OPEN ACCESS Freely available online

PLoS one

A New Method for Species Identification via Protein-Coding and Non-Coding DNA Barcodes by Combining Machine Learning with Bioinformatic Methods

Ai-bing Zhang[1]*, Jie Feng[2], Robert D. Ward[3], Ping Wan[1], Qiang Gao[1], Jun Wu[1], Wei-zhong Zhao[2]

1 College of Life Sciences, Capital Normal University, Beijing, People's Republic of China, **2** School of Mathematical Sciences, Capital Normal University, Beijing, People's Republic of China, **3** Wealth from Oceans Flagship, CSIRO Marine and Atmospheric Research, Hobart, Tasmania, Australia

Abstract

Species identification via DNA barcodes is contributing greatly to current bioinventory efforts. The initial, and widely accepted, proposal was to use the protein-coding cytochrome c oxidase subunit I (COI) region as the standard barcode for animals, but recently non-coding internal transcribed spacer (ITS) genes have been proposed as candidate barcodes for both animals and plants. However, achieving a robust alignment for non-coding regions can be problematic. Here we propose two new methods (DV-RBF and FJ-RBF) to address this issue for species assignment by both coding and non-coding sequences that take advantage of the power of machine learning and bioinformatics. We demonstrate the value of the new methods with four empirical datasets, two representing typical protein-coding COI barcode datasets (neotropical bats and marine fish) and two representing non-coding ITS barcodes (rust fungi and brown algae). Using two random sub-sampling approaches, we demonstrate that the new methods significantly outperformed existing Neighbor-joining (NJ) and Maximum likelihood (ML) methods for both coding and non-coding barcodes when there was complete species coverage in the reference dataset. The new methods also out-performed NJ and ML methods for non-coding sequences in circumstances of potentially incomplete species coverage, although then the NJ and ML methods performed slightly better than the new methods for protein-coding barcodes. A 100% success rate of species identification was achieved with the two new methods for 4,122 bat queries and 5,134 fish queries using COI barcodes, with 95% confidence intervals (CI) of 99.75–100%. The new methods also obtained a 96.29% success rate (95%CI: 91.62–98.40%) for 484 rust fungi queries and a 98.50% success rate (95%CI: 96.60–99.37%) for 1094 brown algae queries, both using ITS barcodes.

Citation: Zhang A-b, Feng J, Ward RD, Wan P, Gao Q, et al. (2012) A New Method for Species Identification via Protein-Coding and Non-Coding DNA Barcodes by Combining Machine Learning with Bioinformatic Methods. PLoS ONE 7(2): e30986. doi:10.1371/journal.pone.0030986

Editor: David S. Milstone, Brigham and Women's Hospital, United States of America

Received July 1, 2011; **Accepted** December 29, 2011; **Published** February 20, 2012

Funding: This study was supported by Beijing Municipal Natural Science Foundation Key Projects (Grant No. KZ201010028028 to AZ), by Natural Science Foundation of China (Grant No. 31071963 to AZ), by Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (Grant No. PHR201107120 to AZ), and by The Research Fund for the Doctoral Program of Higher Education (Grant No. 20101108120002 to AZ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: zhangab2008@mail.cnu.edu.cn

Introduction

DNA barcoding has become increasingly popular as a tool for species discrimination and identification [1–19], although some aspects remain controversial [20–34]. As of October 2011, there were 1, 381, 970 barcodes from 114, 873 species in the Barcode of Life Database (BOLD, www.barcodinglife.org), covering a very wide spectrum of species from algae, fungi, bacteria and plants to invertebrates and vertebrates. The COI barcode has proven to be a successful species-discriminator in most animal groups, but is generally less successful elsewhere. BOLD therefore also includes internal transcribed spacer (ITS) sequences for fungal identification and the two chloroplast-encoded genes ribulose bisphosphate carboxylase (rbcL) and maturaseK (MatK) for plants.

A fundamental issue in DNA barcoding is how best to assign a query sequence from an unknown specimen to the correct species in the reference sequence database [15,19,24,25,35–43]. Currently, most empirical studies employ traditional phylogenetic methods such as Neighbour-joining [1,2,44] to construct an evolutionary tree with both query and reference sequences. A sequence visually falling in a single-species clade is treated as the conspecific of that species. However, if the query falls into a polyphyletic or paraphyletic clade, assignation to correct species becomes ambiguous.

More recently, other statistical approaches to assignation have been suggested including decision theory [11] and Bayesian methods [36,38,39]. Zhang and colleagues have proposed a neural network based approach [15,45]. Neural networks were originally developed to model the function of connected neurons in the brain [46]. However, their utility as a general computational tool was realized with the development of the back-propagation method [47–50]. It has been applied successfully in many fields, including speech synthesis, handwriting recognition and medical diagnostics. In molecular genetics it has been applied to some aspects of DNA/RNA and protein sequence analysis [51,52], such as protein and ribosomal RNA classification [53–55] and phylogenetic reconstruction [56]. Some machine learning techniques have also been proposed for the analysis of DNA sequences, including Classification and Regression Trees (CART) [57,58], Random Forest (RF) [58,59], and Support Vector Machines [60]. All these

methods, and those based on tree construction, require a prior alignment of sequences. Sequence alignment is generally straightforward for protein-coding regions, such as the COI sequence proposed as the universal animal barcode, but can be difficult when barcodes are based on non-coding regions such as 28 S or ITS which have variable length and indels (gaps). A robust alignment of non-coding regions can be extremely hard to achieve. Even if an alignment can be obtained using existing algorithms, such as those employed in ClustalW (http://www.clustal.org/) [61], the computation of genetic distances among sequences is still problematic since there is, so far, no molecular evolutionary model which simulates the evolution of DNA sequences with indels. The indels are generally removed or treated as missing data in the subsequent analysis. Sometimes, indels may be coded as fifth states or given other codes, introducing extra assumptions. While it is necessary, for some taxa, to incorporate non-coding barcodes into the BOLD system, it would be advantageous to eliminate the need to align these sequences for species identification.

In an attempt to overcome these difficulties, we propose here a new species identification strategy taking advantage of both bioinformatics and machine learning as an extension of our prior back-propagation neural network application [15]. It is especially aimed at identifying species with non-coding barcodes, a topic little explored in the current barcoding literature. We test our methods with four empirical datasets, two representing typical protein-coding COI barcodes and two using the non-coding barcode ITS, and compare the results to those from two traditional barcoding strategies, Neighbor-joining (NJ) [44] and Maximum likelihood (ML) [62]. We used more than 21,220 random queries against the corresponding reference libraries. We demonstrate that the new procedures outperform the two traditional barcoding methods and BP-based methods [15]. This is largely because sequence alignments are no longer required - a big advantage for non-coding sequences - and to the saving of computational time compared to previous BP-based methods [15].

Results

Neotropical bat and Marine fish COI datasets

In total, 8,120 random queries from 766 bat COI sequences were examined with two traditional methods (NJ and ML) and the two newly proposed methods (DV-RBF and FJ-RBF) against corresponding reference libraries. 5,180 of these queries were carried out using 5 repeated random splits, representing complete/balanced species coverage in the reference library (meaning that all species from the original database remain in the reference library, see Materials and Methods). The remaining 2,940 queries were conducted using five-fold cross-validation, representing incomplete/unbalanced species coverage in the reference library (meaning that some species from the original database might be absent from the reference library). For the two new methods, 4,122 queries were performed against the corresponding reference databases (Table 1, Table S1). In the case of balanced species coverage, both DV-RBF and FJ-RBF methods achieved 100% success rates (95% CI: 99.70–100%) with 1,295 random queries each (Figure 1a), while the NJ and ML methods obtained success rates of 95.75% (95% CI: 94.51–96.72%), and 87.25% (95% CI: 85.33–88.96%) respectively. For unbalanced species coverage, with 766 random queries for each of DV-RBF and FJ-RBF, the NJ method outperformed all other methods (94.86% with a 95% CI: 92.93–96.28%) compared with ML 88.97% (95% CI: 86.53–91.01%); DV-RBF 86.18% (95% CI: 83.53–88.46%) and FJ-RBF 81.54% (95% CI: 78.61–84.14%) (Figure 1a). The slightly better performance of ML than either DV-RBF and FJ-RBF was without statistical significance (Figure 1 1a).

More than 10,000 (10,040) random replications of queries were performed for the fish dataset against the corresponding reference libraries. 6,340 random queries were carried out with 5 repeated random splits, representing complete/balanced species coverage. The remaining 3,700 random queries were assigned with five-fold cross-validation, representing incomplete/unbalanced species coverage in the reference library. For the two new methods, 5,134 queries were performed against the corresponding reference databases (Table 2, Table S2).

In the situation of complete species coverage, the two new methods (DV-RBF and FJ-RBF) had 100% success rates (95% CI: 99.75–100%), significantly outperforming the two traditional methods that gave success rates of 99.05% (95% CI: 98.44–99.42%) and 93.37 (95% CI: 92.04–94.49%) for NJ and ML respectively (Figure 1b). However, traditional NJ and ML approaches significantly outperformed both DV-RBF and FJ-RBF under the circumstance of unbalanced species coverage (Figure 1b) (NJ, 98.81% with 95% CI: 97.88–99.33%; ML, 93.72% with 95% CI:91.97–95.11% ; DV-RBF, 88.00% with 95% CI: 85.74–89.93%; FJ-RBF, 87.35% with 95% CI: 85.05–89.04%). Our results from the bat and fish protein-coding COI datasets showed that the structure of reference libraries (balanced versus unbalanced species coverage) could affect species identification success rates. The two newly proposed methods perform very well in the former situation, but less well in the latter.

