Bioinformatics Advance Access published June 8, 2011

FACIL: fast and accurate genetic code inference and logo

Bas E. Dutilh[1,*], Rasa Jurgelenaite[1], Radek Szklarczyk[1], Sacha A.F.T. van Hijum[1,2], Harry R. Harhangi[3], Markus Schmid[3,4], Bart de Wild[3], Kees-Jan Françoijs[5], Hendrik G. Stunnenberg[5], Marc Strous[6,7], Mike S.M. Jetten[3], Huub J.M. Op den Camp[3] and Martijn A. Huynen[1]

[1] Centre for Molecular and Biomolecular Informatics, Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Geert Grooteplein 28, 6525 GA, Nijmegen, the Netherlands.

[2] NIZO food research BV, P.O. Box 20, 6710 BA, Ede, The Netherlands.

[3] Department of Microbiology, IWWR, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ, Nijmegen, the Netherlands.

[4] Department of Microbial Ecology, University of Vienna, Althanstrasse 14, 1090, Vienna, Austria.

[5] Department of Molecular Biology, Faculty of Sciences, Nijmegen Centre of Molecular Life Sciences, Radboud University Nijmegen, Geert Grooteplein 28, 6525 GA, Nijmegen, the Netherlands.

[6] Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359, Bremen, Germany.

[7] Center for Biotechnology, University of Bielefeld, 33594, Bielefeld, Germany.

Associate Editor: Prof. Martin Bishop

ABSTRACT

Motivation: The intensification of DNA sequencing will increasingly unveil uncharacterized species with potential alternative genetic codes. A total of 0.65% of the DNA sequences currently in Genbank encode their proteins with a variant genetic code, and these exceptions occur in many unrelated taxa. **Results:** We introduce FACIL, a fast and reliable tool to evaluate nucleic acid sequences for their genetic code that detects alternative codes even in species distantly related to known organisms. To illustrate this, we apply FACIL to a set of mitochondrial genomic contigs of *Globobulimina pseudospinescens*. This foraminifer does not have any sequenced close relative in the databases, yet we infer its alternative genetic code with high confidence values. Results are intuitively visualized in a Genetic Code Logo. **Availability and Implementation:** FACIL is available as a web-based service at http://www.cmbi.ru.nl/FACIL/ and as a stand-alone program. **Contact:** dutilh@cmbi.ru.nl.

1 INTRODUCTION

The recent increases in read lengths have established next-generation DNA sequencing as a mature technique, with the first machines capable of single molecule sequencing currently being shipped to researchers. In most studies, the researchers' interests lie beyond translation: the focus is on proteins that are encoded in the DNA, and their function. Most analyses consider the translation between DNA and protein a trivial exercise. After all the genetic code or codon table, i.e. the "dictionary" that translates codons (nucleic acid triplets) into amino acids (AAs), is largely universal and unambiguous (Koonin and Novozhilov, 2009). However, exceptions in the code of bacteria (Bove, 1993), eukaryotic nuclei (Helftenbein, 1985; Meyer, et al., 1991), organelles (Barrell, et al., 1979) and their associated viruses (Shackelton and Holmes, 2008) have been reported and, given the increasing phylogenetic breadth of sequenced taxa (Wu, et al., 2009), many more such findings may be anticipated. A quick survey of the DNA sequences currently in Genbank shows that a total of 0.65% are annotated as being alternatively translated. If a novel sequence uses a non-canonical code, open reading frames may be different than anticipated due to the reassignment of stop codons and alternative translations of coding codons. This affects both protein sequence and function prediction, so these considerations demand an easy way to assess the genetic code used on a sequenced fragment or assembled contig.

Non-canonical codes are generally identified by inspecting an alignment of the codons on the DNA against homologous protein sequences identified by e.g. BlastX. The program Gendecoder (Abascal, et al., 2006) automates this process, but it focuses on metazoan mitochondria and requires an annotated Genbank file as input. DNA sequencing increasingly yields fragments and assembled contigs that contain sufficient information for reliable genetic code prediction, but performing BlastX searches before knowing the correct translation table is untenable with the rate of DNA sequencing accelerating faster than CPU power. Moreover, the alignment may introduce errors that need to be addressed, preferably by an automated and validated approach.

2 METHODS

2.1 Training data

We used 5,866 annotated DNA sequences to construct the training data set: 3,269 bacterial, 176 archaeal and 2,421 organellar genomes (Supplementary Table 3), representing all such genomes available on July 13th 2010 in the Entrez genome database. From these genomes, we composed a training

[*] To whom correspondence should be addressed.

