An interactive DNA barcode browser

Roderic D. M. Page
https://orcid.org/0000-0002-7101-9767
IBAHCM, MVLS, University of Glasgow, Glasgow, UK
Email address: Roderic.page@glasgow.ac.uk

Abstract

This paper describes an interactive web application to display DNA barcode data. It supports both query by sequence and query by geographic area. By using *n*-gram indexing of DNA sequences, and alignment-free phylogeny construction, the user can interactively explore DNA barcode data in real time.

Introduction

One of the most impressive achievements of biodiversity informatics is the Global Biodiversity Information Facility (https://www.gbif.org, GBIF), which at the time of writing has aggregated data on 1.6 billion occurrences of organisms from across the planet. Yet a time-travelling naturalist from the 19th century would find little in the GBIF interface that they weren't already familiar with. The two core visualisations GBIF provides are (a) lists of species arranged in a Linnaean hierarchy, and (b) species distributions plotted on a map. These methods of summarising biodiversity data are appropriate for much of the data in GBIF, but a growing proportion of data in GBIF comes from DNA barcoding (Ratnasingham & Hebert, 2013) and metagenomics. Displaying data from these sources in this way ignores the very thing that makes sequence data unique, namely the nucleotide sequence itself. Ideally, given DNA sequences we should be able to query GBIF and find out where similar sequences occur. Given a geographic area, we should be able to find sequences known from samples in the area. Furthermore, we should be able to explore sequences using the appropriate tools, such as phylogenetic trees. Not only are trees a useful way summarise sequence data, they also enable us to go beyond counting species and to measure phylogenetic diversity (Faith, 1992)

There are already some striking visualisations of phylogenies and geography, such as Microreact (https://microreact.org) (Argimón et al., 2016) and Nextstrain (https://nextstrain.org) (Hadfield et al., 2018). These tools typically compute all the results needed for a visualisation (e.g., the phylogenetic tree) offline, then generate a web app to display the results. They are highly interactive, but only for the predefined set of sequences. In contrast, the application described here creates trees "on the fly" depending on the sequences the user has selected (either by searching by sequence or geography). To enable this "on the fly" experience the application treats the DNA sequences as text strings indexed as *n*-grams, so that in effect we are "Googling" DNA sequences (Hajibabaei & Singer, 2009). *N*-grams are substrings of characters, where *n* is the number of characters in the substring. In the context of biological sequences, *n*-grams are more commonly known as "*k*-mers", where *k* is the length of a subsequence. Having retrieved a set of sequences we can rapidly compute a phylogenetic tree based on pairwise k-mer distances between sequences, which enables us to avoid the costly stop of aligning the sequences. These two steps (*n*-gram indexing and alignment-free phylogeny construction) enable us to interactively explore DNA barcode data in real time.

This paper describes a DNA barcode browser based on these two technologies (n-gram indexing and alignment-free phylogenetic trees). This application was entered in the GBIF 2020 Ebbe Nielsen Challenge.

Methods

Data

For the entry into the GBIF Challenge a small subset of approximately 50,000 DNA barcodes for animals was retrieved from the BOLD website and converted to Darwin Core Archive format (Wieczorek et al., 2012). Each record was then extracted from the Darwin Core Archive, converted into JSON documents and uploaded to an instance of Elasticsearch.

Indexing DNA sequences

Each DNA sequence was treated as a set of *k*-mers where each sequence is partitioned into all substrings of length *k*. For DNA sequences with four bases, the number of possible substrings of length *k* is 4^k. Searching was implemented using Elasticsearch's n-gram indexing (a *n*-gram is a *k*-mer where $n = k$). To keep the size of the index manageable a value of $n = 5$ was chosen, which results in a maximum of $4^5 = 1024$ different *n*-grams for a given sequence.

Indexing by geographic location

Each DNA barcode that had a latitude and longitude value was indexed by Elasticsearch so that we can search for sequences within a given polygon. This index was also used to display a tiled map of DNA barcodes.