Rust fungi ITS dataset

Since ITS barcodes were only recently developed as alternative barcode markers, there are relatively limited data available. We obtained 85 clean sequences from 14 species of rust fungi and performed 872 random queries with the four barcoding methods. 540 queries were conducted under the situation of balanced species coverage (5 repeated random splits) and 332 queries for the case of unbalanced species coverage (five-fold cross validation, Table 3). The two new methods (DV-RBF and FJ-RBF) significantly outperformed the two traditional methods (NJ and ML) whether or not species coverage in the reference library is balanced (Figure 1c). For instance, both DV-RBF and FJ-RBF methods achieved a 96.29% success rate (95% CI: 91.62–98.40%) for unbalanced coverage while NJ and ML only obtained success rates of 25.92% (95% CI: 19.27–33.91%) and 14.81% (95% CI: 9.79–21.77%) respectively (Figure 1c). In the situation of balanced species coverage, traditional NJ and ML methods obtained higher but still less than 60.00% success rates (NJ, 57.00% with 95% CI: 47.09–67.87% ; ML, 42.16% with 95% CI: 32.12–52.90%; Figure 1c), again much less than the success rates for the two new methods (DV-RBF, 75.90% with 95% CI: 65.19–83.82%; FJ-RBF, 78.31% with 95% CI: 68.30–85.81; Figure 1c; Table S3). Thus in the case of non-coding barcodes, the two newly proposed methods (DV-RBF and FJ-RBF) considerably outperformed the two traditional methods (NJ and ML) regardless of the structure of reference libraries (balanced versus unbalanced species coverage).

Brown algae ITS dataset

207 ITS sequences of brown algae data from 16 species were obtained after data cleansing. We performed 2,188 random queries against corresponding reference libraries with the four barcoding methods, of which 1,360 were conducted using repeated random splits (5 times, each 340 queries for each method), and 828 using five-fold cross-validation (Table 4 and Table S4). As in the case of the rust fungi ITS dataset, both DV-RBF and FJ-RBF methods outperformed with statistical significance the two traditional methods (Figure 1d). A success rate of 98.52% (95% CI: 96.60–99.37%) was achieved for both

Table 1. Species assignments for Neotropical bats [81] based on COI sequences for all 4122 random queries using DV-RBF and FJ-RBF methods.

No.	Category of Random Tests[a]	Query[b]	DV-RBF[c]	Status	FJ-RBF[d]	Status
1	random	BCBNT34706-Rhynchonycteris naso	*Rhynchonycteris naso*	(✓[e])	*Rhynchonycteris naso*	(✓[e])
2	splits	BCBNT35706-Rhynchonycteris naso	*Rhynchonycteris naso*	(✓)	*Rhynchonycteris naso*	(✓)
3		BCBNT13006-Diclidurus isabellus	*Diclidurus isabellus*	(✓)	*Diclidurus isabellus*	(✓)
4		BCBNT37906-Diclidurus isabellus	*Diclidurus isabellus*	(✓)	*Diclidurus isabellus*	(✓)
5		BCBNT14306-Diclidurus isabellus	*Diclidurus isabellus*	(✓)	*Diclidurus isabellus*	(✓)
6		BCBNT92206-Chrotopterus auritus	*Chrotopterus auritus*	(✓)	*Chrotopterus auritus*	(✓)
7		BCBNT59706-Chrotopterus auritus	*Chrotopterus auritus*	(✓)	*Chrotopterus auritus*	(✓)
8		BCBNT04006-Cormura brevirostris	*Cormura brevirostris*	(✓)	*Cormura brevirostris*	(✓)
9		BCBNT05606-Cormura brevirostris	*Cormura brevirostris*	(✓)	*Cormura brevirostris*	(✓)
10		BCBNT39906-Pteronotus personatus	*Pteronotus personatus*	(✓)	*Pteronotus personatus*	(✓)
11		BCBNT09706-Pteronotus personatus	*Pteronotus personatus*	(✓)	*Pteronotus personatus*	(✓)
12		BCBNT36906-Noctilio albiventris	*Noctilio albiventris*	(✓)	*Noctilio albiventris*	(✓)
...		...	...	...	...	...
1295	(259×5)	BCBNT55406-Lophostoma silvicolum	*Lophostoma silvicolum*	(✓)	*Lophostoma silvicolum*	(✓)
1	n-fold	BCBNT29806-Trachops cirrhosus	*Trachops cirrhosus*	(✓)	*Trachops cirrhosus*	(✓)
2	cross-	BCBNT63906-Platyrrhinus helleri	*Platyrrhinus helleri*	(✓)	*Platyrrhinus helleri*	(✓)
3	validation	BCBNC12906-Rhinophylla pumilio	*Rhinophylla pumilio*	(✓)	*Rhinophylla pumilio*	(✓)
4		BCBNT94306-Molossus molossus	*Molossus molossus*	(✓)	*Molossus molossus*	(✓)
5		BCBNC01906-Rhinophylla pumilio	*Rhinophylla pumilio*	(✓)	*Rhinophylla pumilio*	(✓)
6		BCBNT70306-Phyllostomus discolor	*Phyllostomus discolor*	(✓)	*Phyllostomus discolor*	(✓)
7		BCBNT99106-Platyrrhinus aurarius	*Platyrrhinus aurarius*	(✓)	*Carollia perspicillata*	(✗)
8		BCBNC16806-Platyrrhinus aurarius	*Platyrrhinus aurarius*	(✓)	*Platyrrhinus aurarius*	(✓)
9		BCBN31305-Rhinophylla pumilio	*Rhinophylla pumilio*	(✓)	*Rhinophylla pumilio*	(✓)
10		BCBNC06506-Lionycteris spurrelli	*Lionycteris spurrelli*	(✓)	*Lionycteris spurrelli*	(✓)
11		BCBN55205-Trachops cirrhosus	*Trachops cirrhosus*	(✓)	*Trachops cirrhosus*	(✓)
12		BCBNC16106-Platyrrhinus aurarius	*Platyrrhinus aurarius*	(✓)	*Platyrrhinus aurarius*	(✓)
...		...	...	...	...	...
766	$(153 \times 5 + 1)$	BCBNT94606-Glyphonycteris daviesi	*Carollia brevicauda*	(✗)	*Carollia perspicillata*	(✗)

[a]: Two categories of randomization were performed in this study. One is random splits which were conducted at species level (5 times) and the other is n-fold cross-validation which was performed on the whole dataset ($n = 5$ was used). 4122 random queries were generated based on the original 766 bat COI sequences, see text and Online Appendix I for details.

[b]: The names of query sequences consist of BOLD sequence accession numbers (a dash was removed before the last two numbers) and their true species names. Only part of the results were presented here, see Online Appendix I for all 4122 queries and corresponding assignments (singletons were excluded since they can only be assigned to the wrong speices).

[c]: DV denotes DV-Curve, RBF indicates RBF neural network, see text for details.

[d]: FJ denotes FJ-Curve.

[e]: Ticks and crosses indicate correct and wrong assignments respectively.

doi:10.1371/journal.pone.0030986.t001

DV-RBF and FJ-RBF methods in the case of balanced species coverage while NJ and ML methods obtained extremely low success rates of 10.29% (95% CI: 7.49–13.98%) and 7.35% (95% CI: 5.02–10.62%) respectively (Figure 1d). In the situation of unbalanced species coverage, DV-RBF and FJ-RBF obtained somewhat reduced success rates (DV-RBF, 93.71% with 95% CI: 89.55–96.29%; FJ-RBF, 88.40% with 95% CI: 83.32–92.08%), but they were nevertheless much larger than those of the two traditional methods (NJ, 15.45% with 95% CI: 11.16–21.00%; ML, 38.64% with 95% CI: 32.27–45.43%; Figure 1d).

Processing time

The data analyses in this study were performed on a 3.00 GHz desktop computer (Intel(R) Core (TM)2, DuoCPU, E8400 @ 3.00 GHz×2). DV-RBF and FJ-RBF each spent 2.88-7.56 seconds per assignment, while the ML method spent 6.75–9.08 seconds per assignment exclusive of alignment time, depending on dataset size (from 68 to 785 reference sequences in this study). NJ spent less than one second per assignment, but the necessary sequence alignments can take several hours for a few hundred sequences.

Discussion

The new methods proposed in this study for barcode-based species assignations, which combine bioinformatics and machine learning, provide several advantages over existing methods, including the earlier BP-based method [15].