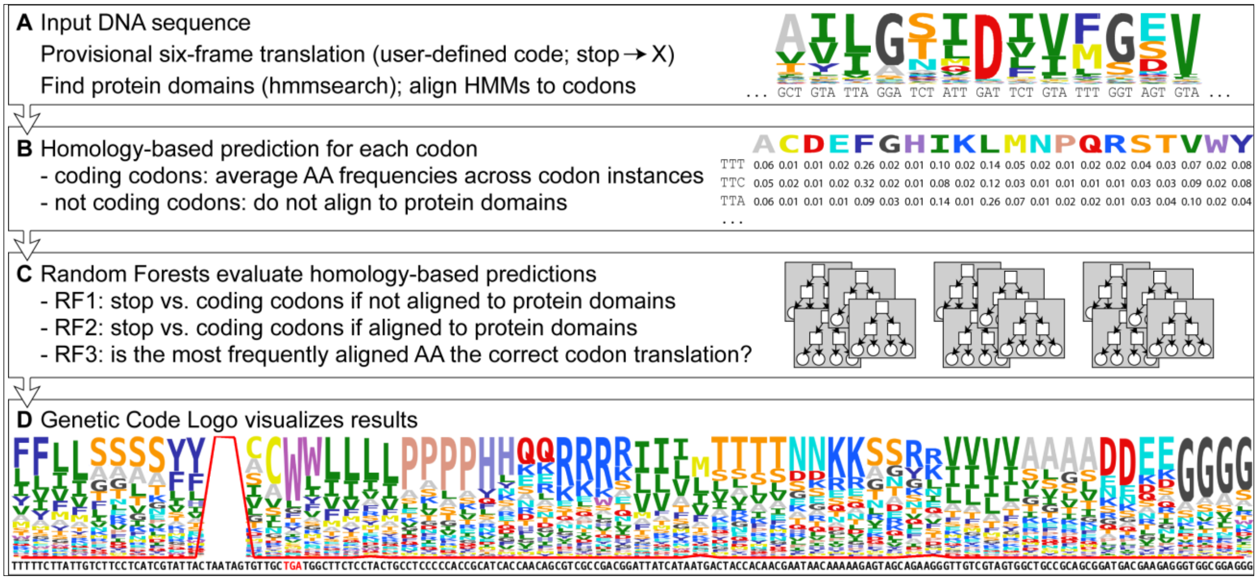

A Input DNA sequence
Provisional six-frame translation (user-defined code; stop → X)
Find protein domains (hmmsearch); align HMMs to codons
... GCT GTA TTA GGA TCT ATT GAT TCT GTA TTT GGT AGT GTA ...
B Homology-based prediction for each codon
- coding codons: average AA frequencies across codon instances
- not coding codons: do not align to protein domains
A C D E F G H I K L M N P Q R S T V W Y
TTT 0.06 0.01 0.01 0.02 0.26 0.02 0.01 0.10 0.02 0.14 0.05 0.02 0.01 0.02 0.02 0.04 0.03 0.07 0.02 0.08
TTC 0.05 0.02 0.01 0.02 0.32 0.02 0.01 0.08 0.02 0.12 0.03 0.01 0.01 0.01 0.03 0.03 0.09 0.02 0.08
TTA 0.06 0.01 0.01 0.01 0.09 0.03 0.01 0.14 0.01 0.26 0.07 0.01 0.02 0.02 0.01 0.03 0.04 0.10 0.02 0.04
...
C Random Forests evaluate homology-based predictions
- RF1: stop vs. coding codons if not aligned to protein domains
- RF2: stop vs. coding codons if aligned to protein domains
- RF3: is the most frequently aligned AA the correct codon translation?
D Genetic Code Logo visualizes results
TTTTTCTTATTGTCTTCCTCATCGTATTACTAATAGTGTTGCTGATGGCTTCTCCTACTGCCTCCCCCACCGCATCACCAACAGCGTCGCCGACGGATTATCATAATGACTACCACAACGAATAACAAAAAGAGTAGCAGAAGGGTTGTCGTAGTGGCTGCCGCAGCGGATGACGAAGAGGGTGGCGGAGGG

