Presentation and app available at:

https://atorresgalvis.shinyapps.io/MolecularDiagnoses/

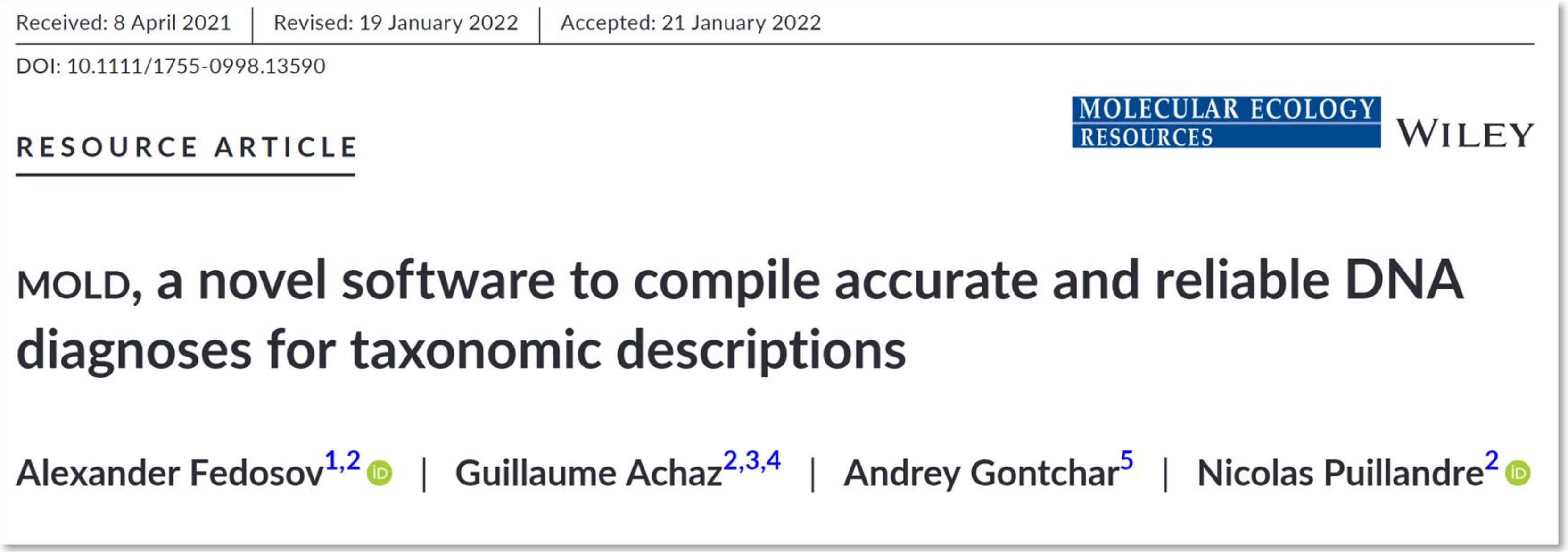
A new method for using molecular diagnoses in Taxonomy

Torres, Ambrosio[1*]; Srivathsan, Amrita[1]; Meier, Rudolf[1,2]
[1]Center for Integrative Biodiversity Discovery (Museum für Naturkunde - Leibniz Institute for Evolution and Biodiversity Science), Berlin, Germany. [2]Department of Biological Sciences, National University of Singapore, Singapore, Singapore. *atorresgalvis@gmail.com

Taxonomists generally use techniques to diagnose taxa and identify species that rely on morphological or anatomical comparisons, especially by trying to propose primary homology hypotheses. More recently, as part of the development of the so-called "integrative taxonomy", more sources of information have begun to be included for taxonomic purposes (e.g., DNA sequencing, chromosome mapping, biochemical profiling, and phylogeography). However, it is uncommon the inclusion of Molecular Diagnostic Combinations (MDCs) in published descriptions of new species, as well as its utilization for identifying unknown specimens. In this study, we propose a method to overcome the main technical and algorithmic issues associated with obtaining reliable MDCs, especially those arising from using large datasets. This method assigns a weight to each site based on the Jaccard index (i.e., the number of sequences in which the site states differ from the site state in the query taxon); then, it uses a weighted random sampling approach to build the candidate combinations to become MDCs. At the same time, the method uses a stability measure of the specificity of the candidate combinations to assess their reliability. This measure relies on the suboptimal match values of each one of the combinations with the sequences that do not belong to the query taxon (the user can select the maximum suboptimal match accepted). As a final step, with the most frequent sites in the combinations that meet the preceding criterion, the final MDC is built (which must also meet the same criterion). We implemented the previously described method using the R programming language (script version and Shiny app version available at https://atorresgalvis.shinyapps.io/MolecularDiagnoses/). In addition, we included a module for identifying unknown specimens by evaluating the matching patterns between a pool of available MDCs and each (aligned or unaligned) sequence. Using different empirical datasets, we demonstrate that the proposed method is efficient for obtaining reliable MDCs, which in turn are helpful in identifying unknown specimens. Finally, we briefly discuss some theoretical and philosophical concerns raised by some traditional taxonomists regarding the use of composite molecular characters for diagnoses and taxonomic descriptions.

