A Profile-Based Probabilistic Approach for the Detection of Anomalies in the Cytochrome C Oxidase I Amplicon Sequences

Mahdi Belcaid*
Hawaii Institute of Marine Biology
University of Hawaii at Manoa
Honolulu, Hawaii
mahdi@hawaii.edu

Guylaine Poisson
Information and Computer Sciences
University of Hawaii at Manoa
Honolulu, Hawaii
guylaine@hawaii.edu

ABSTRACT

The cytochrome c oxidase 1 (COI) gene is among the most popular markers for molecular biodiversity estimation. In essence, COI-based approaches for taxonomic identification rely on comprehensive reference databases to assign unknown sequences to known species and/or to enhance the identification of new species. As such, for COI-based methods to be effective, the accuracy and integrity of reference databases are critical. However, as COI repositories grow, it becomes difficult to manually curate and validate user-contributed data. This, in turn, propagates prediction errors, therefore reinforcing the cycle. Here, we propose a new computationally efficient approach for identifying anomalies which are either due to systematic biases (indels and chimeras) or to user error (mistranslation and misclassification). Our approach uses COI reference alignments to model substitutions across the marker. The resulting model is subsequently used to screen and identify sequences with incongruous fit to the model. Analysis of the complete set of curated Insecta COI reference sequences identify the presence of numerous anomalous sequences, which makes a strong case for the importance of new strategies to screen publicly available COI references.

CCS Concepts

•**Applied computing** → *Molecular sequence analysis; Sequencing and genotyping technologies; Bioinformatics;*

Keywords

COI; Diversity estimation; Sequencing anomalies.

*To whom correspondence should be addressed

SAC 2017,April 03-07, 2017, Marrakech, Morocco

http://dx.doi.org/10.1145/3019612.3019614

1. INTRODUCTION

Accurate species identification is critical for estimating biodevirsity and for better understanding the effects of various anthropogenic and natural disturbances on the underlying biological landscape. For taxa that are well studied, visual cataloguing and identification of morphological features by an expert taxonomist is the gold standard for species identification [2]. However, in addition to intimate knowledge of particular life stages and gender differences of each of the studied species, expert-led taxonomic identification is time consuming and requires high sample integrity [10]. These constraints render this approach inadequate in highly biodiverse or complex environments from which organisms cannot be extracted intact. Increasingly, large biomonitoring efforts have resorted to molecular approaches where DNA-based markers are brought to bear on evaluating species diversity [10]. The mitochondrial gene cytochrome c oxidase I (COI) is a popular marker for the identification of members in the kingdom Animalia, due to its ease of sequencing in most, if not all, animal phyla [10, 9]. Additionally, the primary COI sequence, which consists of a 648-bp coding region, offers greater range of phylogenetic signal, particularly at the highly diverse third-position nucleotides compared to other markers such as 12S or 18S ribosomal markers [13].

For molecular-based methods to be effective, the assembly of a comprehensive reference database is critical. This is particularly important in light of current diversity estimation procedures in high-throughput sequencing projects, which rely on sequence similarity, as well as sequence divergence patterns within and across taxonomic levels, to estimate the number and nature of species in a sample [18]. The set of available COI reference sequences is collated in specialized archives such as: the Barcode of Life Database (BOLD: http://www.boldsystems.org), a library of barcodes for eukaryotic life [18]; and the Moorea BioCode Project (http://mooreabiocode.org/), a comprehensive inventory of all non-microbial life in a complex tropical ecosystem. COI markers are also submitted to public sequence databases, such as the NCBI's Genbank. As is the case with most DNA sequence databases, while the quality of the submitted sequences is traditionally screened for major discrepancies, ultimate responsibility for quality and origin accuracy of the data lies with the project participants [18]. As such, as the number of COI sequences submitted to specialized databases

increases, the likelihood for introducing erroneous references also increases in more than trivial ways. For instance, library preparation and sequencing error rates can exceed 4% in state-of-the-art protocols for amplicon sequencing on an Illumina MiSeq instrument [19]. As such, Inaccuracies in reference databases are inevitable.