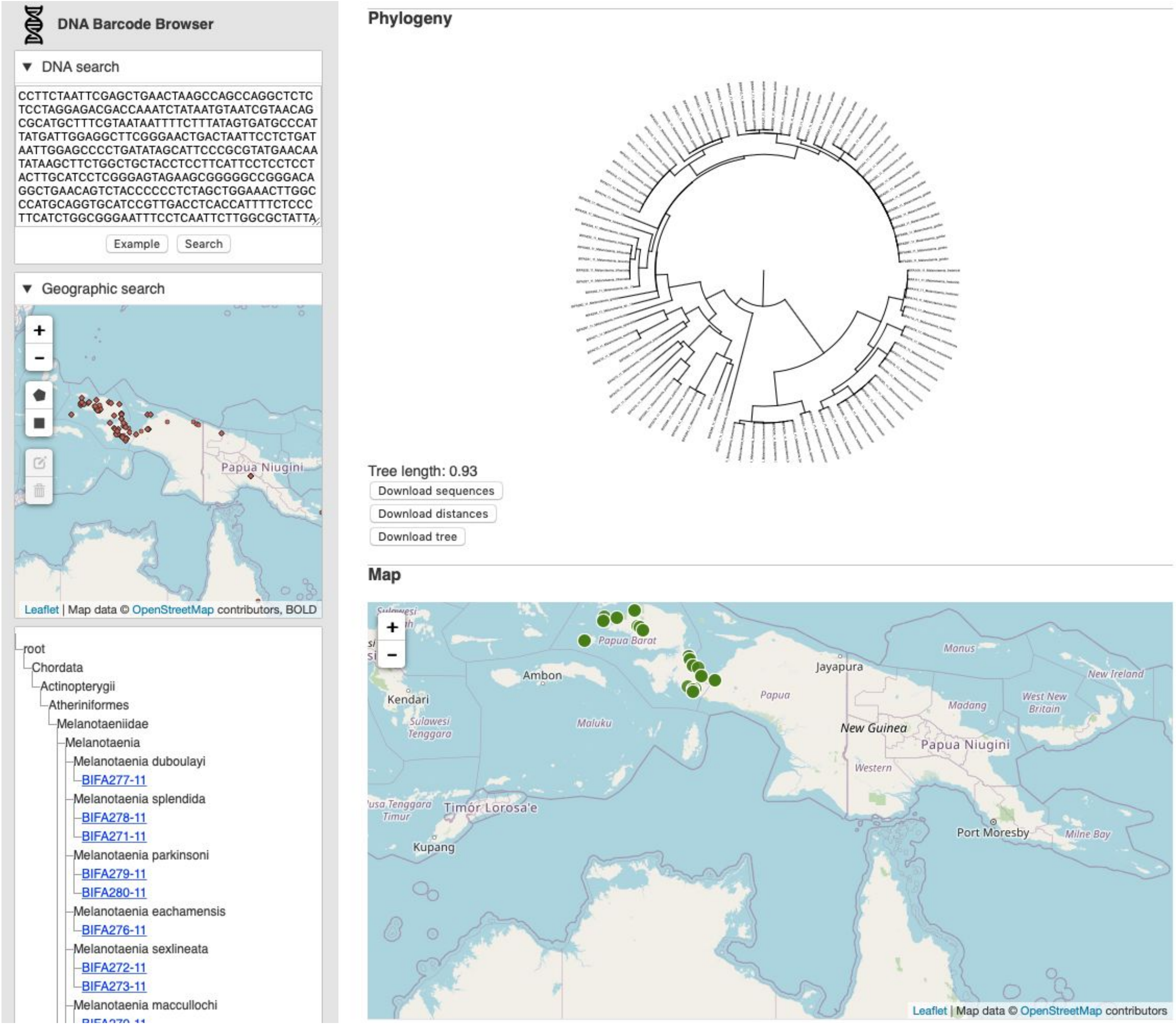
Phylogeny

Given a set of sequences (returned from either a sequence similarity or geographic search) pairwise distances between those sequences are computed based on *k*-mers (Edgar, 2004). The same value of *k* (= *n*) used to index the sequences was also used to compute pairwise distances, based on Yang and Zhang (2008) who showed that $k = 5$ gives reasonable results for phylogenetic inference. A phylogenetic tree is computed from these pairwise distances using Simonsen et al.'s (2008) "Rapid Neighbour-Joining" algorithm (code from https://github.com/biosustain/neighbor-joining).

Implementation

The DNA web browser is implemented in PHP, HTML, and Javascript, with an Elasticsearch search engine running in a Docker container hosted in the cloud. Source code is available on Github https://github.com/rdmpage/dna-barcode-browser. An instance of the application is running at https://dna-barcode-browser.herokuapp.com.

Results



```
CCTTCTAATTCGAGCTGAACTAAGCCAGCCAGGCTCTC
TCCTAGGAGACGACCAAATCTATAATGTAATCGTAACAG
CGCATGCTTTCGTAATAATTTTCTTTATAGTGATGCCCAT
TATGATTGGAGGCTTCGGGAACTGACTAATTCCTCTGAT
AATTGGAGCCCCTGATATAGCATTCCCGCGTATGAACAA
TATAAGCTTCTGGCTGCTACCTCCTTCATTCCTCCTCCT
ACTTGCATCCTCGGGAGTAGAAGCGGGGGCCGGGACA
GGCTGAACAGTCTACCCCCCTCTAGCTGGAAACTTGGC
CCATGCAGGTGCATCCGTTGACCTCACCATTTTCTCCC
TTCATCTGGCGGGAATTTCCTCAATTCTTGGCGCTATTA
```

- root
  - Chordata
    - Actinopterygii
      - Atheriniformes
        - Melanotaeniidae
          - Melanotaenia
            - Melanotaenia duboulayi
              - BIFA277-11
            - Melanotaenia splendida
              - BIFA278-11
              - BIFA271-11
            - Melanotaenia parkinsoni
              - BIFA279-11
              - BIFA280-11
            - Melanotaenia eachamensis
              - BIFA276-11
            - Melanotaenia sexlineata
              - BIFA272-11
              - BIFA273-11
            - Melanotaenia maccullochi

Figure 1. Screenshot of DNA barcode browser https://dna-barcode-browser.herokuapp.com. The user is searching for matches to an existing sequence. These matches are displayed as a phylogenetic tree, and map, and a list in the form of a taxonomic classification.