Molecular Diagnoses: composite diagnostic molecular characters (i.e. *"....a combination of nucleotides at selected sites that are shared by all members of the query taxon and by no member of the reference taxa."*, as stated by Fedosov et al. 2022):

- ~~Theoretical, philosophical and biological concerns/issues.~~
- Methodological issues and computational implementation. 👀



Received: 8 April 2021 | Revised: 19 January 2022 | Accepted: 21 January 2022

DOI: 10.1111/1755-0998.13590

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

MOLD, a novel software to compile accurate and reliable DNA diagnoses for taxonomic descriptions

Alexander Fedosov[1,2] | Guillaume Achaz[2,3,4] | Andrey Gontchar[5] | Nicolas Puillandre[2]

Other tools for molecular diagnoses:

- R package spider (Brown et al., 2012)
- Fastachar (Merckelbach & Borges, 2020)
- R package quiddich (Kühn & Haase, 2020)
- DeSignate (Hütter et al., 2020)
- Cladistic Haplotype Analysis (CHA—Brower, 1999)

Main problems regarding `MOLD` aproach:

1. The hunting of minimal diagnostic nucleotide combinations or mDNCs. From Fedosov et al., 2022:

> *“An mDNC-based diagnosis can be invalidated either by a low-frequency polymorphism in the query species or by a convergent emergence of the same nucleotide combination in any of the reference taxa.”*

> *“We thus explore combinations of DNA characters that contain more than the minimal number of nucleotide sites necessary to assign a sequence to a query taxon. We term such combination redundant DNC, or rDNC. Because rDNCs are longer than mDNCs, the probability of finding the same nucleotide combination among the reference taxa due to convergence is lower[...] Therefore, incomplete match of an rDNC is acceptable. In this context, we developed the `MOLD` algorithm that compiles rDNCs.”*

As there are only four character-states for DNA data (A, C, G, and T) this problem is even more dramatic.

Yes, but…the actual issue is not the attempt of the authors to prevent this, the real issue was the method implemented in `MOLD` to ‘overcome’ this problem:

> *“To test an rDNC after each elongation step, `MOLD` repeatedly creates simulated test data sets that are generated by introducing artificial mutations into the original DNA sequences. This procedure aims to evaluate whether hypothetical larger data sets with sequences that were not sampled in the original data set would still validate a candidate rDNC. It evaluates which are the more relevant rDNCs, despite the limited number of sampled specimens/sequences.*
> *Each artificial sequence is generated by introducing p nucleotide substitutions into an existing sequence, where p is a random natural number drawn from a uniform distribution `[1, k*L/100]`. […] For each rDNC evaluation step, `MOLD` generates 100 test data sets.”*

- Shorter combinations will be probabilistically less vulnerable to the artificial mutations, because it is more possible that artificial mutations reach combinations with more characters.