Current protocols for the identification of COI data anomalies are principally based on tools for the analysis of the non-coding, microbial 16S marker. These tools do not leverage the coding-nature of the COI sequence, other than to test for the absence of a stop in the most likely open reading frame (ORF) of a sequence [18]. To our knowledge, no COI specific approach has been proposed for detecting systematic errors that do not cause the emergence of stop codon in all ORFs of an amplicon sequence.

In preliminary simulations on a subset of 1000 sequences downloaded from the BOLD database (data not included), random insertions or deletions (indels) did not give rise to a stop codon in more than 15% of the sequences. This shows that relying solely on the detection of premature stops in ORFs is not sufficient. In addition to indels, chimeras, hybrid sequences from multiple parents, are a major source of bias in specialized marker databases. Chimeric sequences arise predominantly during the PCR amplification stage when an incomplete PCR fragment serves as a primer by binding the template DNA of different species [21]. According to Porazinska *et al.* [17], chimeras can account for as much as 46% of the sequence data. Chimeras can also arise, albeit to a lesser extent, from the fusion of non-chimeric sequences [20]. Popular tools in microbial ecology, such as uchime [4] or ChimaraSlayer [6] can be used for detecting COI chimeras. These tools can achieve great sensitivity when the chimera fragments originate from evolutionarily distant sequences. However, due to their computational complexity, chimera detection tools can be omitted from pipelines in favor of validation using a sequence's cardinality as proxy for its quality – for example, by discarding any sequence that does not occur in more than one sample. This, at least partially, explains why chimeras are still routinely identified in curated 16S databases despite the abundance of tools for detecting them [15].

Exogenous contamination is yet another source of bias in the public databases. By exogenous contamination, we refer to kingdom-level COI sequences mismatches. This would, for instance, arise if a bacterial sequence was recovered in a project focused on animals. A common practice for handling exogenous contamination is by screening against known, exogenous references [18]. This approach is ideal for filtering out known contamination, but cannot detect novel sequences for which a reference is not readily available.

Given the crucial role that specialized reference databases play in diversity estimation, rigorous inspection and detection of erroneous sequences is critical. Indeed, systematic errors not only lead to the significant overestimation of diversity, but can also bias measures of intra- and inter-taxonomic level sequence similarity, which, in turn, leads to future erroneous taxonomic assignments.

In this work, we present a COI sequence anomaly detection model that leverages the functional constraints on the COI amino acid sequence to identify regions of the protein that deviate from the profile of amino acid substitions observed at a taxonomic level. Our hypothesis is based on the assumption that given a large enough sample of correct sequences, anomalous sequences will appear as statistical outliers. Our approach works in three distinct steps: 1- sequence filtering and core COI set identification 2- score profile inference and 3- score profiles modeling and identification of sequence anomalies. In the first step, a taxonomic group's multiple sequence alignment (MSA) is constructed using an iterative approach. In the second step, the similarity score for each sequence in the group is computed against the profile. In the third step, a non-parametric, probabilistic model is used to detect regions that are dissimilar to the the group's multiple sequence alignment. Sequences with large stretches of deviation from the profile are considered anomalous.

By working at the amino acid level, we are able to detect artifacts of the protein which are not detectable using the nucleotide sequence. Furthermore, by choosing to model amino acid differences in each column of the multiple sequence alignment, we account for localized variability that results from functional constraints of different regions on the COI protein [14].

Our results show that this approach is effective in detecting insertions, deletions, chimeras, mis-translations and contaminations in sequences from the BOLD database. Furthermore, this approach is computationally efficient and will be a good addition to existing strategies for maintaining the integrity of specialized databases.

2. METHODS

All the COI amino acid sequences annotated to the Insecta class were downloaded from the BOLD database (Bold System v3. Dec 2015). These sequences were then preprocessed and analyzed as described below.

