Detecting Anomalies in the Cytochrome C Oxidase I Amplicon Sequences Using Minimum Scoring Segments

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 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 challenging to curate and validate user-contributed data manually. This, in turn, propagates prediction errors, leading to future erroneous taxonomic assignments. 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 multiple sequence alignments to model insertions, deletions, and substitutions and flag sequences with incongruous fit to the model. Our analysis of a complete set of curated Insecta COI reference sequences identifies the presence of numerous anomalous sequences, which makes a strong case for the validity of our approach and 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

1. INTRODUCTION

Accurate species identification is critical for estimating biodiversity and for better understanding the effects of various anthropogenic and natural disturbances on the underlying biological landscape. For taxa that are well studied, visual cataloging and identification of morphological features by an expert taxonomist is the gold standard for species identification [2]. However, expert-led taxonomic identification is time-consuming and requires high sample integrity [10] as well as intimate knowledge of phenotypic differences, including those resulting from changes in life stages or gender. 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].

High-throughput sequencing projects rely on similarity and divergence patterns within and across taxonomic levels to estimate the number and the nature of species in a sample [18]. For these estimations to be accurate, the assembly of a comprehensive reference database is critical. 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 commonly submitted to public sequence databases, such as the NCBI's GenBank. While the quality of the sequences deposited to public databases is traditionally screened for trivial discrepancies – such as the occurrence of a stop codon in the open reading frame (ORF) — ultimate responsibility for quality and taxonomic 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 of 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 [19]. As such, inaccuracies in reference databases are inevitable.

In preliminary simulations on a subset of 1,000 sequences downloaded from the BOLD database (data not included), random insertions or deletions (indels) did not produce 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; or hybrid sequences from multiple parents, are a significant 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 unbound sequences [20]. Popular bioinformatics tools such as uchime [4] or ChimaraSlayer [6] are often used for detecting COI chimeras. These tools can achieve great sensitivity when the chimera fragments originate from evolutionarily distant sequences. However, due to their computationally intensive nature, chimera detection tools are often omitted from pipelines and replaced by cross-sample abundance checks, under the assumption that sequences that only occur in low-abundance in a single sample are erroneous. This approach perhaps 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 public databases. By exogenous contamination, we refer to kingdom-level COI sequences mismatches, such as the recovery of bacterial sequence 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 sequences not available in the contamination database.

Given the crucial role that specialized reference databases play in diversity estimation, rigorous inspection and detection of erroneous sequences they may contain 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 which leverages the functional constraints on a COI amino acid sequence to identify potentially anomalous sequences that deviate significantly from the profile(s) compiled at a given taxonomic level. We hypothesize 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 dereplication, 2- score profile inference and modeling and 3- identification of sequence anomalies. During the first step, we construct a taxonomic group's multiple sequence alignment from a preprocessed dataset. In the second step, we compute the similarity for each sequence in the alignment against the remaining sequences, and use these values in a non-parametric model to describe sequence variability in each column of the multiple sequence alignment. In the last step, we use the model to flag sequences that exhibit poor statistical fit against the multiple sequence alignment.

By working at the amino acid level, we can detect artifacts are not challenging to identify using the nucleotide sequences. 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 specific to different regions of the COI protein [14].

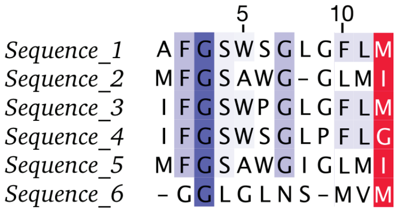
Our results show that this approach is effective for detecting insertions, deletions, chimeras, mis-translations and contaminations in Insecta sequences from the BOLD database. Furthermore, this method is computationally efficient and is a good addition to existing strategies for monitoring the integrity 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 preprocessed and analyzed as described below.