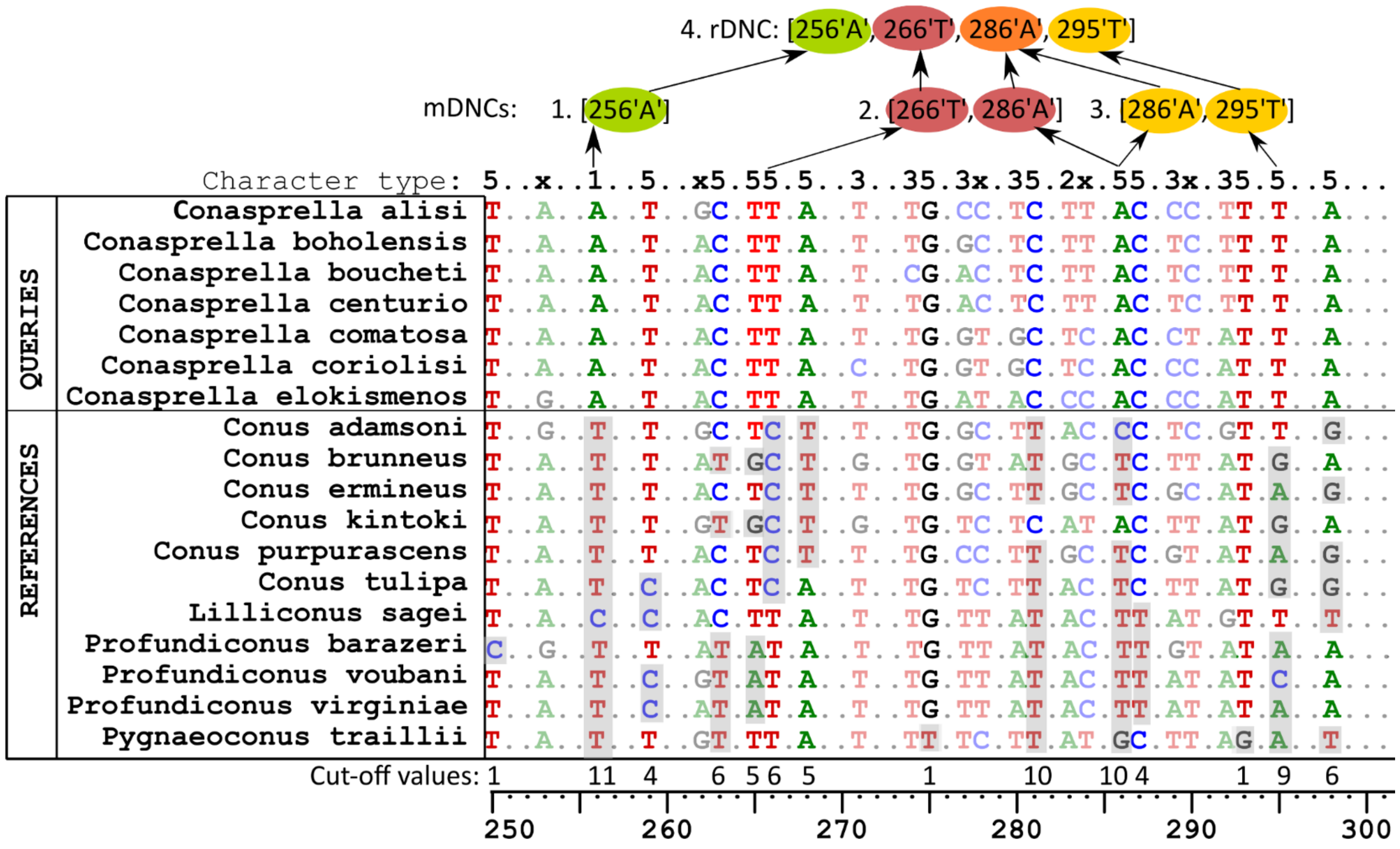
What is the probability to catch Buzz Lightyear if he is surrounded by tons of Martians, especially if you cannot see inside the arcade claw machine (as it would be the case in the MOLD simulations)?

- Moreover, why do they assume/model the mutations based on a uniform distribution?
- This procedure is very time consuming, especially for huge datasets. And at the end, it did not use the observed data to test the combinations.

> *"We show that the shorter the mDNC, the more reliable it is. Also, MOLD core functions are designed in such a way that the shorter the mDNC, the higher the probability that it will be identified."*

In addition, using MOLD we found combinations of two nucleotides for relatively small datasets (~1000 sequences X 300 bp). Obviously, these combinations were not useful for diagnoses, nor for identifying specimens.