2.1 Sequence Filtering and Core COI Set Identification

Sequences with more than 3% of ambiguous amino acids (represented as X in the sequence) or shorter than 180 a.a. were discarded. The remaining sequences were clustered with cd-hit [5] using 99.5% similarity threshold. Representative sequences from each cluster were aligned using MAFFT's fast approximate alignment mode [11, 12]. To further reduce the size of the dataset, sequences which aligned with no indels with a parent sequence – except for possibly terminal gaps –, were temporarily removed. The remaining parent sequences were re-aligned using MAFFT's slow but sensitive mode. In what follows, we refer to this alignment as the *reference MSA*.

2.2 Score Profile Inference

Given a reference MSA of n sequences and l columns, we compute the similarity score S_{ij} of sequence i's j^{th} amino acid, where $0 < i < n$ and $0 < j < l$ as:

$$S_{ij} = \sum_{d=1}^{20} \delta(i,d) \times \frac{\#\ of\ amino\ acids\ d\ in\ position\ j}{n}$$

where δ is the BLOSUM80 substitution matrix.

The score S_{ij} represents the overall similarity of sequence i's j^{th} amino acid, to the remaining amino acids observed in that column [3]. For example, from Figure 1, the score S_{ij} for *sequence_6* at column 42 (highlighted in red in the figure) is: $S_{ij} = 3/6{\cdot}\delta(M,M)+2/6{\cdot}\delta(M,I)+1/6{\cdot}\delta(M,G) = 2/6 \cdot 6 + 2/6 \cdot 1 + 1/6 \cdot -4 = 1.66$.

To account for data biases, such as when the amino acid at position j in sequence i is rare or comes from an under sampled species in the dataset, we smooth, or average, $S_{i,j}$ using a moving average approach with a window of size m [22]. Thus, for a sequence i, the smoothed score SS_{ik} of col j is computed as:

$SS_{ik} = \sum_{j=k-m/2}^{k+m/2} S_{ij}$ for $k < (l - m/2)$

We refer in what follows to the set of smoothed scores for sequence i, $SS_{i*} = [SS_{i1}, SS_{i2}, \ldots, SS_{il}]$ as the *score profile* (See Figure 2A).

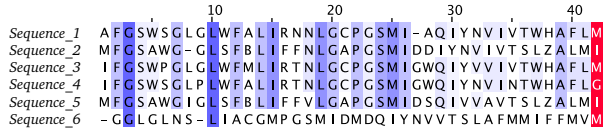

Figure 1: Multiple Sequence alignment of 6 Sequences

2.3 Score Profiles Modeling and Identification of Sequence Anomalies

At each column k of the reference MSA, we model the distribution of score profiles $SS_{*,k}$, i.e., $SS_{1,k}$, $SS_{2,k}$, ..., using a univariate *Kernel Density Estimation* (KDE) with a Gaussian Kernel [16]. KDE is a non-parametric way to estimate the probability density function of a random variable from a set of observations. Conceptually, a KDE is similar to a histogram. However, instead of discretizing the data, a KDE is smooth and continuous, which allows for better density estimation of individual observations. The probability of observing a specific SS value at the same column can be quickly approximated using that column's KDE. Figure 2B, gives an example of two KDEs fitted using the scores observed at columns $k = 25$ and $k = 48$ respectively.

We chose to model $SS_{*,k}$ using a non-parametric KDE for two reasons: 1- it does not make any assumptions about the true distributions of the scores at each column; and 2- it is resilient to multimodal distributions of scores that can arise due to subclustering of sequences according to taxonomic levels. This latter consideration is important when working at higher taxonomic levels, particularly in old lineages, where sequence diversification is potentially greater and where sequences are most likely to cluster tightly according to the underlying taxonomic levels. For example, due to the large phylogenetic distance between the orders Mecoptera and Diplura in the Insecta class, COI sequences from these orders are more likely to form independent clusters, which will be represented as multimodal distribution of SS scores.

