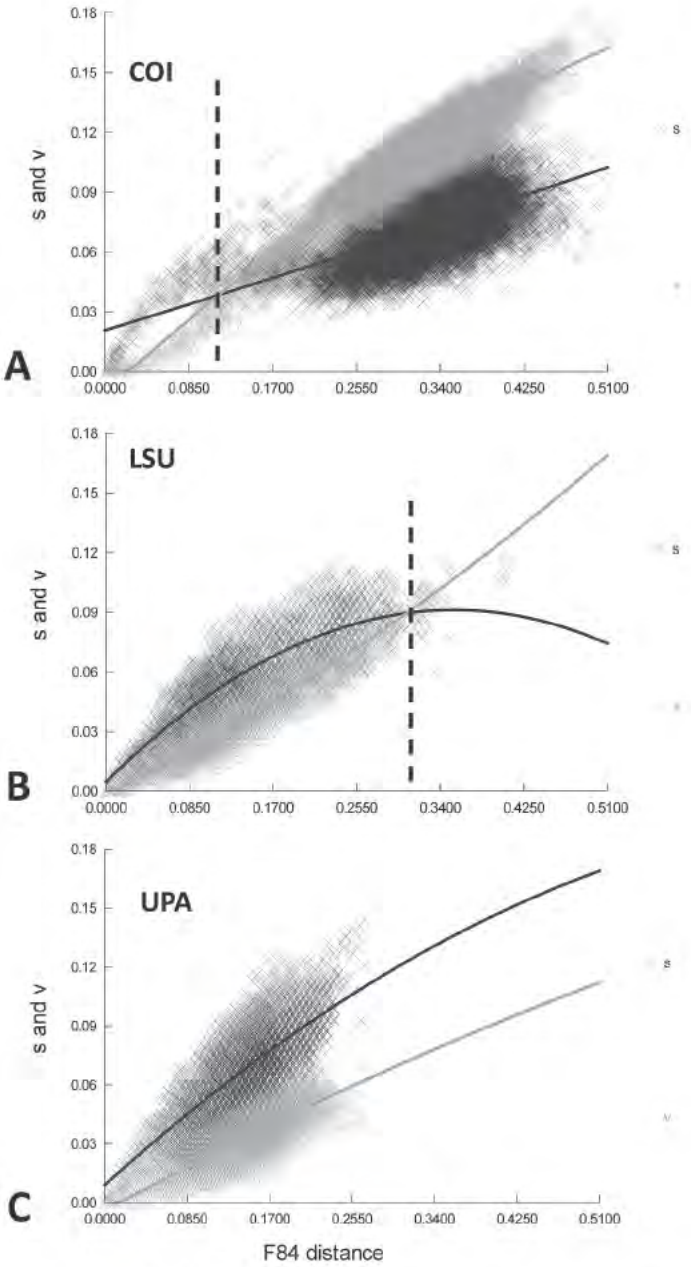

Fig. 2. Saturation curves for A) COI, B) LSU and C) UPA markers for the Hawaiian florideophyte red algae. The estimated number of transitions (indicated by "×") and transversions (indicated by triangles) for each pairwise sequence comparison was plotted against F84 genetic distance. Marker sequences are saturated at the point where transversions outnumber transitions, indicated by a dashed line (distances of approximately 0.11 for COI and 0.30 for LSU, with sequences not reaching saturation for UPA comparisons).