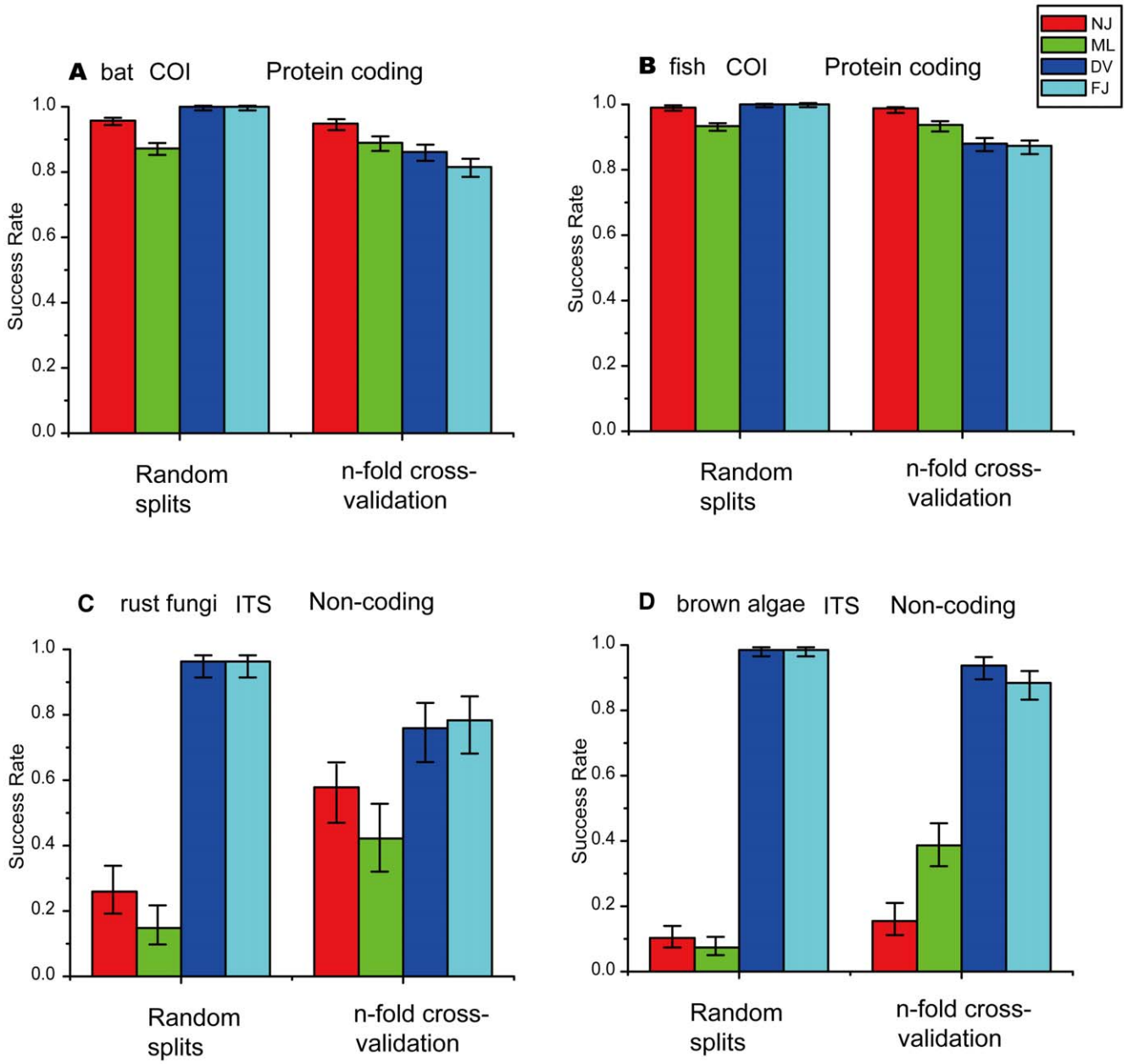

Figure 1. Success rate of species identification and 95% confidence intervals with the new methods (DV-RBF or FJ-RBF) proposed in this study based on COI barcodes and ITS barcodes for four empirical datasets.
doi:10.1371/journal.pone.0030986.g001

The first advantage is that no sequence alignment is required. Alignment algorithms and interpretations have been highly debated topics in the field of evolutionary studies over the past several decades [63–66]. This reflects the difficulties faced in aligning homologs, especially from variable-length non-coding gene regions [64]. Most of the commonly used DNA barcoding approaches to species identification, including classical phylogenetic approaches such as neighbour joining [1,2,44], and decision theory [11] and Bayesian methods [36,38,39], rely heavily on an initial robust alignment. Our new methods circumvent this complex issue by taking advantage of graphical representations of DNA sequences via a DV-Curve [67] or the newly-developed (herein) FJ-Curve approach. We demonstrated their successful applications to four empirical datasets, two of which are based on the commonly used coding COI barcodes, and two on the more-recently proposed non-coding ITS barcodes. The new methods strongly outperformed the existing Neighbor-joining (NJ) and Maximum likelihood (ML) methods for non-coding barcodes, while the latter two performed slightly better than the new methods for coding barcodes in circumstances of potentially unbalanced species coverage in the reference library. The very large discrepancy in success between the traditional and the new methods proposed here in identifying species by ITS sequences is largely attributable to the former, especially the model-based methods, relying heavily on molecular evolutionary models which generally ignore the evolution of indels/gaps. The phylogenetic signals contained in the indels/gaps will be lost during the analysis. In the case of balanced species coverage in the reference database,

Table 2. Species assignments for Pacific Canadian marine fish [82] based on COI sequences for all 5134 random queries using DV-RBF and FJ-RBF methods.

No.	Category of Random Tests[a]	Query[b]	DV-RBF[c]	Status	FJ-RBF[d]	Status
1	random	TZFPA15007-Eptatretus stoutii	*Eptatretus stoutii*	(✓[e])	*Eptatretus stoutii*	(✓[e])
2	splits	TZFPB55006-Eptatretus stoutii	*Eptatretus stoutii*	(✓)	*Eptatretus stoutii*	(✓)
3		TZFPB57806-Eptatretus stoutii	*Eptatretus stoutii*	(✓)	*Eptatretus stoutii*	(✓)
4		TZFPB21505-Eptatretus deani	*Eptatretus deani*	(✓)	*Eptatretus deani*	(✓)
5		TZFPB32505-Eptatretus deani	*Eptatretus deani*	(✓)	*Eptatretus deani*	(✓)
6		TZFPB04605-Porichthys notatus	*Porichthys notatus*	(✓)	*Porichthys notatus*	(✓)
7		TZFPB46906-Porichthys notatus	*Porichthys notatus*	(✓)	*Porichthys notatus*	(✓)
8		TZFPB04305-Porichthys notatus	*Porichthys notatus*	(✓)	*Porichthys notatus*	(✓)
9		TZFPB53606-Squalus acanthias	*Squalus acanthias*	(✓)	*Squalus acanthias*	(✓)
10		TZFPB56706-Squalus acanthias	*Squalus acanthias*	(✓)	*Squalus acanthias*	(✓)
11		TZFPB55906-Squalus acanthias	*Squalus acanthias*	(✓)	*Squalus acanthias*	(✓)
12		TZFPB42505-Cyclothone atraria	*Cyclothone atraria*	(✓)	*Cyclothone atraria*	(✓)
...		...	...	...	...	...
1585	(317×5)	TZFPA19707-Malacocottus	*Malacocottus zonurus*	(✓)	*Malacocottus zonurus*	(✓)
1	n-fold	TZFPB55306-Lycodes diapterus	*Lycodes diapterus*	(✓)	*Lycodes diapterus*	(✓)
2	cross-	TZFPB69106-Sebastes pinniger	*Sebastes pinniger*	(✓)	*Sebastes pinniger*	(✓)
3	validation	TZFPA14506-Talismania bifurcata	*Talismania bifurcata*	(✓)	*Talismania bifurcata*	(✓)
4		TZFPB71206-Ronquilus jordani	*Ronquilus jordani*	(✓)	*Ronquilus jordani*	(✓)
5		TZFPB56606-Sebastes aleutianus	*Sebastes aleutianus*	(✓)	*Sebastes aleutianus*	(✓)
6		TZFPA19407-Nectoliparis pelagicus	*Oncorhynchus tshawytscha*	(✗)	*Bathyagonus infraspinatus*	(✗)
7		TZFPB82006-Sebastes reedi	*Sebastes reedi*	(✓)	*Sebastes reedi*	(✓)
8		TZFPB46706-Alosa sapidissima	*Alosa sapidissima*	(✓)	*Alosa sapidissima*	(✓)
9		TZFPB87508-Oligocottus maculosus	*Oligocottus maculosus*	(✓)	*Oligocottus maculosus*	(✓)
10		TZFPB32805-Alepocephalus tenebrosus	*Alepocephalus tenebrosus*	(✓)	*Alepocephalus tenebrosus*	(✓)
11		TZFPB58306-Theragra chalcogramma	*Theragra chalcogramma*	(✓)	*Theragra chalcogramma*	(✓)
12		TZFPB86908-Cyclothone atraria	*Sebastolobus alascanus*	(✗)	*Bathyagonus infraspinatus*	(✗)
...		...	...	...	...	...
982	$(196 \times 5 + 2)$	TZFPB16505-Sebastes flavidus	*Sebastes flavidus*	(✓)	*Sebastes flavidus*	(✓)

[a]: Two categories of randomization were performed in this study. One is random splits which were conducted at species level (5 times) and the other is n-fold cross-validation which was performed on the whole dataset ($n = 5$ was used). 5134 random queries were generated based on the original 982 fish COI sequences, see text and Online Appendix II for details.

[b]: The names of query sequences consist of BOLD sequence accession numbers (a dash was removed before the last two numbers) and their true species names. Only part of the results were presented here, see Online Appendix II for all 5134 queries and corresponding assignments (singletons were excluded since they can only be assigned to the wrong speices).

[c]: DV denotes DV-Curve, RBF indicates RBF neural network, see text for details.

[d]: FJ denotes FJ-Curve.