Given that the scores at each column are modeled independently, column specific KDEs can account for the differences in amino-acid diversification rates, and therefore distributions, across functionally distinct sites in the COI protein. Figure 3 shows one monomodal ($k = 164$) and two multimodal distributions ($k = 58$ and $k = 119$) from an MSA of Insecta sequences. While all three probability density functions have a similar domain (approximately [-6,6]), the probability density for some values in the range are substantially different across distributions. For example, in the distribution of column $k = 119$, the occurrence of an SS score of -4 has a higher probability, and is, therefore, less likely to represent an anomaly. However, a score of -4 is much less likely to arise in the distributions of columns $k = 58$ and $k = 164$. Similarly, an SS score of 3.75 is more probable and less likely to be anomalous in column $k = 58$ than in the distributions of columns $k = 119$ and $k = 64$

For each sequence, we were interested in identifying subsequences exhibiting poor fit against the reference MSA. This amounts to identifying outlier SS values, according to the KDE distribution of profile scores. The occurrence of a single low-probability window is not *per se* indicative of an anomalous mutation. In an MSA with a large number of sequences, a single mutation in a highly conserved amino-acid, such as in Tryptophan (W) can lead to low probability window. However, the occurrence of multiple consecutive low-probability windows is not to be expected in a normal sequence. In order to identify the longest stretches of low-probability windows in each sequence, we used Kadane's algorithm, as described in [1]. This algorithm finds the continuous subsequence that maximizes the sum of its elements, i.e., no other subsequence can add up to a value larger than that one. However, given that Kadane's algorithm takes as input positive and negative numbers, we mapped our probabilities into scores using the following equations:

$$\sigma(p,\epsilon) = \begin{cases} 50(\epsilon - p)^2 & \text{if } p \leq \epsilon \\ -50((p-\epsilon)/\sqrt{1+\mathrm{e}^{(p-\epsilon)}} & \text{if } p > \epsilon \end{cases} \quad (1)$$

This function allows us to convert a KDE-computed probability p, such that values of p where $(p \leq \epsilon)$ are converted into increasing positive values and probabilities of good-fit $(p > \epsilon)$ into increasing negative values. Here we choose ϵ as 0.05. For example, the probabilities 0.001, 0.05, 0.4 and 0.9 are converted using Equation 1 into the values 6, 0, -11, -23. We refer to the continuous subsequence that maximizes the sum of σ-converted probabilities as *Kadane's Longest Subsequence* (KLS). Briefly, a KLS represents the longest stretch where the fit between a sequence and a reference MSA is minimal – i.e., the alignment is of poor quality – according the distributions of SS scores.

3. RESULTS AND DISCUSSION

The dataset downloaded from the BOLD database consisted of 512,915 sequences from the class Insecta, spread across 16 taxonomic orders (See Table 1 for a breakdown of the number of sequences per order).

Due to the high similarity of the COI sequences at the amino-acid level, both dereplication steps achieved over 1,000-fold reduction of the dataset size (See Table 1). The resulting 1,326 core set sequences were aligned into a reference

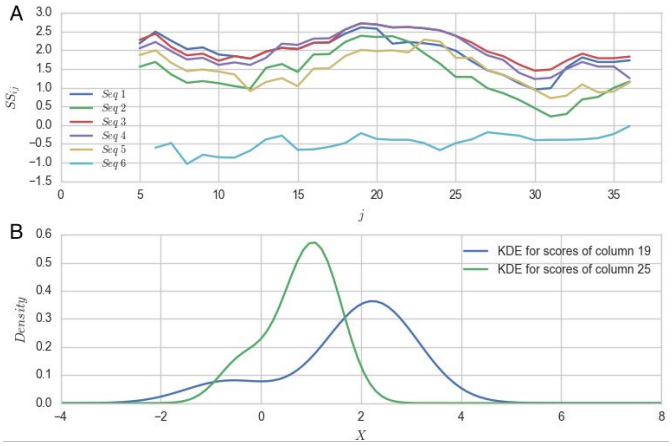

Figure 2: A) Score profiles for the sequences in the MSA depicted in Figure 1. B) KDE distribution of windows of size 5 centered around columns $k = 24$ and $k = 29$ of the same MSA.