Table 2. Estimated saturation points (expressed as F84 distances) for Hawaiian florideophyte red algal sequences of the COI, LSU and UPA markers. Only taxonomic groups that had substantial representation in the Hawaiian Rhodophyta Biodiversity Survey are included here

Taxonomic group	*COI*	*LSU*	*UPA*
Bonnemaisoniales	0.19	none	none
Ceramiales (Rhodomelaceae)	0.12	0.34	none
Ceramiales (Ceramiaceae *s.l.*)	0.17	0.33	none
Ceramiales (Dasyaceae and Delesseriaceae *s.l.*)	0.06	0.50	none
Corallinales	0.01	none	none
Gelidiales	0.25	none	none
Gracilariales	0.20	0.07	none
Gigartinales	0.17	none	none
Halymeniales	Insufficient data	0.05	none
Nemaliales	0.12	0.03	none
Nemastomatales	0.19	none	none
Rhodymeniales	0.10	0.18	none

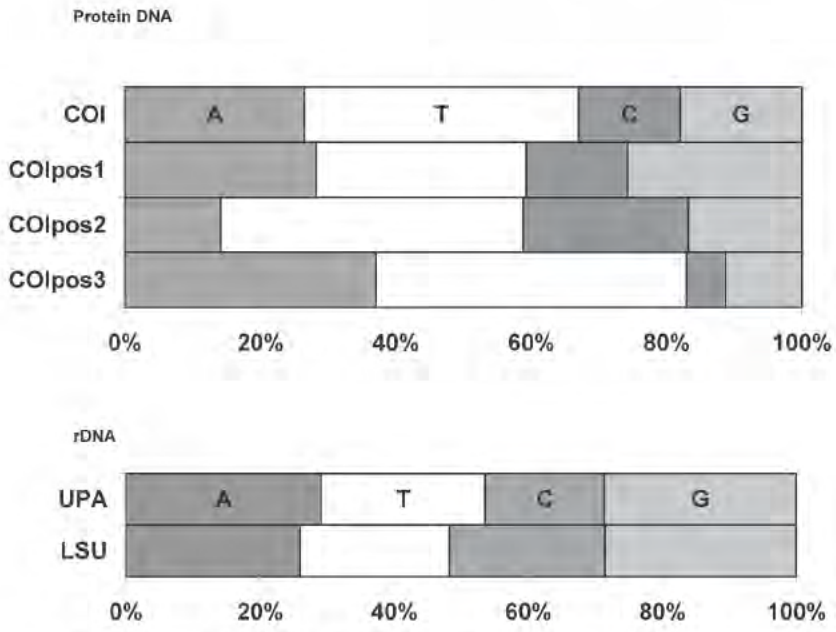

Fig. 3. Base frequencies for the COI (protein-coding marker; top panel) and UPA and LSU (rDNA markers, bottom panel) markers. The average percent composition of each marker (and marker position, for COI) by nucleotide is coded as follows: white = thymine (T), light grey = guanine (G), medium grey = adenine (A) and dark grey = cytosine (C).

Neighbor-joining analyses – Overall, nodal support increased with the addition of nucleotide characters from additional markers to the neighbor-joining analyses (Fig. 4a-c). The neighbor-joining analysis based on the UPA marker alone yielded a total of 96 nodes with ≥ 70% support while the UPA+LSU analysis yielded a total of 129 nodes and the UPA+LSU+COI analysis yielded 182 nodes (Fig. 4, Table 3). The greatest increase in nodal support was in the 90-100% category, with 64 nodes in that category in the first analysis, 93 in the second, and 133 in the fully concatenated data set (Fig. 4a-c, Table 3). Support values at different taxonomic levels (order, family, genus and species) also generally increased with the addition of data (Table 4), although many of the bootstrap values are at nodes that do not represent taxonomic groupings (Table 4, "other" column).