2.1 Sequence Filtering and Reference MSA Generation

Sequences with more than 3% of ambiguous amino acids (represented as X in the sequences) 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 divided by order and aligned using MAFFT's fast approximate alignment mode [11, 12]. Each MAFFT alignment was subsequently parsed to identify *seed* sequences and *partial replicates*. The seeds are references against which other sequences aligned with no insertions and with at least 93% similarity, whereas the partial replicates are the sequences that align against the seeds with no insertions and with at least 93% similarity. The seeds were subsequently aligned using MAFFT's sensitive mode, yielding what we refer to in what follows as the *reference MSA*.



```
                  5          10
Sequence_1   A F G S W S G L G F L M
Sequence_2   M F G S A W G – G L M I
Sequence_3   I F G S W P G L G F L M
Sequence_4   I F G S W S G L P F L G
Sequence_5   M F G S A W G I G L M I
Sequence_6   – G G L G L N S – M V M
```

Figure 1: Sample multiple sequence alignment with $n = 6$ and $l = 12$

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 (i_j), where $0 < i < n$ and $0 < j < l$ as:

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

where δ is a function for scoring a pair of amino acids based 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, in Figure 1, the score S_{ij} for *sequence_6* at column 12 (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 simialrity biases such as when the amino acid at position j in sequence i is rare or comes from an undersampled species in the dataset, we smooth, $S_{i,j}$ using a moving average approach with a window of size m [22]. Thus, for a sequence i, the smoothed score SS_{ik} at position j is computed as:

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

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).

2.3 Score Profiles and Sequence Anomalies

At each column k of the reference MSA, we model the distribution of score profiles $SS_{*,k}$ ($SS_{1,k}$, $SS_{2,k}$, ...) using a univariate *Kernel Density Estimation* (KDE) with a Gaussian Kernel [16]. KDE is a non-parametric method for estimating 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.

We chose to model $SS_{*,k}$ using a non-parametric KDE since this approach does not make any assumptions about the structure of underlying distributions of scores at each column. This makes the method particularly resilient to multimodal distributions of scores which could arise due to taxonomic subclustering of the sequences. This consideration is essential when working at higher taxonomic levels, particularly in old lineages, where sequence diversification is potentially non-negligible and where sequences are most likely to cluster tightly according to the underlying taxonomic levels. For example, due to the considerable phylogenetic distance between the Mecoptera and Diplura orders of the Insecta class is likely to translate for some columns as a multimodal distribution of the smoothed similarity scores.

Using a column's KDE, we can quickly estimate the probability of observing an SS value at a particular column. Figure 2 (Bottom) gives an example of two KDEs fitted using the scores observed at columns $k = 330$ and $k = 350$ respectively. The distribution at column 330 shows less variance than that at 350, which reflects the distribution of scores seen over both columns. The probability density of the red curve quantifies our confidence in the quality of the alignment of the red sequences over that at 350.

Given that we model the scores independently at each column, the resulting KDEs capture the differences in amino-acid diversification rates across functionally distinct sites in the COI protein.

For each sequence, we were interested in identifying subsequences exhibiting poor fit against the reference MSA. This amounts to finding SS values with low probability, as computed by KDE distributions of profile scores. The occurrence of a single low-probability position is not *per se* indicative of an anomalous mutation. In an MSA with a large number of sequences, a mutation in a highly conserved amino-acid, such as in Tryptophan (W) is assinged a low probability density. However, the occurrence of multiple, consecutive low-probability density positions is unlikely to occur in a normal sequence. To identify the longest stretches of low-probability density windows in each sequence, we used Kadane's algorithm, as described in [1]. Given a sequence of positive and negative values, Kadane's algorithm finds the continuous subsequence that maximizes the sum of its elements, i.e., no other subsequence can add up to a larger value. Given that Kadane's algorithm requires positive and negative values as input, we mapped our probability densities into scores using the following equations:

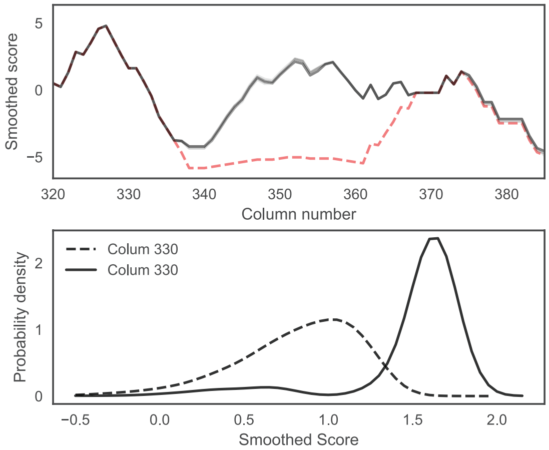


Figure 2: Top: The black curve represents the profiles of smoothed scores for 20 sequences from the BOLD COI dataset. The red curve represents a profile which substantially deviates from the cloud of black curves over a substring of length 22 amino acids. Bottom: KDE distributions derived from windows of size m=11 and centered around columns $k = 330$ (solid line) and $k = 350$ (dashed line)

$$\sigma(l, e, k) = k(\sqrt{x} - \sqrt{e}) \tag{1}$$

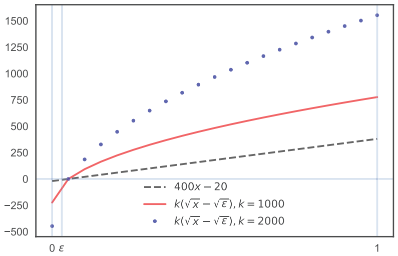


Figure 3: Increasingly large probability densities above threshold e (0.03 in this example) are assigned increasingly large, positive scores whereas increasingly small probability densities below e are assigned increasingly small, negative scores. The function is shown here over the 0-1 interval.

Equation 1 allows us to convert the probability densities, p, such that values of p where $(p \leq \epsilon)$ are converted into

Table 1: Breakdown of the number of sequences per order at each stage of the analysis.

Order	Raw	Cleaned	Unique	Seeds
Archaeognatha	216	204	86	6
Blattodea	2,024	1,853	431	33
Coleoptera	72,731	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
Isoptera	2,218	2,138	259	19
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	515,133	466,227	177,833	5,729

decreasing negative values, while the probability densities describing a good-fit ($p > \epsilon$) are converted into increasing positive values (Figure 3). Here we choose ϵ as 0.03. For example, the probability densities 0.001, 0.05, 0.4 and 0.9 are converted using Equation 1 into the values -141, 50, 459, 775. We refer to the continuous subsequence that maximizes the sum of σ-converted probability densities as *Minimum Scoring Segment* (MSS). A MSS value represents the size of the stretch where the fit between a sequence and a reference MSA is minimal according to the distributions of SS scores, i.e., the alignment is of poor quality.

3. RESULTS AND DISCUSSION

The dataset downloaded from the BOLD database consisted of 515,133 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 among COI sequences at the amino-acid level, the clustering- and alignment-based dereplication steps allowed us to reduce the dataset by over 1,000-fold, and resulted in 5,729 seed sequences (See Table 1). The seeds were aligned into a reference MSA using MAFFT's sensitive mode. The reference MSA consisted of 1057 columns and was, overall, of poor quality; 79% of the columns in the reference MSA contained gaps in 5,000 or more sequences. The number of gaps varied abruptly throughout the alignment and was notably higher at both ends of the MSA, suggesting that gaps were introduced in bulk. Through visual inspection of the alignment, we observed that the majority of the gaps were indeed inserted due to a few sequences which contained rare or unique regions.

The resulting reference alignment would most likely be different if it included the partial replicate sequences. However, we conjecture that since the partial replicates do not contain insertions, their addition would have introduced minor, confined substitutions, which would not impact our approach, particularly since we average the score profiles across a sliding window.

The SS scores and their probability densities were computed for all sequences using a window size $m = 11$. This step took less than two minutes to complete on a laptop with 16GB of RAM and 2.8 GHz CPU.