Table 1: Breakdown of number of sequences per order during cleaning and dereplication.

Order	Raw	Cleaned	Clustered	Dereplicated
Order	Raw	Cleaned	Clustered	Dereplicated
Archaeognatha	216	204	86	6
Blattodea	2,024	1,853	690	33
Coleoptera	7,2731	68,440	25,895	140
Dermaptera	173	153	71	7
Diplura	39	34	15	3
Diptera	105,850	93,659	42,262	336
Embioptera	201	197	92	4
Ephemeroptera	3,450	3,212	933	18
Grylloblattodea	3	3	1	1
Hemiptera	42,674	37,887	15,433	180
Hymenoptera	192,672	170,616	68,998	353
Lepidoptera	91,627	86,609	22,856	226
Mantodea	947	925	418	9
Mantophasmatodea	2	2	1	1
Mecoptera	165	157	50	4
Megaloptera	141	138	32	5
Total	512,915	464,089	77,833	1,326

MSA using MAFFTs slow but sensitive mode. This reference alignment will most likely be different from what would be expected if all unique sequences were aligned. However, we conjecture that the mutation-containing sequences that were temporarily discarded are unlikely to have a substantial impact on the alignment. Thus, while minor, confined differences are inevitable between our core set alignment and one that would contain all sequences, these differences are not likely to affect the steps in our approach since our score profiles are averaged across a sliding window. Furthermore, we show below that this approach still allows us to detect numerous sequence anomalies in the dataset.

The reference MSA consisted of 957 columns and was, overall, of poor quality; it contained 238 columns represented by gaps in 1,000 or more sequences. Visual inspection of the alignment revealed that the majority of these gaps were introduced by a few misaligned sequences. The SS scores and their probabilities were computed for all sequences against the reference alignment using a window, m, of size 7. Here, the value of m is used to control the length of the anomalous region in a sequence. Smaller values of m can be used to identify short regions that do not align well against the reference MSA. On the other hand, larger values of m can smooth out short anomalies. Here we chose $m = 7$ to buffer, on average three deletions or mutations of an evolutionarily conserved amino acid, such as Tryptophan (W) or a Proline (P) in our reference MSA. The relationship between the reference MSA size, the number of allowed errors and m will be further explored in future work. This step took only a few seconds to run on a commodity computer.

Under the naïve assumption that the probabilities for a sequence i's SS scores are independent, various approaches can be used to decide wether sequence i represents an anomaly. For example, one can compound, average, or sum the individual probabilities. Alternatively, one can set a minimum threshold on the lowest observed probability among all windows, or set a maximum allowable number of consecutive low-probability windows above which a sequence is deemed an anomaly. We chose to use a modification of the latter procedure. However, rather than an "all-or-none" approach where a window is dropped if it is below a certain probability, we assigned positive and negative scores to each window in a manner that is proportional to the observed probabilities, and used those positive and negative scores as input to Kadane's algorithm. This strategy is more appropriate here as it allows us to detect the longest misaligned subsequence, which can then be manually investigated. This is not possible with the other methods, since they would either flag a sequence as an anomaly without indicating where the misalignment occurred, or return poorly aligning window(s). The intuition here is similar to that of a dynamic programming local alignment algorithm.

A critical choice for transforming the probabilities into scores is the conversion function. In essence, any function that assigns large increasingly positive scores to increasingly unlikely events and increasingly small negative values to increasingly probable events is a potential candidate. The simple and most intuitive such function is a linear function with a negative slope which crosses the $x - axis$ at a predetermined threshold ϵ (ex. $y = -400x + 20$ crosses x at $\epsilon = 0.05$). Our choice of function, as described in Equation 1, allowed us to put quadratically more weight on highly likely or highly improbable windows.