[e]: Ticks and crosses indicate correct and wrong assignments respectively.

doi:10.1371/journal.pone.0030986.t002

the new methods outperformed the traditional NJ and ML methods for both coding and non-coding barcodes. This indicates that a complete reference library with balanced species coverage will improve species identification success rates: a well-curated reference database is an essential prerequisite for accurate species identification.

The second advantage, like the BP-based method [15], is that the new methods are based on fewer assumptions when making inferences. Most other current methods rely on a number of more or less restrictive assumptions that may not apply to real data [15,36]. For example, the decision theory method [11] assumes an ideal panmictic population for all species or groups without recombination, migration, and so on, so that the evolutionary process within each group is governed by only one parameter: the number of mutational steps between two individuals within that group [15].

Whether it is worthwhile to adopt a biological, populational and/or phylogenetic rationale for DNA barcode sequence assignation, or whether pure statistical approaches are more efficient, remains largely unaddressed [68]. Species identifications via DNA barcoding can be complex both in theory and in practice [19]. Some authors [58] have argued that no one method can

Table 3. Species assignments for rust fungi (BOLD project CHITS) based on ITS sequences for 484 random queries using DV-RBF and FJ-RBF methods.

No.	Category of Random Tests[a]	Query[b]	DV-RBF[c]	Status	FJ-RBF[d]	Status
1	random	CHITS08008-Chrysomyxa wereii	*Chrysomyxa wereii*	(✓[e])	*Chrysomyxa wereii*	(✓[e])
2	splits	CHITS07708-Chrysomyxa wereii	*Chrysomyxa wereii*	(✓)	*Chrysomyxa wereii*	(✓)
3		CHITS11109-Chrysomyxa pirolata	*Chrysomyxa pirolata*	(✓)	*Chrysomyxa pirolata*	(✓)
4		CHITS01308-Chrysomyxa pirolata	*Chrysomyxa pirolata*	(✓)	*Chrysomyxa pirolata*	(✓)
5		CHITS11009-Chrysomyxa pirolata	*Chrysomyxa pirolata*	(✓)	*Chrysomyxa pirolata*	(✓)
6		CHITS09509-Chrysomyxa arctostaphyli	*Chrysomyxa arctostaphyli*	(✓)	*Chrysomyxa arctostaphyli*	(✓)
7		CHITS04108-Chrysomyxa arctostaphyli	*Chrysomyxa arctostaphyli*	(✓)	*Chrysomyxa arctostaphyli*	(✓)
8		CHITS03208-Chrysomyxa empetri	*Chrysomyxa empetri*	(✓)	*Chrysomyxa empetri*	(✓)
9		CHITS03308-Chrysomyxa empetri	*Chrysomyxa empetri*	(✓)	*Chrysomyxa empetri*	(✓)
10		CHITS03108-Chrysomyxa chiogenis	*Chrysomyxa chiogenis*	(✓)	*Chrysomyxa chiogenis*	(✓)
11		CHITS02408-Chrysomyxa chiogenis	*Chrysomyxa chiogenis*	(✓)	*Chrysomyxa chiogenis*	(✓)
12		CHITS06208-Chrysomyxa ledicola	*Chrysomyxa ledicola*	(✓)	*Chrysomyxa ledicola*	(✓)
...		...	...	...	...	...
135	(27×5)	CHITS06508-Chrysomyxa nagodhii	*Chrysomyxa nagodhii*	(✓)	*Chrysomyxa nagodhii*	(✓)
1	n-fold	CHITS05608-Chrysomyxa ledi	*Chrysomyxa rhododendri*	(✗)	*Chrysomyxa rhododendri*	(✗)
2	cross-	CHITS01208-Chrysomyxa cassandrae	*Chrysomyxa cassandrae*	(✓)	*Chrysomyxa cassandrae*	(✓)
3	validation	CHITS04008-Chrysomyxa arctostaphyli	*Chrysomyxa arctostaphyli*	(✓)	*Chrysomyxa arctostaphyli*	(✓)
4		CHITS02308-Chrysomyxa chiogenis	*Chrysomyxa chiogenis*	(✓)	*Chrysomyxa chiogenis*	(✓)
5		CHITS02108-Chrysomyxa nagodhii	*Chrysomyxa cassandrae*	(✗)	*Chrysomyxa ledi*	(✗)
6		CHITS05308-Chrysomyxa arctostaphyli	*Chrysomyxa ledicola*	(✗)	*Chrysomyxa ledi*	(✗)
7		CHITS06208-Chrysomyxa ledicola	*Chrysomyxa ledicola*	(✓)	*Chrysomyxa ledicola*	(✓)
8		CHITS06008-Chrysomyxa ledicola	*Chrysomyxa ledicola*	(✓)	*Chrysomyxa ledicola*	(✓)
9		CHITS09509-Chrysomyxa arctostaphyli	*Chrysomyxa ledicola*	(✗)	*Chrysomyxa ledi*	(✗)
10		FUCUI00608-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
11		CHITS11009-Chrysomyxa pirolata	*Chrysomyxa pirolata*	(✓)	*Chrysomyxa pirolata*	(✓)
12		CHITS05708-Chrysomyxa ledi	*Chrysomyxa rhododendri*	(✗)	*Chrysomyxa rhododendri*	(✗)
...		...	...	...	...	...
107	$(21 \times 5 + 2)$	CHITS08909-Chrysomyxa ledicola	*Chrysomyxa ledicola*	(✓)	*Chrysomyxa ledicola*	(✓)

[a]: Two categories of randomization were performed in this study. One is random splits which were conducted at species level (5 times) and the other is n-fold cross-validation which was performed on the whole dataset ($n = 5$ was used). 484 random queries were generated based on the original 107 rust fungi ITS sequences, see text and Online Appendix III for details.

[b]: The names of query sequences consist of BOLD sequence accession numbers (a dash was removed before the last two numbers) and their true species names. Only part of the results were presented here, see Online Appendix III for all 484 queries and corresponding assignments (singletons were excluded since they can only be assigned to the wrong speices).

[c]: DV denotes DV-Curve, RBF indicates RBF neural network, see text for details.

[d]: FJ denotes FJ-Curve.

[e]: Ticks and crosses indicate correct and wrong assignments respectively.

doi:10.1371/journal.pone.0030986.t003

perform equally well in all circumstances of DNA barcoding. Machine learning based approaches [15,45] which are neither classical population nor phylogeny based approaches, present fresh insights. The newly developed method here may be thought of as an extension of BP-based species identification [15], in the sense that both are based on machine learning, but it uses entirely different algorithms that apply the power of both bioinformatics and RBF neural networks (NN). The reason for choosing RBF NN is that it has been shown to work well when there are complex or highly non-linear relationships and relatively small training sets, which is the case for the sophisticated process of species assignments from DNA sequences. When the input data to an algorithm is too large to be processed, then the input data will be transformed into a reduced representation set of features (termed the features vector). Transforming the input data into the set of features is termed feature extraction. In DNA sequences, each site is treated as a feature. A simple n-gram approach is also commonly used for creating feature vectors, but this proved to be five times slower than NN methods in text categorization classification [69]. In Zhang et al. [15], DNA sequences were digitized simply using the codes A-0.1, T-0.2, G-0.3, C-0.4, and this proved to be successful. However, the converted input matrices are so huge that the training of NN becomes quite slow especially for large datasets. Both the DV-Curve and the FJ-Curve substantially reduce the data matrix dimensions from, for example, the 648 of standard COI barcodes to 24 (DV-Curve) or less (FJ-Curve). This property greatly improves computational speed when processing large datasets compared to BP-based species identification [15].

We also note that while this is a powerful approach, and one that is especially well suited for non-coding sequences such as ITS,

Table 4. Species assignments for the brown algae (BOLD project PHAEP) based on ITS sequences for 1094 random queries using DV-RBF and FJ-RBF methods.

No.	Category of Random Tests[a]	Query[b]	DV-RBF[c]	Status	FJ-RBF[d]	Status
1	random	FUCUI04008-Fucus distichus	*Fucus distichus*	(✓[e])	*Fucus distichus*	(✓[e])
2	splits	FUCUI03708-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
3		FUCUI04408-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
4		FUCUI03408-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
5		FUCUI00308-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
6		FUCUI05508-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
7		FUCUI02608-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
8		FUCUI04508-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
9		FUCUI05708-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
10		FUCUI04608-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
11		FUCUI02708-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
12		FUCUI00108-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
...		...	...	...	...	...
340	(68×5)	MACRO97608-Scytosiphon cylindricus	*Scytosiphon cylindricus*	(✓)	*Scytosiphon cylindricus*	(✓)
1	n-fold	MACRO69407-Saccharina latissima	*Saccharina latissima*	(✓)	*Saccharina latissima*	(✓)
2	cross-	MACRO12106-Saccharina latissima	*Saccharina latissima*	(✓)	*Saccharina latissima*	(✓)
3	validation	MACRO77607-Scytosiphon sp	*Scytosiphon sp*	(✓)	*Scytosiphon sp*	(✓)
4		MACRO11406-Scytosiphon cylindricus	*Scytosiphon cylindricus*	(✓)	*Scytosiphon cylindricus*	(✓)
5		MACRO12806-Scytosiphon cylindricus	*Scytosiphon cylindricus*	(✓)	*Scytosiphon cylindricus*	(✓)
6		MACRO49807-Saccharina latissima	*Saccharina latissima*	(✓)	*Saccharina latissima*	(✓)
7		MACRO17406-Petalonia sp	*Petalonia sp*	(✓)	*Petalonia sp*	(✓)
8		MACRO94108-Petalonia sp	*Petalonia sp*	(✓)	*Petalonia sp*	(✓)
9		FUCUI05308-Fucus spiralis	*Fucus spiralis*	(✓)	*Fucus spiralis*	(✓)
10		FUCUI00608-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)
11		MACRO104108-Saccharina latissima	*Saccharina latissima*	(✓)	*Saccharina latissima*	(✓)
12		MACRO73607-Scytosiphon cylindricus	*Scytosiphon cylindricus*	(✓)	*Scytosiphon cylindricus*	(✓)
...		...	...	...	...	...
207	$(41 \times 5 + 2)$	FUCUI00708-Fucus distichus	*Fucus distichus*	(✓)	*Fucus distichus*	(✓)

[a]: Two categories of randomization were performed in this study. One is random splits which were conducted at species level (5 times) and the other is n-fold cross-validation which was performed on the whole dataset ($n = 5$ was used). 1094 random queries were generated based on the original 207 brown algae ITS sequences, see text and Online Appendix IV for details.