Under the naïve assumption that the observed probability densities at consecutive positions of a sequence are independent, various approaches can be used to decide whether sequence i is anomalous. For example, one can set either a minimum threshold on the lowest allowed probability density in a sequence, or a maximum allowable number of consecutive low-probability densities above which we deem a sequence anomalous. Our approach consists of identifying in each sequence the longest stretch with the lowest probability densities – this is analogous to finding the longest substring with the weakest fit against the reference MSA. To achieve this, we convert the probability density in each position into a positive or negative score in a manner that is proportional to its probability densities, i.e., increasingly large probability densities are assigned increasingly large, positive scores and increasingly small probability densities are assigned increasingly large, negative scores. We then use the scores as input to Kadane's algorithm.

In essence, any monotonically increasing function that assigns increasingly positive scores to increasingly probably events and increasingly small negative values to increasingly unlikely events is a potential candidate for converting probability densities to scores. A simple and intuitive example is a linear function with a positive slope which crosses the $x - axis$ at a given threshold ϵ (ex. $y = 400x - 20$ crosses x at $\epsilon = 0.05$). However, a linear function cannot increase sufficiently fast to assign a large range of values to the low-confidence interval [0,ϵ] without assigning extreme values to the high-confidence interval where $x > \epsilon$. By contrast, the function described in Equation 1 can allow sufficient control over the range of values assigned to the low- and high-confidence intervals.

We used the converted probability densities to compute the MSS values for each sequence in reference MSA. A total of 1,622 sequences had an MSS value of at least 1. The largest MSS value was 196 and the smallest MMS value was 0. An MSS value of 0 indicates an overall good quality fit to the reference MSA and, overall, no poorly aligned regions. On the other hand, an MSS value of 196 indicates a low-quality alignment against the reference MSA over a window of at least 196 amino acids. Except for trivial cases, confirming whether a sequence is anomalous requires manual review and supporting evidence. As such, the MSS value should be viewed as an indicator of the length of the misaligned region, rather than an absolute predictor of whether a sequence is anomalous. For our reference MSA, gap-free substrings are, on average, 11 amino acids long. As such, we use an MSS value of 25, which arises most frequently in out data from the misalignment of a subsequence of length 19-13 amino acids, using a window of size $m = 11$.

3.1 Investigation of Anomalous Sequences

A total of 170 seeds contained MSS values greater than or equal to 25. These sequences served as seeds for 2,062 partial replicates, with four seeds (GBLN4189-14, GBMIN13163-13, GBGL1368-06, ASHMT277-11) accounting for 1,687 sequences (See supplementary files). For sequences with MSS values below 25, the mean MSS was 3.69 with a standard deviation of 5.

We randomly selected and manually inspected a subset of 70 sequences with an MSS higher than 25. A sequence was confirmed as anomalous if we could classify it unequivocally into one of following four error classes: 1- mistranslated amino acid sequence, 2- containing an insertion or a deletion (indel), 3- chimera, or 4- misclassification. We considered a sequence suspicious if its MSS aligned with low-quality to the reference MSA, but its assignment to any of the four error classes was not evident.

3.1.1 Mistranslated Sequences

The MSS values and the score profile curves are intuitive representations of anomalies. For instance, mistranslated sequences are trivial to identify visually using score profile curves; their score profile is consistently low, and their MSS value is close to their sequence length. These sequences are also trivial to validate, as they have an alternative, full ORF which produces a better overall alignment and substantially reduces the MSS value. One such sequence, CNKOC293-14, has a consistently negative profile score using the submitted translation (See Figure 4). When using the first ORF on the reverse strand, the same sequence yields a consistently positive score.

3.1.2 Indel Containing Sequences

The score profile curve of an indel-containing sequence is different. The DNA sequences pre- and post-indel either produce well or poorly aligned regions. The subsequence that aligns well has a higher profile score, which leads to a notable change in the score profile curve either from high to low or the inverse. In sequences with indels, the MSS length is a function of the location of the indel in the sequence. Indels represent the most abundant anomalies in our dataset. Sequence LEPIN057-14 is a good example as it shows an evident change in the sequence's score profile near $k = 220$ (See Figure 4). Translation of this sequence into all reading frames shows that the second ORF produces a better fit for the first 220 amino acids of the reference MSA, whereas the first ORF results in a better fit for amino acids located after position 220 in the reference MSA. An early occurrence of an indel can yield a predominantly non-homologous amino-acid sequence with a large MSS and results in a score profile that is most similar to that of a mistranslated sequence.