We converted the probabilities into scores using Equation 1 and computed the KLS scores for each sequence in the core set. the largest KLS value was 196 and the smallest KLS value was 0. A KLS of length 0 indicates an overall good quality fit to the reference MSA – no poor aligning regions. On the other hand, a KLS of length 196 represents a subsequence of 210 amino acids (we need to account for the windows of size m=7 at both edges of the KLS) with a poor alignment to the reference MSA. Based on careful examination of sequences and their KLS values, we selected a KLS threshold of 25, as it allows for a moderate level of variance in each sequence – a KLS of 25 can arise due to a low quality alignment in a substring of 13 amino acids using a window size of $m = 7$. Naturally, the window size can be decreased to identify short MSA anomalies, or increased to focus only on anomalies that span longer ranges.

There were 72 sequences in the reference MSA (core set)

with a KLS greater than or equal to 25. These match at 98% similarity with 324 sequences in raw dataset (prior to dereplication). For sequences with KLS values below 25, the mean KLS was 3.69 with a standard deviation of 5. All the sequences with an SCS greater than 25 were manually inspected. A sequence was deemed to be anomalous if it was classified into one of following four biases: 1- mistranslations, 2- indels, 3- chimera, or 4- misclassifications. A sequence was deemed suspicious if its KLS-region aligned with poor quality to the reference MSA, but its assignment to any of the 4 categories was not unequivocal. The core COI set, the reference MSA and a complete analysis of the 72 sequence anomalies can be found at the following url: https://figshare.com/s/99bc4f716da30bba9327.

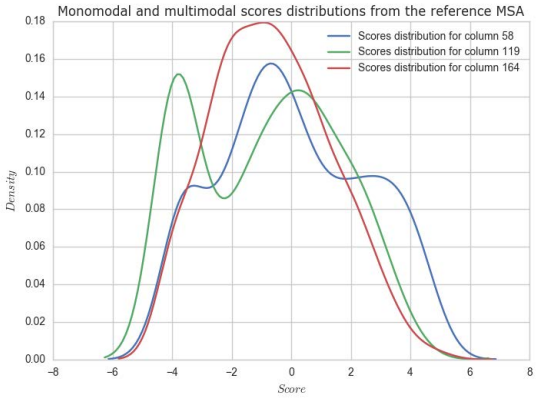

Figure 3: Steps form fit probability computation

The KLS values and the curves of profile scores are intuitive ways to think about anomalies. For instance, mistranslated sequences are trivial to identify using profile score curves; their score profile is consistently low across the complete score profile and their KLS value is close to the sequence length. These sequences are also trivial to validate, as they have an alternative, full ORF which produces a better overall alignment. One such sequence, CNKOC293-14, has a consistently negative profile score using the submitted tranlslation (second ORF on the leading strand). When using the first ORF on the reverse strand, the same sequence yields a consistently positive score.

The profile score of an indel-containing sequence is different. The DNA sequences pre- and post-indel can either produce homologous or non-homologous amino acid subsequence. The subsequence producing a homologous protein sequence has a higher profile score than the non-homologuous subsequence. This leads to a regime change in the profile score curve either from high to low or the inverse. This pattern is similar, whether the sequence is in the forward or reverse strand. For this type of error, the longest window identified by Kadane's algorithm matches the low regime in the profile score and the KLS length is a function on the location of the indel in the sequence. Indels represented the most abundant anomalies in our dataset. Sequence LEPIN057-14 is a good example as it shows a clear regime change near $k = 41$. Translation of this sequence into all reading frames shows that the second ORF produces a better fit for the first 41 amino acids of the reference MSA, whereas the first ORF produces a better fit for amino acid 41 and up against the reference MSA.