There are two ways to use the DNA web browser (Fig. 1). The first is modelled on BLAST (Altschul et al., 1990). In the "DNA search" box the user pastes a DNA sequence and clicks on "Search". The application searches the Elasticsearch index of DNA barcodes for similar sequences, if it finds any it returns them and then computes a phylogenetic tree for those sequences. The location of the sequences is also plotted on a map, and the taxonomic classification of the sequences is displayed as a list.

The second method is to search spatially. Using the "Geographic Search" map the user can draw a polygon on the map that corresponds to the area within which they want to search.

The database retrieves sequences in that area, and as before, computes and displays a tree, shows the sequences on a map, and displays a list.

The tree and map are interactive, the user can pan and zoom both the tree and the map. The total branch length of the phylogeny (in units of pairwise *k*-mer distance) is displayed as a crude measure of phylodiversity (Faith, 1992) to help give a sense of how much genetic diversity the tree represents. The user can also download the (unaligned) DNA barcodes in FASTA format, retrieve the pairwise distances in NEXUS format (Maddison, Swofford & Maddison, 1997) suitable for input into programs such as PAUP (https://paup.phylosolutions.com/) and Splitstree (http://www.splitstree.org/), and download the tree in Newick format.

Future directions

There are several directions this work could be taken. The interface could be refined and the displays of tree, map, and list linked together using "brushing" (that is, selecting a node in a tree could also highlight the corresponding location in the map and in the list). There are other visualisations that could be used, such as Klee-plots (Stoeckle & Coffran, 2013), histograms of pairwise sequence distance (Meyer & Paulay, 2005), and plotting the phylogeny on the map (Page, 2015).

For the DNA barcode browser to be the basis of more analytical approaches, we would need to introduce measures of phylogenetic diversity, and sample geographically comparable areas. For example, we could divide the globe into comparable size areas using a discrete grid (Sahr, White & Kimerling, 2003) and use the *k*-mer approach to build phylogenies for samples of sequences taken from each cell in that grid. This would enable a grid of phylodiversity values to be computed for the whole planet.

Acknowledgments

I thank GBIF for continuing to support innovation in biodiversity informatics through initiatives such as the Ebbe Nielsen Challenge.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial Genomics,* 2:e000093. DOI: 10.1099/mgen.0.000093.

Edgar RC. 2004. Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Research* 32:380–385. DOI: 10.1093/nar/gkh180.

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61:1–10. DOI: 10.1016/0006-3207(92)91201-3.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34:4121–4123. DOI: 10.1093/bioinformatics/bty407.

Hajibabaei M, Singer GA. 2009. Googling DNA sequences on the World Wide Web. *BMC Bioinformatics* 10:S4. DOI: 10.1186/1471-2105-10-S14-S4.

Maddison DR, Swofford DL, Maddison WP. 1997. Nexus: An Extensible File Format for Systematic Information. *Systematic Biology* 46:590–621. DOI: 10.1093/sysbio/46.4.590.

Meyer CP, Paulay G. 2005. DNA Barcoding: Error Rates Based on Comprehensive Sampling. *PLOS Biology* 3:e422. DOI: 10.1371/journal.pbio.0030422.

Page R. 2015. Visualising geophylogenies in web maps using GeoJSON. *PLoS currents* 7.

Ratnasingham S, Hebert PDN. 2013. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLOS ONE* 8:e66213. DOI: 10.1371/journal.pone.0066213.

Sahr K, White D, Kimerling AJ. 2003. Geodesic Discrete Global Grid Systems. *Cartography and Geographic Information Science* 30:121–134. DOI: 10.1559/152304003100011090.

Simonsen M, Mailund T, Pedersen CNS. 2008. Rapid Neighbour-Joining. In: Crandall KA, Lagergren J eds. *Algorithms in Bioinformatics*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 113–122. DOI: 10.1007/978-3-540-87361-7_10.

Stoeckle MY, Coffran C. 2013. TreeParser-Aided Klee Diagrams Display Taxonomic Clusters in DNA Barcode and Nuclear Gene Datasets. *Scientific Reports* 3:1–6. DOI: 10.1038/srep02635.

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D.

2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* 7:e29715. DOI: 10.1371/journal.pone.0029715.

Yang K, Zhang L. 2008. Performance comparison between k -tuple distance and four model-based distances in phylogenetic tree reconstruction. *Nucleic Acids Research* 36:e33–e33. DOI: 10.1093/nar/gkn075.