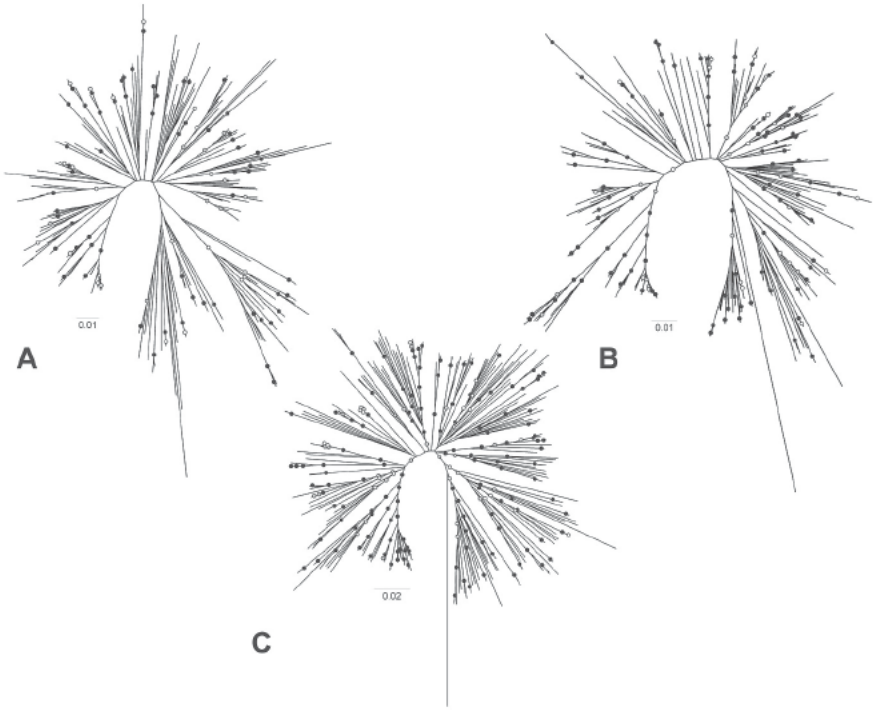

Fig. 4. Neighbor-joining trees of Hawaiian florideophyte red algae based on A) UPA sequences only, B) UPA+LSU sequences, and C) a concatenated alignment of the COI, LSU and UPA sequences. Nodal support (bootstrap proportions) is represented as follows: white = 70-79%, grey = 80-89%, black = 90-100%. Scale bars = substitutions per site.

Table 3. Number of nodes with bootstrap support based on neighbor-joining analyses of sequence data for the UPA marker, UPA + LSU markers, and UPA + LSU + COI markers, with 2,000 bootstrap replicates.

Analysis	*# with 70-79% support*	*# with 80-89% support*	*# with 90-100% support*	*Total*
UPA	22	10	64	96
UPA + LSU	18	18	93	129
UPA + LSU + COI	23	26	133	182

Table 4. Numbers of supported nodes (at each support level) for each taxonomic level and each analysis (*other refers to support values for groups of sequences without taxonomic standing)

Analysis	*Level of support*	*Order*	*Family*	*Genus*	*Species*	*Other**
UPA	Black (> 90%)	4	7	22	30	21
UPA	Grey (80-89%)	0	1	1	1	7
UPA	White (70-79%)	0	1	2	5	14
UPA + LSU	Black (> 90%)	4	11	33	43	36
UPA + LSU	Grey (80-89%)	1	1	0	2	14
UPA + LSU	White (70-79%)	0	1	0	3	14
UPA + LSU + COI	Black (> 90%)	5	14	28	44	69
UPA + LSU + COI	Grey (80-89%)	0	1	5	0	18
UPA + LSU + COI	White (70-79%)	1	1	1	1	19

The fully concatenated data set (UPA + LSU + COI) roduced a neighbor-joining tree with strong support values at the familial level and below (Fig. 5), but the backbone of the tree was completely unresolved. At the ordinal level the Batrachospermales, Bonnemaisoniales, Ceramiales, Colaconematales, Corallinales, Gracilariales, Gelidiales, Halymeniales, Nemastomatales and Rhodymeniales were monophyletic (although the Acrosymphytales, Hildenbrandiales, Peyssonneliales, Plocamiales and Sporolithales were represented by only a single accession in the analysis), but the Gigartinales and Nemaliales were polyphyletic (Fig. 5). At the familial level, however, support tended to be higher. For example, the Liagoraceae, Galaxauraceae and Scinaiaceae (Nemaliales) were all monophyletic, but the order Nemaliales was not due to the nesting of the Colaconematales. Only a single family was represented in the orders Acrosymphytales, Hildenbrandiales, Plocamiales, Sporolithales, Peyssonneliales, Gracilariales, Corallinales, Gelidiales and Halymeniales. The families Callithamniaceae, Rhodomelaceae and Spyridiaceae within the Ceramiales were monophyletic, although the Ceramiaceae, Dasyaceae, Delesseriaceae and Wrangeliaceae were not (Fig. 5). Both the Naccariaceae and the Bonnemaisoniaceae were monophyletic in the Bonnemaisoniales. Although representation within the Rhodymeniales was not strong considering the familial diversity, all represented families were monophyletic.