2. Character choosing based on Jaccard index (“cut-off values” in Fedosov et al., 2022).



```
         Character type: 5..x..1..5..x5.55.5..3..35.3x.35.2x.55.3x.35.5..5...
       Conasprella alisi T..A..A..T..GC.TT.A..T..TG.CC.TC.TT.AC.CC.TT.T..A...
  Conasprella boholensis T..A..A..T..AC.TT.A..T..TG.GC.TC.TT.AC.TC.TT.T..A...
    Conasprella boucheti T..A..A..T..AC.TT.A..T..CG.AC.TC.TT.AC.TC.TT.T..A...
    Conasprella centurio T..A..A..T..AC.TT.A..T..TG.AC.TC.TT.AC.TC.TT.T..A...
    Conasprella comatosa T..A..A..T..AC.TT.A..T..TG.GT.GC.TC.AC.CT.AT.T..A...
   Conasprella coriolisi T..A..A..T..AC.TT.A..C..TG.GT.GC.TC.AC.CC.AT.T..A...
 Conasprella elokismenos T..G..A..T..AC.TT.A..T..TG.AT.AC.CC.AC.CC.AT.T..A...
           Conus adamsoni T..G..T..T..GC.TC.T..T..TG.GC.TT.AC.CC.TC.GT.T..G...
          Conus brunneus T..A..T..T..AT.GC.T..G..TG.GT.AT.GC.TC.TT.AT.G..A...
          Conus ermineus T..A..T..T..AC.TC.T..T..TG.GC.TT.GC.TC.GC.AT.A..G...
           Conus kintoki T..A..T..T..GT.GC.T..G..TG.TC.TC.AT.AC.TT.AT.G..A...
      Conus purpurascens T..A..T..T..AC.TC.T..T..TG.CC.TT.GC.TC.GT.AT.A..G...
            Conus tulipa T..A..T..C..AC.TC.A..T..TG.TC.TT.AC.TC.TT.AT.G..G...
        Lilliconus sagei T..A..C..C..AC.TT.A..T..TG.TT.AT.AC.TT.AT.GT.T..T...
  Profundiconus barazeri C..G..T..T..AT.AT.A..T..TG.TT.AT.AC.TT.GT.AT.A..A...
   Profundiconus voubani T..A..T..C..GT.AT.A..T..TG.TT.AT.AC.TT.AT.AT.C..A...
 Profundiconus virginiae T..A..T..C..AT.AT.A..T..TG.TT.AT.AC.TT.AT.AT.A..A...
    Pygnaeoconus traillii T..A..T..T..GT.TT.A..T..TT.TC.TT.AT.GC.TT.AG.A..T...
         Cut-off values: 1     11 4   6 56 5      1    10   104     1 9  6
                         250       260       270       280       290       300
```

“[… some] sites are a priori poor candidates to construct short mDNCs, as they will need to be combined with many others to assemble an mDNC […]. Therefore, each site is assigned a score that corresponds to the number of reference taxa members that differ from the query taxon for the nucleotide at this site (Figure 1, numbers below the alignment). […] The scores are then ranked in descending order and the user defines how many of the top-ranking sites are used for assembling a draft combination [which will be subsequently tested].” (Fedosov et al., 2022)

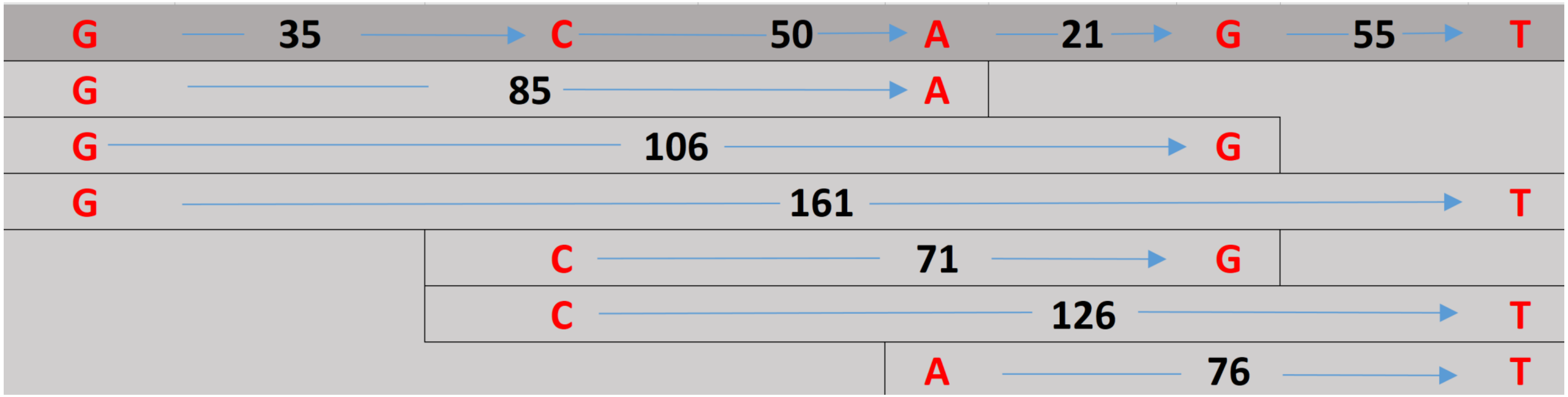
- How many characters are good enough? (default is 100)
- What happens with multiple ties (i.e. Multiple characters with the same [‘good’ and ‘bad’] index value), especially for highly conserved loci?

