**Codetta: predicting the genetic code from nucleotide sequence**

Yekaterina Shulgina\(^1\) and Sean R. Eddy\(^{1,2,*}\)

\(^1\)Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA,
\(^2\)Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA,
\(*\)To whom correspondence should be addressed

**Abstract**

**Summary:** Codetta is a Python program for predicting the genetic code of an organism from a nucleotide sequence. Codetta can analyze an arbitrary nucleotide sequence and needs no sequence annotation or taxonomic placement. The most likely amino acid decoding for each of the 64 codons is inferred from alignments of profile hidden Markov models of conserved proteins to the input sequence.

**Availability and implementation:** Codetta 2.0 is implemented as a Python 3 program for MacOS and Linux, and is available from http://eddylab.org/software/codetta/codetta2.tar.gz and at http://github.com/kshulgina/codetta.

**Contact:** seaneddy@fas.harvard.edu

**Supplementary information:** Supplementary files are available at *Bioinformatics* online.

1 Introduction

The genetic code defines how mRNA codons are interpreted into an amino acid sequence. Almost all of life uses the same decoding scheme; however, dozens of clades have evolved alternative genetic codes where the meaning of one or more codons is changed (Knight *et al.*, 2001). Knowing the genetic code is necessary to correctly predict protein sequences, but common practice is to assume the standard genetic code. New alternative genetic codes continue to be reported (Swart *et al.*, 2016; Shulgina and Eddy, 2021; Borges *et al.*, 2022), especially as advances in
metagenomic sequencing expand known microbial diversity. Predicted proteins from newly sequenced organisms are added to protein sequence databases which primarily consist of in silico translations assuming a genetic code, underscoring the need for a broadly usable genetic code prediction method.

Computational methods for genetic code prediction have been developed since the early 2000’s (Telford et al., 2000; Abascal et al., 2006; Dutilh et al., 2011; Noutahi et al., 2017). The general approach is to align the input nucleotide sequence to profiles of conserved proteins, and then, for each codon, to tally the most frequently aligned amino acid. These methods have led to the discovery of new genetic codes, but are limited in usage due to phylogenetic constraint or high error rates.

Codetta is a Python program for genetic code prediction that can scale to analyze large numbers of genomes. In a recent publication, we used an early version of Codetta to screen over 250,000 bacterial and archaeal genomes and found five clades using new genetic codes (Shulgina and Eddy, 2021). Codetta improves upon previous efforts by using a probability model with a robust statistical footing to infer the amino acid for each codon and by aligning to profile hidden Markov models (HMMs) from large publicly available datasets. Here we describe Codetta 2.0, substantially revised for easier setup and usage with new options for using custom profile HMM databases and for simple parallelization on a computing cluster.

2 Usage

Codetta takes any nucleotide sequence as input (DNA or RNA; any level of assembly completion; no annotation needed) and predicts the genetic code from coding regions with recognizable homology.

Codetta performs three main steps, in order:

1. **Alignment.** A preliminary six-frame standard genetic code translation of the input sequence is aligned to profile HMMs of conserved proteins using the HMMER *hmm*scan program
A typical source of profile HMMs is the Pfam database (Mistry et al., 2020), but custom profile HMMs can be provided with the --profiles option.

2. **Processing.** Correspondence between codons and aligned profile HMM consensus columns is processed into a single file.

3. **Inference.** For each codon, the most likely amino acid is chosen based on a probability model of how profile HMM consensus columns match with codons in the nucleotide sequence (Shulgina and Eddy, 2021). Codetta considers 20 amino acid models and one model of non-specific translation. If an amino acid has probability above a threshold (default 0.9999), then the amino acid meaning is selected. Otherwise the codon is left un-inferred (‘?’ output), which results from insufficient or ambiguous information. This is also the expected outcome for stop codons, which are not explicitly predicted by Codetta. Simulations show that observing a codon 10-30 times in Pfam alignments suffices to infer the correct amino acid (Shulgina and Eddy, 2021).

These three steps are bundled in the `codetta.py` program and are also provided as separate programs.