DISCUSSION

The adjusted success rates of obtaining target sequence data for the three markers (46.8% for COI, 64.7% for UPA and 79.6% for LSU) clearly illustrate that the ease of marker data acquisition is not equal. Given that a large number of accessions were processed and studied over the course of the Hawaiian Rhodophyta Biodiversity Survey (> 2,000), we aimed to standardize primer, PCR amplification and sequencing protocols for each marker as much as possible, such that only 1-2 variations per marker were attempted, which is also one of the desired characteristics of DNA barcoding protocols (Hajibabaei *et al.*, 2005). Others have noted the necessity of heavily modified primers for even closely

Fig. 5. Neighbor-joining tree of Hawaiian florideophyte red algae based on a concatenated alignment of the COI, LSU and UPA sequences, with taxon labels indicating family and order. Nodal support (bootstrap proportions) is represented as follows: white = 70-79%, grey = 80-89%, black = 90-100%. Scale bar = substitutions per site.

related species of red algae for the COI marker (Clarkston & Saunders, 2010), and this, in combination with our finding that the COI sequences were the most difficult to obtain using standardized protocols, implies that large-scale DNA barcoding of red algae with COI will not be as straightforward as for some other groups of organisms (e.g. birds — Hebert *et al.*, 2004b; fishes — Ward *et al.*, 2009; lepidopterans – Hebert *et al.*, 2003). The breakdown of nucleotide divergence by codon position for COI, with the third position harboring much of the divergence (Fig. 1), is to be expected given that it is a coding gene (Zardoya & Meyer, 1996).

In contrast, the rRNA markers used in the biodiversity survey (UPA and LSU) are easier to amplify and sequence, but have the tradeoff of being more conserved in sequence (Fig. 1), such that it is much less likely that sufficient divergence will be detected between closely related taxa to reliably differentiate them (see Sherwood *et al.*, in press for examples), which is critical for the purposes of DNA barcoding.

As would be expected based on the divergences displayed in the nucleotide divergence plot (Fig. 1) and base pair frequencies (Fig. 3), the COI marker exhibited the largest degree of substitutional saturation, with the saturation point being reached at approximately one fifth of the maximum F84 distance of the COI data set (Fig. 2a). Interestingly, however, although the LSU marker was the most conserved (Fig. 1), some saturation was observed for the most distant sequences in the data set (Fig. 2b), which we have considered negligible for barcoding purposes, whereas the UPA marker did not display saturation for any of the comparisons (Fig. 2c). Although saturation is likely of greater importance in phylogenetic reconstruction than DNA barcoding, the analysis is worthy in the context of exploring additional uses for barcode data, such as our concatenated phylogeny (Fig. 5), and in identifying levels of divergence at which caution should be used in interpreting barcode comparisons of the three markers (e.g. Cywinska *et al.*, 2006). The issue of saturation can be further addressed in phylogenetic reconstruction with alternative strategies, for example, by partitioning the data set by marker (and by codon position for the COI sequences) to allow different models of evolution to be used (Verbruggen *et al.*, 2010), or through phylogenetic reconstruction of translated sequences (Salemi, 2003; Xia *et al.*, 2003) or R-Y coding (Verbruggen & Theriot, 2008).