3. Other associated problems:
   - Lack of identification/verification tools.
   - Time-consumption (especially for the simulations step).
   - The software is very unfriendly/difficult to use. Moreover, its operation is not intuitive, and in some cases, subjective.

The `OpDMC` function identifies combinations of characters that can accurately distinguish different species.
It follows these steps:

- Input: FASTA file with molecular sequences and CSV file with the list of the query species.
- Unique sequences: extract unique sequences from the FASTA file.
- Variable sites: identify and keep the positions that differ among the sequences.
- Jaccard index: calculate the similarity between query and other species for each character.
- **Weighted sampling:** build potential Diagnostic Molecular Combinations (DMCs) for each species, based on Jaccard index scores.
- **Specificity check:** ensure DMCs are highly specific to the query species, taking into account the maximum suboptimal matching for non-query species (i.e. the minimum allowed mismatches for non-query species). This approach is similar to the Bremer support measure used in phylogenetics. E.g. two species, two potential DMCs of four nucleotides, and subopt value = 2:
  a. *Megaselia_lawrenceleei*= suboptimal match with non-query [3/4]
  b. *Megaselia_lawrenceleei*= suboptimal match with non-query [2/4]
  a and b. *Megaselia_zestinsohi*= perfect Match with query taxon [4/4]
- **Refinement:** if needed, improve selected DMCs using sub-combinations without losing the specificity.
- **Optimal DMC:** build the most optimal combination based on the occurrence frequency of the sites in the checked DMCs (the specificity of the Optimal DMC is also ensured/guaranted).
- Output: Generate a CSV file with the optimal DMC and associated information.

Combination of 5 characters as distance matrix:



G 35 C 50 A 21 G 55 T

G 85 A

G 106 G

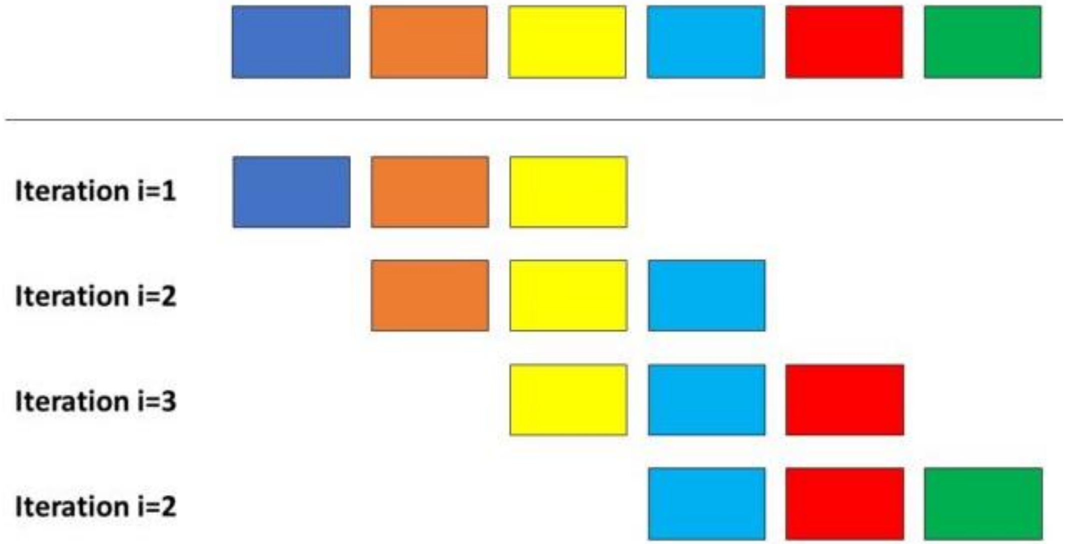
G 161 T

C 71 G

C 126 T

A 76 T

Sliding Window n=3



Iteration i=1

Iteration i=2

Iteration i=3

Iteration i=2