[b]: The names of query sequences consist of BOLD sequence accession numbers (a dash was removed before the last two numbers) and their true species names. Only part of the results were presented here, see Online Appendix IV for all 1094 queries and corresponding assignments (singletons were excluded since they can only be assigned to the wrong speices).

[c]: DV denotes DV-Curve, RBF indicates RBF neural network, see text for details.

[d]: FJ denotes FJ-Curve.

[e]: Ticks and crosses indicate correct and wrong assignments respectively.

doi:10.1371/journal.pone.0030986.t004

it is not without problems. Like most currently used barcoding methods, it will assign a query to ''the most like'' species when the true species is not represented in the reference library. The issue of an incomplete reference database has been well explored in Ross *et al.* [40] and Ekrem *et al.* [70]. A new fuzzy set based species identification protocol has shed some light on this issue [71]. Unlike the BP-based method [15], the second limitation of the new method is that neither DV-RBF nor FJ-RBF approaches have the potential to incorporate non-DNA data into the system. Where several different sources of data are available, such as morphological characters or behavioural data, we would instead suggest using the BP-based approach proposed earlier [15].

Materials and Methods

Graphical Representation of DNA Sequences via Bioinformatic Approaches

The DV-Curve. The DV-Curve (Dual-Vector Curve) was proposed by Zhang [67] as a 2D graphical representation for the visualization and analysis of DNA sequences (Figure 2). It proved to be a good visualization for representing DNA sequences without degeneracy and loss of information. Let us consider a DNA sequence $S = s_1 s_2 \cdots s_n$ consisting of n nucleotide sites. Let (X_i, Y_i) be the point of the DV-Curve, where $(X_0, Y_0) = (0,0)$ is the start point. The DV-Curve is uniquely determined by the following formula [67]:

$$Y_{2i-1}=\begin{cases} Y_{2i-2}+1, & if\ s_i=A\ or\ T \\ Y_{2i-2}-1, & if\ s_i=C\ or\ G \end{cases} \quad (1)$$

$$Y_{2i}=\begin{cases} Y_{2i-1}+1, & if\ s_i=A\ or\ C \\ Y_{2i-1}-1, & if\ s_i=T\ or\ G \end{cases} \quad (2)$$

$$X_{2i-1}=2i-1 \quad (3)$$

$$X_{2i}=2i \quad (4)$$

where $i = 1,2,\cdots,n$.

The FJ-Curve. In this section, motivated by Jeffrey's ingenious work of chaos game representation (CGR) of DNA sequences [72], we propose a 3D representation of DNA sequences. Let $S=s_1s_2\cdots s_n$ be a DNA sequence, n is the length of S. First we assign the four nucleotides to the four corners of a regular tetrahedron, i.e. A, G, C, T are assigned coordinates $(-1, 1, -1)$, $(1, 1, 1)$, $(1, -1, -1)$ and $(-1, -1, 1)$ respectively. Then we construct a curve for the given DNA sequence S. The point $P_i(x_i,y_i,z_i)$ corresponding to s_i is calculated by the following formula:

$$\begin{cases} x_i=\frac{1}{2}(x_{i-1}+x_{s_i}), \\ y_i=\frac{1}{2}(y_{i-1}+y_{s_i}), \\ z_i=\frac{1}{2}(z_{i-1}+z_{s_i}). \end{cases} \quad (5)$$

$i=1,2,.....,n$, $(x_0,y_0,z_0)=(0,0,0)$ and x_{s_i}, y_{s_i} and z_{s_i} are calculated by the following formula:

$$x_{s_i}=\begin{cases} -1, & \text{if } s_i\in\{\text{A,T}\}, \\ 1, & \text{if } s_i\in\{\text{C,G}\}. \end{cases}$$
$$y_{s_i}=\begin{cases} -1, & \text{if } s_i\in\{\text{T,C}\}, \\ 1, & \text{if } s_i\in\{\text{A,G}\}. \end{cases} \quad (6)$$
$$z_{s_i}=\begin{cases} -1, & \text{if } s_i\in\{\text{C,A}\}, \\ 1, & \text{if } s_i\in\{\text{T,G}\}. \end{cases}$$

In this way, S is converted into a series of points $P_1,P_2,\cdots,P_n$. Let the origin (0,0,0) be the point P_0. As the index i runs from 0 to n, we connect the points $P_0,P_1,P_2,\cdots,P_n$ in turn and get a zigzag 3-D curve within a regular tetrahedron. This is the FJ-Curve of DNA sequence S (named after one of the Authors (Dr. Feng Jie) of this study).

From the FJ-Curve, some information on the base distribution and composition of the DNA sequence can be intuitively gathered. As an example, the FJ-curve for the twenty base length sequence GCCTCCGCCCAGACTTCTTC is shown in Figure 3. It is evident that most points are located near the vertex C $(1, -1, -1)$, a consequence of the high proportion of C content in the sequence. On the other hand, because the A content is the lowest, the points near the vertex A are sparse.

Numerical Characterizations of the DV-Curve and the FJ-Curve. To numerically characterize a DNA sequence via the DV-Curve, a 24-component vector $\vec{D}$ as described by Zhang [67] was used:

$$\vec{D}_{DV}=[CM1_{xy},CM2_{xy},\cdots,CM24_{xy}] \quad (7)$$

The CM_{xy} value [73] is calculated as follows:

$$(X_c,Y_c)=(\frac{1}{2n+1}\sum_{j=0}^{2n}X_j,\frac{1}{2n+1}\sum_{j=0}^{2n}Y_j) \quad (8)$$

$$CM_{xy}=\frac{1}{2n+1}\sum_{j=0}^{2n}(X_j-X_c)(Y_j-Y_c) \quad (9)$$

To get equation (7), we need to assign A,T,G,C to basic Dual-Vectors in 4! different ways to have $4!=24$ different DV-Curves for a given DNA sequence. The vector $\vec{D}_{DV}$ is further used as the input for a neural network.

We derive a set of numerical characterizations from the FJ-Curve of the DNA sequence as sequence descriptors: $(CM_{xy},CM_{xz},CM_{yz},\lambda_L,\lambda_M)$. The first three descriptors [74] are calculated as follows:

$$\begin{cases} CM_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-x_c)(y_i-y_c), \\ CM_{xz}=\frac{1}{n}\sum_{i=1}^{n}(x_i-x_c)(z_i-z_c), \\ CM_{yz}=\frac{1}{n}\sum_{i=1}^{n}(y_i-y_c)(z_i-z_c), \\ (x_c,y_c,z_c)=(\frac{1}{n}\sum_{i=1}^{n}x_i,\frac{1}{n}\sum_{i=1}^{n}y_i,\frac{1}{n}\sum_{i=1}^{n}z_i). \end{cases} \quad (10)$$

The fourth descriptor is selected from the L/L matrix [74], in which the elements $l_{i,j}$ are defined as the quotient of the Euclidean distance between a pair of vertices (dots) of the FJ-Curve and the sum of distances between the same pair of vertices measured along the characteristic curve. In other words

$$l_{i,j}=\frac{d_{i,j}}{\sum_{k=i}^{j-1}d_{k,k+1}} \quad (11)$$

where $d_{i,j}$ is the Euclidean distance between a pair of vertices. λ_L denotes the leading eigenvalue of the L/L matrix. The last descriptor is selected from the M/M matrix [75], in which the elements $m_{i,j}$ are given as the quotient of the Euclidean distance between two vertices of the FJ-Curve and the graph theoretical distance between the two vertices. In other words

$$m_{i,j}=\frac{d_{i,j}}{|i-j|} \quad (12)$$

where $d_{i,j}$ is the Euclidean distance between a pair of vertices. λ_M denotes the leading eigenvalue of the M/M matrix. To maximally extract information from DNA sequences, we here used both L/L and M/M matrices so there may be some overlap in information (i.e., redundant information) in the matrix representations. We therefore applied Principal Component Analysis (PCA) [76] to the matrix representations in order to reduce the correlation between L/L and M/M matrices. Principal Components whose contributions to total variation are less than 0.01 were ignored in the subsequent analysis.