3.1.3 Out of Class Contamination

Pipelines used for analyzing COI data, such as the one proposed in [18] and utilized in BOLD, suggest screening the COI sequences against a custom-compiled contamination database. This approach does not work when the sources of contamination are not cataloged. 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, was submitted to the BOLD database as originating from a Hymenoptera but showed a score profile curve that was consistently lower than the median score profile. This sequence of 217 amino acids has an MSS value of 196, suggesting a poor alignment over the complete sequence (See Figure 4). A Blast search of this sequence against the NR database (using the NCBI's web BLASTP with default parameters) showed that most returned hits were of proteobacterial origin. More specifically, 49 of the top 50 BLAST hits were against *Legionellaceae* bacteria, with the best hit aligning with 98% similarity over the complete length of the sequence. Further investigation showed that Legionella can reside within insect hemocytes and that some insects are commonly used as models for studying the immune response to Legionella infections [7, 8]. Under the light of such evidence, we hypothesized that this sequence is in fact from a *Legionellaceae* pathogen, rather than from the Hymenoptera host. Other sequences were also predicted using BLASTP to be of bacterial origin. For instance, the top 50 best hits of sequence CNEIE680-12 (MSS= 59) were against bacterial sequences in the *Oxalobacteraceae* family. For sequence SSWLE10476-13 (MSS = 47) which also shows a consistently low score profile that matched closely with that of MANT163-13, the three most significant Blast hits are to arthropods, two annotated at the class level (Insecta) and one annotated at the family level (Staphylinidae). The fourth hit, however, 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 likely artifacts resulting from the propagation of erroneous annotation. These examples of contamination highlight the limitations in using a reference database for contaminant screening.

Indel containing sequences or chimeras are challenging to detect using alignment at the nucleotide level in the absence of references that are similar to it. However, the previous examples clearly illustrate that a switch in a score profile can be a reasonable tell-tale of a sequence anomaly, regardless of whether or not similar references are present in the database.

3.2 Identifying Anomalies in Out of Sample Data

We evaluated the performance of our method on out of sample sequences for which the status (anomalous vs. non-anomalous) is known. First, we generated two datasets from the sequences which we excluded from the reference MSA.

The first dataset consisted of 100 randomly picked partial replicate sequence with a seed containing an MSS value less than 15, i.e., those are random sequences we believe are not anomalous. The second dataset consists of 50 randomly selected partial replicate sequences which aligned with a seed among the 70 we manually confirmed as anomalous. Sequences from both datasets were incorporated into the reference MSA using transitive alignments. Specifically, given a pairwise alignment of the partial replicate a and its seed b, and given a reference MSA describing the alignment of the seeds b and c, we can transitively align a and c by introducing gaps in a so that its alignment in the reference MSA matches its first alignment on b except where both a and b have gaps (See Figure 5 for an example). This transitive operation is unambiguous given that a's alignment against