Note that an early occurrence of an indel in the leading DNA strand (or at the end in the lagging DNA strand) yields a predominantly non-homologous amino-acid sequence with a large KLS. In this case, a score profile is most similar to that of mistranslated sequence. When highly similar sequences are available in the reference database, an indel event in a COI sequence can be computationally trivial to identify by pairwise alignments (using BLAST or other alignment tools). However, indel containing sequences are difficult to detect when highly-similar sequences are lacking in the databases.

Pipelines and COI standard analysis procedures, such as the one proposed in [18] and used in BOLD, suggest screening the COI sequences against a predefined contamination database. This approach is ideal when all the sources of contamination are catalogued, but can fail to identify uncatalogued sources of contamination. For instance, despite the screening carried out by BOLD on a small suite of possible contaminants, we were able to identify sequences with substantially divergent profile scores. For instance, sequence MANT163-13, which was submitted to BOLD as a sequence in the Hymnoptera order, had a profile score curve that stood out as consistently lower that the median score. This sequence of 217 amino acids has a KLS value of 196. Blast analysis (using the NCBI's web BLASTP with default parameters) showed that its most similar sequences in the NR database were of proteobacterial origin– 49 of the top 50 BLAST hits are against bacteria in the *Legionellaceae*, with the best hits aligning with 98% similarity over the complete length of the sequence. Further investigation showed that Legionella can reside within insect hemocytes, or blood cells [8]. In fact, some insects are even used as models for studying the immune response to the Legionella infection [7]. Under the light of such evidence, we hypothesized that this sequence MANT163-13 is, in fact, from a *Legionellaceae* pathogen, rather than from the Hymnoptera host. Other sequences were also predicted using BLASTP to be of bacterial origin. For instance, the top 50 best hits of sequence CNEIE680-12 (KLS= 59) were against bacterial sequences in the *Oxalobacteraceae* family. For sequence SSWLE10476-13 (KLS= 47) which also has a consistently low profile score that matched closely with that of MANT163-13, the three most significant Blast hits are to arthropods (2 annotated at the class level – Insecta– and one annotated at the family – Staphylinidae – level). The fourth hit was against a bacterium *Rickettsiella grylli*. Interestingly, while we couldn't initially neither confirm nor deny that this sequence represents bacterial contamination, a comment posted by a BOLD database curator in November 2015 identified this sequence as contamination. As such, we hypothesize that this sequence's top Arthropoda hits are artifacts that were either used for, or which resulted from the propagation of this erroneous annotation. These examples of contamination highlight the limitations in using a reference database for contaminant screening.

4. CONCLUSION AND FUTURE WORK

The COI gene has proved to be an effective marker for a majority of organisms in the kingdom Animalia. Thus far, the process of cataloguing new species in reference libraries has mostly relied on the use of targeted sequencing. This low-throughput activity has greatly served the data curation process by keeping a handle on the number of errors in the COI reference databases. However, targeted sequencing is painstakingly slow and cannot be used to fill the large gap in sequence diversity. Thus, as researchers turn to high-throughput methods for cataloguing new COI markers – notably environmental sequencing –, automatic methods for scaling the validation of submitted COI sequences will become critical.

Here we propose a new approach that leverages the coding nature of the COI marker to probabilistically identify experimental errors that are challenging to detect at the nucleotide level. Our tests were able to identify numerous sequence anomalies in the well-curated BOLD database. Our results highlight the usefulness of this approach and make a strong case of its use as a complement to existing tools for identifying anomalous sequences in COI reference databases.

Our future goal is to address the lack of threshold for estimating the likelihood of an anomalous sequence. Given that the amino acid diversity is both position and taxonomy specific, we will focus our efforts on including these parameters to dynamically estimate a threshold based on the observed diversity within and outside the processed taxonomic level. Furthermore, we intend to expand our strategy for detecting contamination by employing probabilistic clustering of profile scores to detect, when possible, potentially misclassified sequences from lower taxonomic levels such as the order or the family. Finally, we also plan on providing an interactive graphical user interface driven environment for the exploration of profile score and multiple sequence alignments. Such application would be equipped with tools such as BLAST, DNA translation and realignment, as we have found these tools to be extremely useful during the validation of our own sequences.