It is noteworthy that some florideophyte groups reached saturation of the COI and LSU markers at different points (Table 2), implying that rates of evolution are not constant across lineages for these regions. This variation can also have implications for DNA barcoding if comparisons are made at too broad a taxonomic scale (i.e. beyond the individual threshold of saturation for that group of organisms; *sensu* Cywinska *et al.*, 2006), or if the comparative database has not been sufficiently populated for that taxon. The COI marker reaches saturation at distances between 0.06-0.25 and the LSU marker as low as 0.05 for the Halymeniales, and not at all for a number of other groups. No obvious taxonomic trend can be discerned from these numbers, and more data are likely needed within each group, as well as from different geographical regions, for full interpretation. Until that time, however, it is recommended that caution be used when comparing sequences nearing or exceeding the identified saturation points listed in Table 2.

The COI, LSU and UPA markers were sequenced for the Hawaiian Rhodophyta Biodiversity Survey with the primary intent of using the data for barcode-like comparisons — i.e. distance-based comparisons to identify unknown specimens and to begin clarification of species boundaries for the Hawaiian red algal flora using a standardized molecular data set. The data have indeed been used for these purposes (e.g. Sherwood, 2008 [*Asparagopsis taxiformis*], Conklin *et al.*, 2009 [*Kappaphycus* and *Eucheuma*], Sherwood *et al.*, in press [*Amansia glomerata*]). Nonetheless, when combined, the sequences yield 1,321 bp of data (with no indels and after trimming ambiguous regions), and the phylogenetic "performance" of the combined data was examined in relation to the results from the smaller data sets. This comparison was especially timely given the recent publication of a red algal tree of life based on a heterogeneous sequence data set obtained from mined data (Verbruggen *et al.*, 2010), which provided an up-to-date

basis of comparison. Our marker data represent a ready-to-use data source that could be incorporated into future reassessments of the red algal tree of life. As expected, bootstrap support increased with the addition of nucleotide data, both in terms of numbers of nodes with ≥ 70% support, and in the category of highest support (≥ 90%). These numbers were determined as an initial exploratory analysis, but should ultimately be re-evaluated with a formal phylogenetic analysis. Although topology of each analysis was not compared in detail due to space constraints, the topology of the concatenated data set (Fig. 5) was compared with that of Verbruggen *et al.* (2010). A number of orders were well supported, although in general, support for relationships among the florideophyte orders was not high, including the poorly supported regions of the tree identified by Verbruggen *et al.* (2010). This was especially true for regions "B" [among orders of the Nemaliophycidae] and "C" [among orders of the Rhodymeniophycidae]; representation for other regions was low or not included in the present analyses. Support values were generally higher at the family level, which is similar to the findings of Verbruggen *et al.* (2010), although their phylogeny was better supported overall, especially at the ordinal level. It should be noted, however, that our taxonomic representation was not as broad (being limited to Hawaiian species), our analyses differed in having many accessions for some groups, and our analyses were limited to NJ rather than complex phylogenetic analyses. So why examine the data in this manner? Algal biodiversity surveys sometimes opt to use more than one short molecular marker (e.g. Cianciola *et al.*, 2010), and understanding the limitations of performance for these markers will allow researchers to maximize the utility of their data.

In conclusion, marker selection will depend on the question being asked. For biodiversity surveys such as the one examined here, where taxa spanning the florideophyte red algae need to be efficiently sequenced, the COI barcode may need to be supplemented with additional markers to ensure sufficient data (such as LSU and/or UPA, unless taxon-specific or more universal primer combinations for COI are designed, optimized and made available). In contrast, those researchers working on DNA barcoding projects with more phylogenetically restricted groups of taxa, such as individual genera, may find that working with COI presents fewer problems. Phylogenetic reconstruction at the broad scale of florideophyte algae requires more than one or two short markers (e.g. the 14 loci used by Verbruggen *et al.*, 2010), but a large degree of family-level resolution seems to be achieved through the concatenation of the COI, LSU and UPA markers targeted for the Hawaiian Rhodophyta Biodiversity Survey, allowing greater utility of the data set than the initial aim of distance-based comparisons.