The most computationally intensive step is the `hmmscan` alignment of profile HMMs. Analyzing a 4.6 Mb *Escherichia coli* genome with Pfam 35.0 as a profile HMM database takes about 30 minutes on a single 2.0 GHz CPU. Speed depends on the length and number of input sequences and on the size of the profile HMM database. On a SLURM computing cluster, `hmmscan` processes can be parallelized across machines with the `--parallelize_hmmscan` option.

### 3 Performance

Codetta was designed to predict the genetic code from nucleotide sequence alone for species with potentially no close taxonomic relatives. Most other tools are for a predefined clade (Abascal
Table 1: Per-codon comparison of Codetta and FACIL on (a) 506 yeast genomes (32,384 codon predictions) and (b) 82 mitochondrial genomes (5,248 codon predictions). We show totals separately for sense codons and stop codons and further split reassigned codons and AUG (frequent FACIL error). For stop codons, only amino acid predictions are considered incorrect.

et al., 2006; Mühlhausen and Kollmar, 2014) or require the input of a phylogenetic tree and annotated genes (Noutahi et al., 2017). The only existing method for arbitrary nucleotide sequences is FACIL (Dutilh et al., 2011). FACIL similarly uses Pfam alignments and for each codon simply selects the most frequently aligned amino acid, filtering out low confidence predictions at a later step.

We benchmarked Codetta and FACIL on two test sets: 506 yeast nuclear genomes with three genetic codes varying at the CUG codon (13.8 Mb average genome size) and 82 mitochondrial genomes with 18 alternative genetic codes affecting 14 codons total (27 Kb average genome size). For the mitochondrial comparison, we used a custom-built profile HMM database of mitochondrial proteins to demonstrate usage of the custom profiles feature; we also repeated the Codetta analysis with Pfam. Complete results are provided in Supplementary Files.
In the yeast comparison (Table 1a), Codetta predicted the correct amino acid for 30,357 out of 30,366 sense codons, including the reassigned CUG codons. Nine sense codons were left uninferred by Codetta, which include six cases of likely ambiguous translation of CUG and three cases of rare codon usage (Shulgina and Eddy, 2021). In contrast, FACIL selected the wrong amino acid for 53 codons (mostly AUG codons) and left 563 codons uninferred (including 109 reassigned CUG codons). FACIL explicitly predicts stop codons, while Codetta should leave stop codons uninferred. Only amino acid predictions at stop codons were considered incorrect. Both methods had similar rates of incorrect stop codon predictions. In the mitochondrial comparison (Table 1b), FACIL again had more incorrect sense codon predictions (78, 1.5%), compared to three for Codetta.

Two of the three predictions unique to Codetta were of the rare isoleucine codon AUA as methionine in green algal mitochondria. The third prediction was of the lysine codon AAG as methionine in chytrid mitochondria. Multiple sequence alignments of mitochondrial proteins show all three of the putative reassigned codons at conserved methionine positions (Supplementary Figures), suggesting these may be bona fide reassignments rather than errors, but this requires additional confirmation.

Codetta occasionally makes incorrect predictions stemming from its underlying assumptions. Known sources of error include profile HMMs aligning to non-coding sequence, often at recent pseudogenes; extensive mRNA editing relative to the input sequence; and input sequences that differ significantly in amino acid composition from the profile HMMs (Shulgina and Eddy, 2021). Candidate new genetic codes should be validated, ideally experimentally.

4 Conclusion

In Shulgina and Eddy (2021), we analyzed all available bacterial and archaeal genomes, surpassing previous screens of genetic code diversity in scale. We have now extended Codetta 2.0 to be a well-documented and user-friendly tool with options to use custom profile HMM databases and to parallelize large analyses. As sequenced microbial diversity continues to grow, Codetta will al-
low confirming the genetic code to become a straightforward step in genome annotation, enabling discovery of new genetic codes and ensuring the accuracy of protein sequence databases.

Funding

This work was supported by the National Human Genome Research Institute of the National Institutes of Health [F31-HG010984 to YS and R01-HG009116 to SRE]; and the Howard Hughes Medical Institute.

References