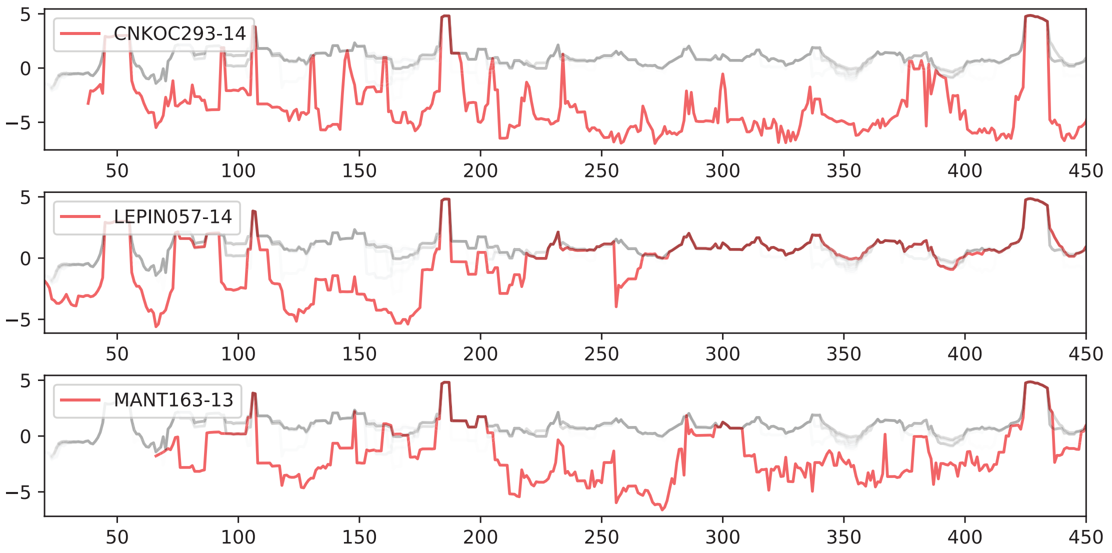


Figure 4: Score profile curves for three anomalous sequences (shown in red) plotted against typical score profile curves. The top and the bottom plots show entirely misaligned sequences representing an erroneously translated sequence and a misclassified sequence respectively. The middle plot shows a sequence with an upshift in the score profile occurring at the position of an insertion.

b does not include any gaps.

```
    a: MK-AP--M-
1   b: MGDAPMKMA

    b: MGD--APMKMA--
2   c: -VDAVAPMK-ARV

    a: MK---AP--M---
3   b: MGD--APMKMA--
    c: -VDAVAPMK-ARV
```

Figure 5: The transitive alignment of partial replicate, a, against the seed c. 1) The Initial alignment used to dereplicate a and keep b as a seed. 2) The alignment of seeds b and c in the reference MSA. 3) In addition to the original gaps introduced in a during the initial alignment (1), gaps from b's alignment with c are also transitively added to a. The transitive gaps are shown in red

For the first dataset, out the 100 sequences inspected, a single sequence (ASWAX567-08) had a MSS value above 15 (MSS value of 33). This suggested that this partial replicate was anomalous despite the fact that its seed (BBHEC816-10) had a MSS of 0. Upon manual inspection, we identified an insertion of an adenosine base at position 628, the removal of which produced a MSS of 0. This confirmed that sequenceASWAX567-08 was indeed anomalous.

For the second dataset, only 26 of the 50 sequences randomly generated from anomalous seeds had MSS values above 25. All the partial replicates with MSS values greater than 25 had the same irregular patterns observed in their seeds. However, the 21 remaining sequences had MSS values between 1 and 23, with an average MSS value of 5.1. After manually inspecting these sequences, we observed that these MSS-scoring sequences were shorter than their seeds and, in most cases, did not or did only partially cover the regions in their seeds where we observed the irregular patterns.

The outcomes form both out of sample datasets above show that partial replicates can have a different classification from that of their assigned seed. Therefore, one cannot screen dereplicated datasets and propagate the status of seeds onto the partial replicates they substitute. This applies whether or not we can confirm the seeds as anomalous. However, although our analyses did not extend to partial replicates – beyond the out of sample tests above – the computational tractability of our method facilitates the subsequent incorporation of partial replicates into the reference MSA and provides an efficient approach to screen additional sequences at a linear computation cost.

The seed sequences, the reference MSA a detailed analysis of the 70 sequence anomalies and all supplementary files can be found at the following url: `https://figshare.com/s/99bc4f716da30bba9327`.