Acknowledgements. This research was supported by the US National Science Foundation (grant DEB-0542508 to ARS and GGP). T.S. was partially supported by The David and Lucile Packard Foundation (grant #2006-30569 to I.A. Abbott), and this grant also supported the collection and processing of many samples from the Northwestern Hawaiian Islands. We thank Norman Wang for building and maintaining the Hawaiian Algal Database and providing data downloads. Napua Harbottle, Jack Fisher and Roy Tsuda of the Bernice P. Bishop Museum are acknowledged for their assistance and support of using herbarium specimens. We also thank Celia Smith, Heather Spalding and Kimberly Peyton for allowing use of specimens from the Hawaii Undersea Research Laboratory cruises, and the organizers of the Census of Marine Life CReefs cruise to French Frigate Shoals in the Northwestern Hawaiian Islands for facilitating participation in this expedition. A.R.S. would like to acknowledge helpful discussions with Gary Saunders. We thank two anonymous reviewers for their helpful comments.

REFERENCES

ABBOTT I.A., 1999 — *Marine Red Algae of the Hawaiian Islands*. Honolulu, Bishop Museum Press, 477 p.

CIANCOLA E.N., POPOLIZIO T.R., SCHNEIDER C.W. & LANE C.E., 2010 — Using molecular-assisted alpha taxonomy to better understand red algal biodiversity in Bermuda. *Diversity* 2: 946-958.

CLARKSTON B.E. & SAUNDERS G.W., 2010 — A comparison of two DNA barcode markers for species discrimination in the red algal family Kallymeniaceae (Gigartinales, Florideophyceae), with a description of *Euthora timburtonii* sp. nov. *Botany* 88: 119-131.

CONKLIN K.Y., KURIHARA A. & SHERWOOD A.R., 2009 — A molecular method for identification of the morphologically plastic invasive algal species *Eucheuma denticulatum* and *Kappaphycus* spp. (Rhodophyta, Gigartinales) in Hawaii. *Journal of applied phycology* 21: 691-699.

CYWINSKA A., HUNTER F.F. & HEBERT P.D.N., 2006 — Identifying Canadian mosquito species through DNA barcodes. *Medical and veterinary entomology* 20: 413-424.

HAJIBABAEI M., DEWAARD J.R., IVANOVA N.V., RATNASINGHAM S., DOOH R.T., KIRK S.L., MACKIE P.M. & HEBERT P.D.N., 2005 — Critical factors for assembling a high volume of DNA barcodes. *Philosophical transactions of the royal society Series B* 360: 1959-1967.

HARPER J.T. & SAUNDERS G.W., 2001 — The application of sequences of the ribosomal cistron to the systematics and classification of the Rhodophyta. *Cahiers de biologie marine* 42: 25-38.

HEBERT P.D.N., CYWINSKA A., BALL S.L. & DEWAARD J.R., 2003 — Biological identifications through DNA barcodes. *Proceedings of the royal society of London, Series B* 270: 313-322.

HEBERT P.D.N., PENTON E.H., BURNS J.M., JANZEN D.H. & HALLWACHS W., 2004a — Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the national academy of sciences of the United States of America* 101: 14812-14817.

HEBERT P.D.N., STOECKLE M.Y., ZEMLAK T.S. & FRANCIS C.M., 2004b — Identification of birds through DNA barcodes. *PLoS Biology* 2: e312.

LE GALL L. & SAUNDERS G.W., 2010 — DNA barcoding is a powerful tool to uncover algal diversity: a case study of the Phyllophoraceae (Gigartinales, Rhodophyta) in the Canadian flora. *Journal of phycology* 46: 374-389.