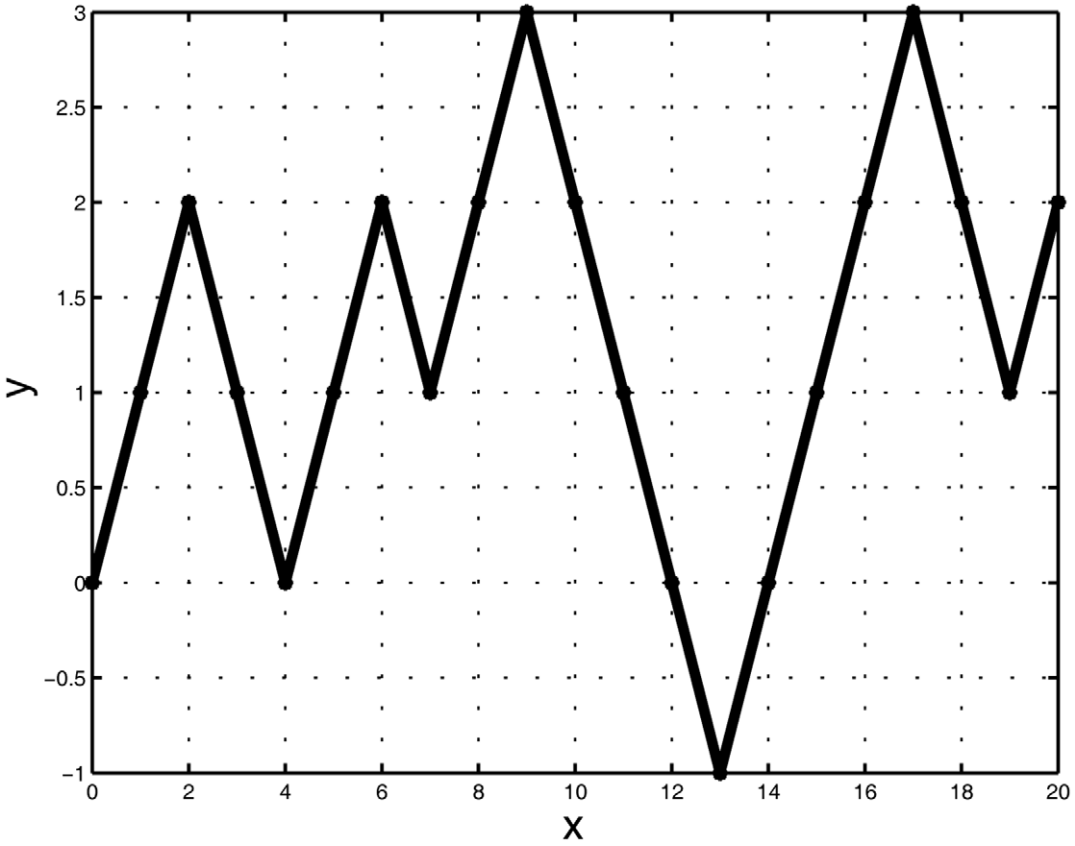

Figure 2. The DV-Curve of the 10 bp sequence 'AGACTGCATC'.
doi:10.1371/journal.pone.0030986.g002

Radial Basis Function Neural Network

The BP-Neural Network was initially proposed by Zhang and colleagues [15,45] to identify species in DNA barcoding and proved to be successful in both computer simulations and empirical studies. However, two inherent drawbacks of the BP-Neural Network preclude its wide application to DNA barcoding campaigns: its slow training for large reference datasets and potential local minimization during network training. In this study, we propose to use the Radial Basis Function (RBF) neural network which creates a network with zero error on training vectors. RBFs are embedded in a two-layer feed-forward neural network that is characterized by a set of inputs and outputs (Figure 4 and 5). Between the inputs and outputs there is a layer of processing units called hidden units, each of which implements a radial basis function [77].

The Gaussian activation function for RBF network is given by

$$a = radbas(\sigma) = e^{-\sigma^2} \tag{13}$$

The input of hidden units is given by

$$\sigma = \sqrt{\sum_{i=0}^{n} (w_i - d_i)^2 \cdot b} \tag{14}$$

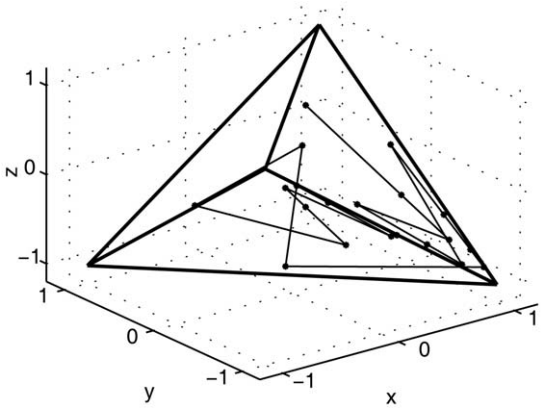

Figure 3. The FJ-Curve of the 20 bp sequence 'GCCTCCGCCCA-GACTTCTTC'.
doi:10.1371/journal.pone.0030986.g003

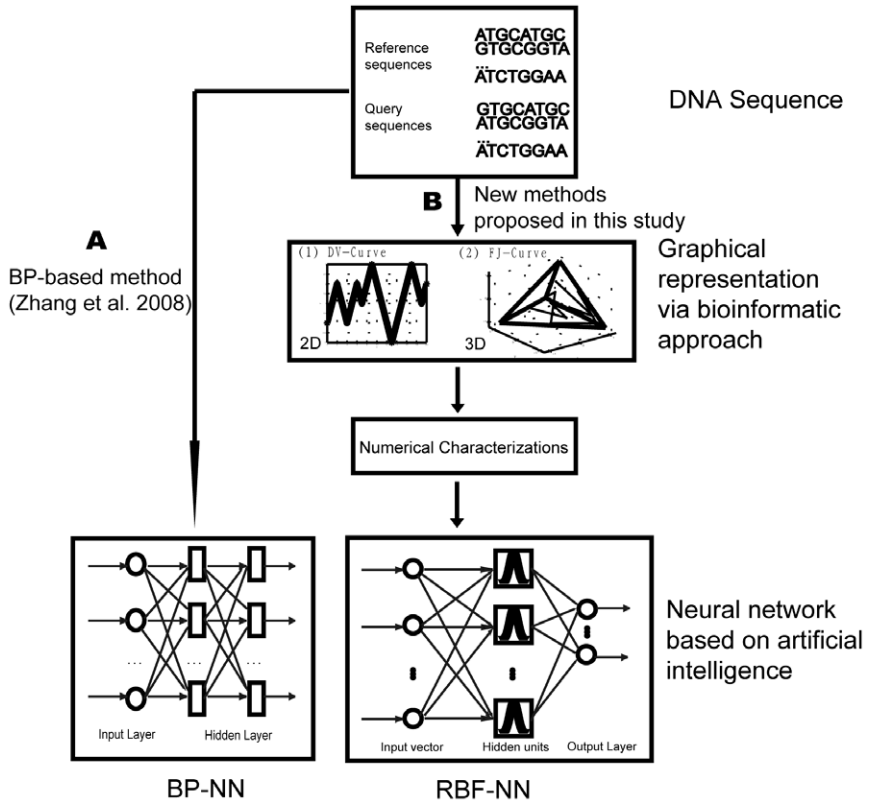

Figure 4. The work flow of the RBF network approach proposed in this study and a comparison with the BP network.
doi:10.1371/journal.pone.0030986.g004

where w_i is the input vector and d_i is the weight of hidden units.

We, therefore, have

$$a = e^{-(\sqrt{\sum_{i=0}^{n}(w_i-d_i)^2}\cdot b)^2} = e^{-(\|W-D\|)^2} \quad (15)$$

The output layer implements a weighted sum of hidden-unit outputs:

$$o_k = \sum_{j=1}^{q} w_{jk}\sigma_j, k = 1,2,\cdots,s \quad (16)$$

Species Identification via DNA sequences in DNA Barcoding

Training a network using reference sequences. Instead of simply encoding the raw DNA sequences as inputs of a neural network [15], we here employed both bioinformatic approaches (the DV-Curve and the FJ-Curve) and machine learning for species identification. The numerical characterizations of the DV-Curves and the FJ-Curves computed earlier were fed into the RBF neural networks as inputs. The former is termed the DV-Curve based RBF (DV-RBF) and the latter the FJ-Curve based RBF (FJ-RBF). The network takes a matrix of input vectors D and target vectors O. The training will return a network with weights and biases b such that the outputs O_k are exactly O when the inputs are D. Generally, during network training, the weights of hidden units are set to D' and each bias in the hidden units is set to a value which is determined by the width of an area in the input space to which each neuron responds. The second-layer (output layer) weights and biases are computed by simulating the first-layer outputs $A = [a_1,...,a_q]$, and then solving a linear expression

$$[W_{2k}, b_2] * [A; 1] = O \quad (17)$$

Since the inputs to the second-layer A and the target (O) and the layer are linear, we can use the following formula to calculate the weights and biases of the output layer to minimize the sum-squared error:

$$W_b = O/[D; ones(1,q)] \quad (18)$$

where W_b contains both weights and biases, with the biases in the last column, and $ones(1,q) = [1,...,1]$.

Identifying query sequences using a trained network. The query sequences were firstly transferred into a numeral matrix using the method described above (the DV-Curve or the FJ-Curve), which served as the input vector (Figs. 4 and 5). Then, the input vector was fed into the trained network, and one output row vector, corresponding to a different species following the formula of Zhang *et al.* [15], was obtained for each input vector. The output vector of the network for one sequence selected from, for example, species 1, could be like '(1,0,0,0)' in the case of four species in the reference library through activation of the competitive function.