3.3 Impact of the Window Size on the Identification of Anomaly Boundaries

The window size, m, plays a critical role in identifying the boundaries of an artifact causing a region of low-quality alignment in a sequence – particularly in the presence of large, conserved gaps. In essence, smaller values of m are ideal for identifying short regions that do not align well with the reference MSA, whereas larger values of m can smooth out short anomalies. For example, using $m = 2$, the m-smoothed scores for $S = [5.5, 1.0, -2.1, 5.9, 5.5, -2.5]$

are $SS = [nan, 3.2, -0.5, 1.9, 5.7, 1.4]$, whereas the smoothed scores with $m = 3$ are $SS = [nan, nan, 2.5, 2.5, 1.6, 2.2]$ (nan are assigned to position where the SS score cannot be computed to window constraints). As the windows size grows, the influence of distant, highly similar positions – including the gaps-rich columns introduced to compensate for rare artifacts – soften the contribution of misaligned columns and either shrink resulting MSS values or lead to completely missing the poorly aligned region. However, the influence of distant columns is desirable for merging windows of low conservation, when those are interspersed with gap rich column. Figure 6 illustrates this issue for a subsequence of the mistranslated seed (CNKOC293-14). Although this sequence has little similarity with the remaining seeds, the gap-rich columns of the reference MSA cause a significant upshift in the score profile curve. With a short window (ex. $m = 5$), the two island of spurious similarity can interrupt the MMS extension, and therefore underestimate the extent of the anomaly.

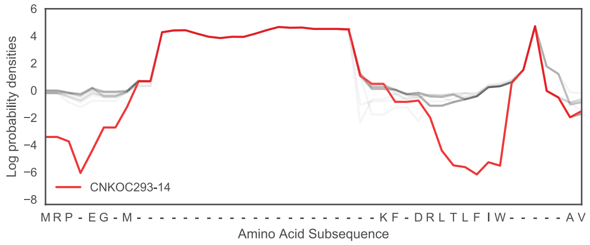


Figure 6: Alignment of mistranslated sequence along gaps-rich columns. Despite low similarity with the remaining seeds, the sequence shows score profile over the gap-rich columns

In the absence of a consistent window size for detecting anomalies, a reasonable strategy consists of employing a large window size to weed out conspicuous anomalies and to decrease the window size gradually to capture shorter anomalies. For instance, by using a window k=15, we were able to identify 45 sequences with MSS values higher than 100. The removal of these sequences leads to more compact alignments (number of columns only 67% columns in the reference MSA contained gaps in 5,000 or more sequences compared to 79% before). Furthermore, removing 45 sequences identified with $m = 15$ leads to an increase in the observed MSS when reanalyzing the data using $m = 11$

4. CONCLUSION

The COI marker has proved effective for metabarcoding projects and its use has made significant contributions in the field of ecology. Thus far, the process of cataloging new species in reference libraries has mostly relied on the use of targeted sequencing. This low-throughput activity has 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 cataloging diversity – notably for sequencing environmental DNA or bulk biodiversity samples –, 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, both due to its computational tractability and it intuitive interpretability, and make a strong case for its use as a complement to existing tools to identify anomalous sequences in COI reference databases.

5. ACKNOWLEDGMENTS

We would like to thank Drs. Emma Ransome and Christopher Meyer for their helpful clarifications on sequencing biases in COI amplicon data. This work was supported by NSF-OCE grant number 1260169 and NIH-NIGMS grant number 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.

ABOUT THE AUTHORS:

Mahdi Belcaid received a Ph.D in Computer Science from the University of Hawaii's Information and Computer Sciences Department in 2012. He is currently an assistant research professor at the Hawaii Institute of Marine of Marine Biology. His current research focuses on the application of computational techniques for the analysis of large marine biology datasets.

Guylaine Poisson received a B.S. degree and a M.S. degree in Biological Sciences from the Univeristé de Montréal, Québec, Canada in 1994 and 1997, and a Ph.D. degree in Cognitive Computer Science from the Université du Québec à Montréal, Québec, Canada in 2005. She is currently an associate professor in the Department of Information and Computer Sciences at the University of Hawai`i at Mānoa. Her research interests include bioinformatics and computational biology.