NEWMASTER S.G. & RAGUPATHY S., 2009 — Testing plant barcoding in a sister species complex of pantropical *Acacia* (Mimosoideae, Fabaceae). *Molecular ecology resources* 9: 172-180.

PARADIS E., CLAUDE J. & STRIMMER K., 2004 — APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289-290.

PHILIPPE H., SÖRHANNUS U., BAROIN A., PERASSO R., GASSE F. & ADOUTTE, A., 1994 — Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. *Journal of evolutionary biology* 7: 247-265.

PRESTING G.G., 2006 — Identification of conserved regions in the plastid genome – implications for DNA barcoding and biological function. *Canadian journal of botany* 84: 1434-1443.

R DEVELOPMENT CORE TEAM, 2010 — R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org).

ROBBA L., RUSSELL S.J., BARKER G.L. & BRODIE J., 2006 — Assessing the use of the mitochondrial *cox1* marker for use in DNA barcoding of red algae (Rhodophyta). *American journal of botany* 93: 1101-1108.

SALEMI M., 2003 — Practice: The PHYLIP and TREE-PUZZLE Software Packages. *In:* Salemi M. & Vandamme, A.-M. (eds), *The Phylogenetic Handbook*. Cambridge, Cambridge University Press, pp. 88-100.

SAUNDERS G.W., 2005 — Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future application. *Philosophical transactions of the royal society Series B* 360: 1879-1888.

SAUNDERS G.W., 2009 — Routine DNA barcoding of Canadian Gracilariales (Rhodophyta) reveals the invasive species *Gracilaria vermiculophylla* in British Columbia. *Molecular ecology resources* 9: 140-150.

SHERWOOD A.R. & PRESTING G.G., 2007 — Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria. *Journal of phycology* 43: 605-608.

SHERWOOD A.R., 2008 — Phylogeography of *Asparagopsis taxiformis* (Bonnemaisoniales, Rhodophyta) in the Hawaiian Islands: two mtDNA markers support three separate introductions. *Phycologia* 47:79-88.

SHERWOOD A.R., KURIHARA A. & CONKLIN K.Y. (in press) — Molecular diversity of Amansieae (Ceramiales, Rhodophyta) from the Hawaiian Islands: a multi-marker assessment reveals high diversity within *Amansia glomerata*. *Phycological Research* (in press).

TAMURA K., DUDLEY J., NEI M. & KUMAR S., 2007 — MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular biology and evolution* 24: 1596-1599.

THOMPSON J.D., GIBSON T.J., PLEWNIAK F., JEANMOUGIN F. & HIGGINS D.G., 1997 — The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids research* 25: 4876-4882.

VERBRUGGEN H. & THERIOT E.C., 2008 — Building trees of algae: some advances in phylogenetic and evolutionary analysis. *European journal of phycology* 43: 229-252.

VERBRUGGEN H., MAGGS C.A., SAUNDERS G.W., LEGALL L., YOON H.S. & DE CLERCK O., 2010 — Data mining approach identifies research priorities and data requirements for resolving the red algal tree of life. *BMC Evolutionary biology* 10: 16.

WANG N., SHERWOOD A.R., KURIHARA A., CONKLIN K.Y., SAUVAGE T. & PRESTING G.G., 2009 — The Hawaiian Algal Database: a laboratory LIMS and online resource for biodiversity data. *BMC Plant biology* 9:117.

WARD R.D., HANNER R. & HEBERT P.D.N., 2009 — Review Paper: The campaign to DNA barcode all fishes, FISH-BOL. *Journal of fish biology* 74: 329-356.

XIA X., 2010 — DAMBE v.5.2.6. Distributed by the author at *http://dambe.bio.uottawa.ca.*

XIA X., XIE Z., SALEMI M., CHEN L. & WANG Y., 2003 — An index of substitution saturation and its application. *Molecular phylogenetics and evolution* 26: 1-7.

ZARDOYA R. & MEYER A., 1996 — Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. *Molecular biology and evolution* 13: 933-942.