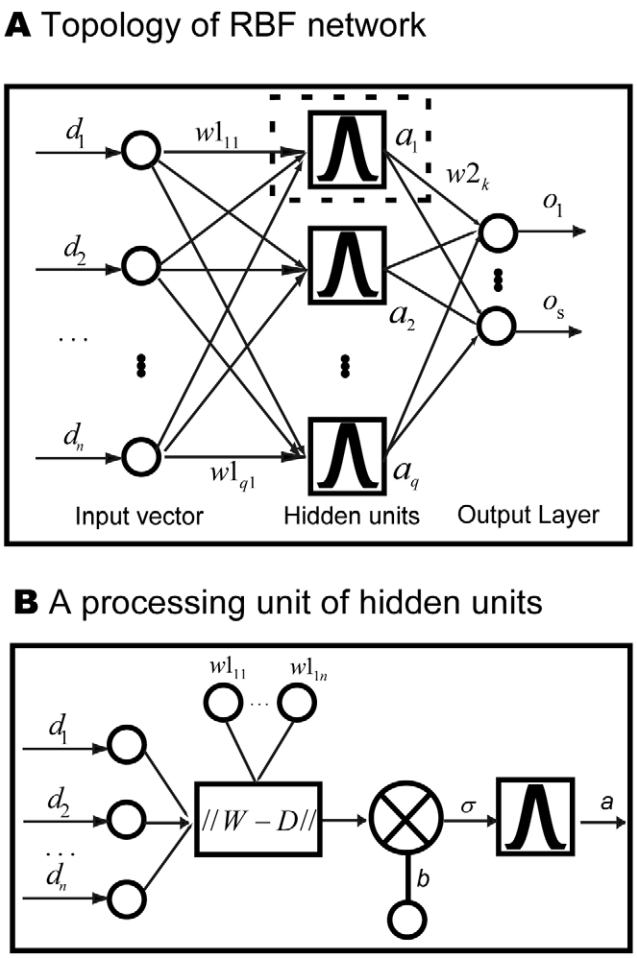

Figure 5. Topology of the RBF network and a processing unit of hidden units.
doi:10.1371/journal.pone.0030986.g005

Success Rate of Species Identification and Confidence Intervals

The success rate of species identification is defined as the following formula [15]:

$$Rate_{success} = \frac{Number_{hit}}{Number_{test}} \tag{19}$$

Binary data indicating the presence (successful identification) or absence (failed identification) of a specific attribute are often modeled as random samples from a Bernoulli distribution with parameter *prob*, where *prob* is the proportion in the population with that attribute. A $(1-\alpha)$-level confidence interval (CI) for *prob* is calculated by the following formula [78]:

$$\frac{(\widehat{prob}-\beta)}{(1+\frac{z^2}{n})} \le prob \le \frac{(\widehat{prob}+\beta)}{(1+\frac{z^2}{n})} \tag{20}$$

where $\alpha=0.05$, $\beta=\sqrt{\frac{\widehat{prob}(1-\widehat{prob})z^2}{n}+\frac{z^4}{4n^2}}$, $z=z_{\alpha/2}$.

Comparison to the Existing Methods with Repeated Random Splits and n-fold Cross-validation

We wished to determine how our new methods for species identification compare with traditional DNA barcoding approaches, including Neighbor-joining (NJ) [44] and Maximum likelihood [62]. We did this utilising both repeated random splits and n-fold cross-validation. There are some differences between these two randomization strategies. For the former, sequences from an original empirical dataset were divided into two sub-datasets: a reference set and a query set. The reference set comprised all sequences from species with samples of 3 or less, together with two thirds (or as close as possible) of sequences from species with samples of 4 or greater. Remaining sequences formed the query set. This process was repeated 5 times. The resultant reference set has a 'complete/balanced' species coverage since random splits were performed at the level of species, and all species in the original database will be kept in the reference library. In n-fold cross-validation, data were split into *n* partitions and a subset from the *nth* partition used to validate the success rate estimated from the remaining data. We here used a five-fold cross-validation to examine all methods under study. The subsequent reference library will have an 'incomplete/unbalanced' species coverage since the random partition was conducted on the whole original dataset, and therefore not all species are guaranteed to be included in the reference library. For the two traditional methods, each query from the query set was selected to form a new dataset with all reference sequences, and a genetic distance matrix was generated with the K2P model [79]. An NJ tree was then constructed with PAUP*beta [80] and an ML tree built with PHYML [62]. A successful identification was counted when a query fell into a monophyletic species clade. Species identification success rates were estimated over all random queries and 95% CI estimated with equation (20). For simplicity, the success rate from all 5-fold cross-validations was combined for the confidence interval estimate, although the pooling of results from 5-fold cross-validations could underestimate the CI. However, this underestimation was treated as trivial in this study since sampling sizes were generally large (much larger than 30). For all methods, singletons in the query set not represented in the reference set were not counted when calculating success rates, since these singletons would necessarily be assigned to the wrong species.

COI Datasets

Neotropical Bat Dataset and Marine Fish Dataset. The COI dataset of 87 Neotropical species from 47 genera of bat in Guyuna contained 819 COI sequences with lengths greater than 600 bp [81]. These were downloaded from the Barcode of Life Database (BOLD, www.barcodinglife.org) on May 10, 2010. We cleaned the dataset by removing sequences with ambiguous sites, such as "Ns", and those whose length were less than 648 bp (the standard length in COI DNA barcoding) [1–4]. This gave 766 COI sequences from 84 species. The second COI dataset was North Pacific fish. Steinke *et al* barcoded 201 North Pacific fish species, yielding 1225 barcode sequences [82]. We downloaded these from BOLD project TZFPC, Fishes of Pacific Canada Part I, on May 10, 2010. Read lengths were about 655 bp long. To reduce the effect of ambiguous sites on the analysis, we again filtered the dataset by removing uncertain nucleotide sites, such as "Ns". The 982 resultant 652 bp alignments were used in the subsequent analysis. Meanwhile, five-fold cross-validation was performed as well (Table 1 and 2).

ITS Datasets

Rust fungi dataset and Brown algae dataset. The rust fungi dataset comprised 108 ITS sequences from 16 species in

BOLD (project CHITS, Chrysomyxa ITS Barcoding; downloaded on June 4, 2010). The length of these sequences varied from 625 bp to 792 bp, excepting one sequence of 333 bp. The dataset was cleaned as above by removing sequences with ambiguous sites (e.g. 'Ns'). 85 sequences representing 14 species remained for the subsequent analysis. An initial alignment of the sequences was made using the program MUSCLE [83] with the default setting to check the homology of the sequences as a whole. All the indels (gaps) introduced during the alignment were eliminated later, the sequences for the subsequent analysis thus remained unaligned. 216 ITS sequences belonging to 20 species from seven genera of brown algae were retrieved from BOLD (project PHAEP; Phaeophyceae published, downloaded on June 20, 2010). These sequences cover a broad diversity of brown algae (six families from four orders). Sequences containing ambiguous sites were removed, and the resultant 207 sequences were highly variable in length (387 bp to 1215 bp).

Supporting Information

Table S1 Species assignments for Neotropical bats [81] based on COI sequences for all 4122 random queries using DV-RBF and FJ-RBF methods in details.
(XLS)

Table S2 Species assignments for Pacific Canadian marine fish [82] (BOLD project TZFPC) based on COI sequences for all 5134 random queries using DV-RBF and FJ-RBF methods in details.
(XLS)

Table S3 Species assignments for rust fungi (BOLD project CHITS) based on ITS sequences for 484 random queries using DV-RBF and FJ-RBF methods in details.
(XLS)

Table S4 Species assignments for the brown algae (BOLD project PHAEP) based on ITS sequences for 1094 random queries using DV-RBF and FJ-RBF methods in details.
(XLS)

Acknowledgments

We appreciate the constructive comments of reviewers and Dr. David S. Milstone on an earlier version of this paper.

Author Contributions

Conceived and designed the experiments: AZ. Performed the experiments: AZ JF. Analyzed the data: AZ JF QG. Wrote the paper: AZ. Participated in discussions: RW PW JW WZ.