5. ACKNOWLEDGMENTS

We would like to thank Dr. Emma Ransome for her helpful discussions on COI amplicon data . This work was supported by NIH-NIGMS grant# P30GM114737.

6. REFERENCES

[1] J. Bentley. *Programming Pearls*. ACM, New York, NY, USA, 1986.

[2] S. Csősz and B. L. Fisher. Toward objective, morphology-based taxonomy: A case study on the malagasy nesomyrmex sikorai species group (hymenoptera: Formicidae). *PloS one*, 11(4):e0152454, 2016.

[3] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. Biological sequence analysis: probabilistic models of proteins and nucleic acids, 1998.

[4] R. C. Edgar, B. J. Haas, J. C. Clemente, and C. Quince. UCHIME improves sensitivity and speed of chimera detection. . . . , 2011.

[5] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, Dec. 2012.

[6] B. J. Haas, D. Gevers, A. M. Earl, M. Feldgarden, D. V. Ward, G. Giannoukos, D. Ciulla, D. Tabbaa, S. K. Highlander, E. Sodergren, B. Methé, T. Z. DeSantis, Human Microbiome Consortium, J. F. Petrosino, R. Knight, and B. W. Birren. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research*, 21(3):494–504, Mar. 2011.

[7] C. R. Harding, G. N. Schroeder, J. W. Collins, and G. Frankel. Use of Galleria mellonella as a model organism to study Legionella pneumophila infection. *Journal of visualized experiments : JoVE*, (81):e50964, 2013.

[8] C. R. Harding, G. N. Schroeder, S. Reynolds, A. Kosta, J. W. Collins, A. Mousnier, and G. Frankel. Legionella pneumophila pathogenesis in the galleria mellonella infection model. *Infection and immunity*, 80(8):2780–2790, 2012.

[9] P. Hebert and A. Cywinska. Login. . . . *B: Biological* . . . , 2003.

[10] P. Hebert and S. Ratnasingham. Login. . . . *of the Royal* . . . , 2003.

[11] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–3066, July 2002.

[12] K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, Apr. 2013.

[13] N. Knowlton and L. A. Weigt. New dates and new rates for divergence across the isthmus of panama. *Proceedings of the Royal Society of London B: Biological Sciences*, 265(1412):2257–2263, 1998.

[14] D. H. Lunt, D. X. Zhang, J. M. Szymura, and G. M. Hewitt. The insect cytochrome oxidase I gene: evolutionary patterns and conserved primers for phylogenetic studies. *Insect molecular biology*, 5(3):153–165, Aug. 1996.

[15] M. Mysara, Y. Saeys, N. Leys, and J. Raes. CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies. *Applied and* . . . , 2015.

[16] E. Parzen. On Estimation of a Probability Density Function and Mode on JSTOR. *The annals of mathematical statistics*, 1962.

[17] D. L. Porazinska, R. M. Giblin-Davis, and W. Sung. The nature and frequency of chimeras in eukaryotic metagenetic samples. *Journal of* . . . , 2012.

[18] S. Ratnasingham and P. D. N. Hebert. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Resources*, 7(3):355–364, May 2007.

[19] M. Schirmer, U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan, and C. Quince. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*,

43(6):gku1341–e37, Jan. 2015.

[20] J. Tu, J. Guo, J. Li, S. Gao, B. Yao, and Z. Lu. Systematic Characteristic Exploration of the Chimeras Generated in Multiple Displacement Amplification through Next Generation Sequencing Data Reanalysis. *PloS one*, 10(10):e0139857, Oct. 2015.

[21] G. Wang and Y. Wang. Login. *Microbiology*, 1996.

[22] B. S. Yandell. Smoothing Methods in Statistics. *Technometrics*, 1997.