References

1. Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc Lond B Biol Sci 270: 313–321.
2. Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc B 270(Suppl): 96–99.
3. Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proc Natl Acad Sci U S A 101: 14812–14817.
4. Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. PLoS Biol 2: 1657–1663.
5. Ebach MC, Holdrege C (2005) DNA barcoding is no substitute for taxonomy. Nature 434: 697.
6. Gregory TR (2005) DNA barcoding does not compete with taxonomy. Nature 434: 1067.
7. Marshall E (2005) Taxonomy-Will DNA bar codes breathe life into classification? Science 307: 1037.
8. Schindel DE, Miller SE (2005) DNA barcoding a useful tool for taxonomists. Nature 435: 17.
9. Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Phil Trans R Soc Lond B 360: 1805–1811.
10. Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. Phil Trans R Soc Lond B 360: 1847–1857.
11. Abdo Z, Golding GB (2007) A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. Syst Biol 56: 44–56.
12. Hajibabaei M, Singer GA, Clare EL, Hebert PDN (2007) Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. BMC Biol 5: 24. doi:10.1186/1741-7007-5-24.
13. Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet 23: 167–172.
14. Meusnier I, Singer GA, Landry JF, Hickey DA, Hebert PDN, et al. (2008) A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 9: 214.
15. Zhang AB, Sikes DS, Muster C, Li SQ (2008) Inferring species membership using DNA sequences with back-propagation neural networks. Syst Biol 57: 202–215.
16. Monaghan MT, Wild R, Elliot M, Fujisawa T, Balke M, et al. (2009) Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Syst Biol 58: 298–311.
17. Ward RD, Hanner R, Hebert PDN (2009) The campaign to DNA barcode all fishes, FISH-BOL J Fish Biol 74: 329–356.
18. Hebert PDN, Dewaard JR, Landry JF (2010) DNA barcodes for 1/1000 of the animal kingdom. Biol Lett 6: 359–62.
19. Zhang AB, He LJ, Crozier RH, Muster C, Zhu CD (2010) Estimation of sample sizes for DNA Barcoding. Mol Phylogenet Evol 54: 1035–1039.
20. Will KW, Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics 20: 47–55.
21. Moritz C, Cicero C (2004) DNA barcoding: promise and pitfalls. PloS Biol 2: 279–354.
22. Prendini L (2005) Comment on 'Identifying spiders through DNA barcoding. Can J Zool 83: 498–504.
23. Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. PLos Biol 3: 2229–2238.
24. Hickerson MJ, Meyer CP, Moritz C (2006) DNA barcoding will often fail to discover new animal species over broad parameter space. Syst Biol 55: 729–739.
25. Brower AVZ (2006) Problems with DNA barcodes for species delimitation: 'ten species' of Astraptes fulgerator reassessed (Lepidoptera: Hesperiidae). Syst Biodivers 4: 127–132.
26. Lefebure T, Douady CJ, Gouy M, Gibert J (2006) Relationship between morphological taxonomy and molecular divergence within Crustacea: proposal of a molecular threshold to help species delimitation. Mol Phylogenet Evol 40: 435–447.
27. Meier R, Shiyang K, Vaidya G Ng PKL (2006) DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. Syst Biol 55: 715–728.
28. Wiemer M, Fiedler K (2007) Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). Front Zool 4: (doi: 10.1186/1742-9994-4-8).
29. Whitworth TL, Dawson RD, Magalon H, Baudry E (2007) DNA barcoding cannot reliably identify species of the blowfly genus Protocalliphora (Diptera: Calliphoridae). Proc R Soc B 274: 1731–1739.
30. Meier R, Zhang G, Ali F (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the barcoding gap and leads to misidentification. Syst Biol 57: 809–813.
31. Song H, Buhay JE, Whiting MF, Crandall KA (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc Natl Acad Sci U S A 105: 13486–91.
32. Silva-Brando KL, Lyra ML, Freitas AV (2009) Barcoding lepidoptera: current situation and perspectives on the usefulness of a contentious technique. Neotrop Entomol 38: 441–51.
33. Spooner DM (2009) DNA barcoding will frequently fail in complicated groups: An example in wild potatoes. Am J Botany 96: 1177–1189.
34. Lou M, Golding GB (2010) Assigning sequences to species in the absence of large interspecific differences. Mol Phylogenet Evol 56: 187–194.
35. DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Phil Trans R Soc B 360: 1975–1980.
36. Nielsen R, Matz M (2006) Statistical approaches for DNA barcoding. Syst Biol 55: 162–169.
37. Elias M, Hill RI, Willmott KR, Dasmahapatra KK, Brower AV, et al. (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proc Biol Sci 274: 2881–9.
38. Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R (2008) Statistical assignment of DNA sequences using Bayesian phylogenetics. Syst Biol 57: 750–757.
39. Munch K, Boomsma W, Willerslev E, Nielsen R (2008) Fast phylogenetic DNA barcoding. Phil Trans R Soc Lond B 363: 3997–4002.

40. Ross HA, Murugan S, Li WLS (2008) Testing the reliability of genetic methods of species identification via simulation. Syst Biol 57: 216–230.
41. Kuksa P, Pavlovic V (2009) Efficient alignment-free DNA barcode analytics. BMC Bioinformatics 10(Suppl 14): S9.
42. Chu KH, Xu M, Li CP (2009) Rapid DNA barcoding analysis of large datasets using the composition vector method. BMC Bioinformatics 10(Suppl 14): S8.
43. O'Meara BC (2010) New heuristic methods for joint species delimitation and species tree inference. Syst Biol 59: 59–73.
44. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.
45. Zhang AB, Savolainen P (2008) BPSI2.0: A C/C++ interface program for species identification via DNA barcoding with a BP-neural Network by calling the Matlab engine. Mol Ecol Res 9: 104–106.
46. Rosenblatt F (1958) The Perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review 65: 386–408.
47. Werbos PJ (1974) Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD thesis Harvard University, Cambridge, Massachusetts.
48. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by backpropagating errors. Nature 323: 533–536.
49. Parker DB (1987) Optimal algorithm for adaptive networks: Second order back propagation, second order direct propagation, and second order Hebbian learning. Proc Int Joint Conference on Neural Networks 2: 593–600.
50. Smith M (1993) Neural networks for statistical modeling. New York: Van Nostrand Reinhold.
51. Wu CH (1997) Artificial neural networks for molecular sequence analysis. Computers Chem 40: 237–256.
52. Wu C, Chen H (1997) Counter-propagation neural networks for molecular sequences classification: Supervised LVQ and dynamic node allocation. Appl Intel 7: 27–38.
53. Wu C, Shivakumar S (1994) Back-progragation and counter-propagation neural networks for phylogenetic classification of ribosomal RNA. Nucleic Acids Res 22: 4291–4299.
54. Wu C, Shivkumar S, Lin H, Veldurti S, Bhatikar Y (1995) Neural Networks for molecular sequence classification. Math Compu Simu 40: 23–33.
55. Wang HC, Dopazo J, de la Fraga LG, Zhu YP, Carazo JM (1998) Self-organizing tree-growing network for the classification of protein sequences. Protein Sci 7: 2613–2622.
56. Dopazo J, Carazo JM (1997) Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. J Mol Evol 44: 226–233.
57. Breiman L, Freidman J, Olsen R, Stone C (1984) Classification and regression trees. New YorkNY: Chapman and Hall.
58. Austerlitz F, David O, Schaeffer B, Bleakley K, Olteanu M, et al. (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. BMC Bioinformatics 10(Suppl 14): S10.
59. Breiman L (2001) Random forests. Mach Learn 45: 5–32.
60. Seo (2010) Classification of Nucleotide Sequences Using Support Vector Machines. J Mol Evol 71: 250–267.
61. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
62. Guindon S, Gascuel O (2003) PhyML: A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
63. Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73: 237–44.
64. Wheeler WC (1996) Optimization alignment: The end of multiple sequence alignment in phylogenetics? Cladistics 12: 1–9.
65. Gladstein DS, Wheeler WC (1997) POY: The optimization of alignment characters. Version 3.0.11. American Museum of Natural History, New York;Program and documentation available at ftp://ftp.amnh.org/pub/molecular/poy/.
66. Giribet G (2001) Exploring the Behavior of POY, a Program for Direct Optimization of Molecular Data. Cladistics 17: S60–S70.
67. Zhang ZJ (2009) DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences. Bioinformatics 25: 1112–7.
68. Frzal L, Leblois R (2008) Four years of DNA barcoding: current advances and prospects. Infect Genet Evol 8: 727–736.
69. Babu AS, Pavan Kumar PNVS (2010) Comparing neural network approach with Ngram approach for text categorization. Int J Comput Sci Engin 2: 80–83.
70. Ekrem T, Willassen E, Stur E (2007) A comprehensive DNA library is essential for identification with DNA barcodes. Mol Phylog enet Evol 43 43: 530–542.
71. Zhang AB, Muster C, Liang HB, Zhu CD, Crozier R, et al. (2011) A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. Mol Ecol;. doi: 10.1111/j.1365-294X.2011.05235.x.
72. Jeffrey HI (1990) Chaos game representation of genestructure. Nucleic Acids Res 18: 2163.
73. Liao B, Tan M, Ding K (2005) Application of 2-D graphical representation of DNA sequence. Chem Phys Lett 414: 296–300.
74. Wang WP, Liao B, Wang TM, Zhu W (2006) A graphical method to construct a phylogenetic tree. Int J Quantum Chem 106: 1998–2005.
75. Randic M, Vracko M, Lers N, Plavsic D (2003) Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett 368: 1–6.
76. Jolliffe IT (2002) Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed. New York: Springer. 28 p.
77. Park J, Sandberg JW (1991) Universal approximation using radial basis functions network, Neural Computation 3: 246–257.
78. Tamhane AC, Dunlop DD (2000) Statistics and Data Analysis: from Elementary to Intermediate, 1st Edition. Saddle RiverNJ: Pearson Education, Inc. Publishing as Prentice Hall. 288 p.
79. Kimura M (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol 16: 111–120.
80. Swofford DL (2002) PAUP*, Phylogenetic Analysis Using Parisimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, MA.
81. Clare El, Lim BK, Engstron MD, Eger JL, Hebert PDN (2007) DNA barcoding of neotropical bats: species identification and discovery within Guyana. Mol Ecol Notes 7: 184–190.
82. Steinke D, Zemlak TS, Boutillier JA, Hebert PDN (2009) DNA barcoding of Pacific Canada's Fishes. Mar Biol 156: 2641–2647.
83. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.