Fig. 1. Outline of the FACIL algorithm, see text for details. The Genetic Code Logo visualizes the results, including the reliability of alternative genetic code predictions. The example shows the predicted code for *Globobulimina pseudospinescens* mitochondrial fragments, generated by entering the "example" input data on the FACIL webserver. The logo shows the 64 codons from left to right (predicted alternatives in red), each with a stack of AAs. The stack height indicates the percentage of RF3 trees supporting the predicted translation, the letter sizes indicate the scaled AA alignment scores and the red line is the percentage of RF1/RF2 trees that predict a stop codon.

set by randomly selecting 1,000 regions each of length 100, 200, 300, ..., 1,000, 2,000, 3,000, ..., 10,000, 20,000, 30,000, ..., 100,000, 200,000, 300,000, ..., and 1,000,000 nt, i.e. a total of 37,000 fragments. We made sure these genomic regions did not overlap to avoid redundant training data. The complete set of training data is available from the FACIL website: http://www.cmbi.ru.nl/FACIL/input/complete_training_table.txt.gz.

2.2 Random Forest analysis

RF is a non-parametric classification algorithm that uses many classification trees in parallel (Breiman, 2001). It uses a random subset of the cases in the training data set and the remainder of the cases for testing and calculating the accuracy scores. The randomForest R package version 2.11.0 was used with 100 trees (using 1,000 trees gave almost the same results, see tab in Supplementary Table 1) and default parameters (63% of codons used for training, 37% for testing, square root of the number of variables to train individual trees). In FACIL, the RFs assess the correctness of the homology-based predictions: the response variables of RF1 and RF2 were "stop codon" or "coding codon", the response of RF3 was either "correct AA translation" or "incorrect AA translation".

2.3 *Globobulimina pseudospinescens* sequencing and assembly

Approximately 10,000 single-cell *Globobulimina pseudospinescens* organisms were isolated by hand from Gullmar Fjord sediment (Risgaard-Petersen, et al., 2006). After washing, total DNA was extracted using the QIAamp DNA Micro Kit and sequenced by Illumina Genome Analyser II. 9,950,730 32 nt reads were assembled using Edena (Hernandez, et al., 2008) with parameters m=16 and M=16, which yielded the highest N50 value (N50=170). The raw data and assembly are available from the Gene Expression Omnibus (GEO) under accession number GSE26664. The total DNA of a eukaryote may contain up to three different translation codes (De Grey, 2005): nuclear, mitochondrial, and plastid (if the organism is photosynthetic, but this is not the case for *G. pseudospinescens*). To avoid mixing these signals, the user can choose to feed individual contigs to FACIL, but this might lead to a bad genetic code prediction due to shortage of data. Thus, we selected those contigs that were likely derived from the *G. pseudospinescens* mitochondrial genome as follows. The 8,456 assembled contigs were queried by BlastX version 2.2.22+ (Camacho, et al., 2009) against all proteins encoded by completely sequenced mitochondria, downloaded from NCBI organelle genome resources (http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=2759) on July 28th, 2010. Importantly, we used the standard genetic code for this BlastX search in order not to impose a bias in the genetic code on the contigs and our results. The 150 contigs with a high-scoring BlastX hit (E-value ≤0.01) were considered to be of mitochondrial origin (average length 223 nt, median length 191 nt). These sequences are available as "example" input data on the FACIL webserver. They contain fragments of mitochondrial genes like cytochrome B and several ATP synthase, cytochrome-c oxidase and NADH dehydrogenase subunits. We found no evidence for multiple copies (e.g. a nuclear and a mitochondrially encoded copy) of the encoded genes after a BlastN search (E-value ≤0.01) of the contigs against themselves.

3 RESULTS

3.1 Homology-based prediction

We present FACIL (Fast and Accurate genetic Code Inference and Logo), a method to predict and evaluate the coding of every codon for any nucleic acid sequence, without requiring a priori annotation of proteins. First, FACIL queries all Pfam-fs protein domain HMMs (local alignment models (Finn, et al., 2010)) against a provisional six-frame translation of the DNA. All known variant codes differ by at most a few codons, so a provisional translation can help to align the AAs in the protein domains to the codons in the DNA. By default, our provisional translation uses the standard code, but by iterating FACIL, using the newly identified codes as input, it is in principle possible to find more distant codes. Stop codons are translated as X to enable the initial alignment of all sites by hmmsearch (HMMER 3.0, http://hmmer.org; default parameters and Pfam trusted score cutoff; Figure 1a). This sensitive profile-based homology search algorithm allows FACIL to identify homologous regions even if codons are consistently mis-translated. Thus, potentially unique codes can be identified even for organisms that are taxonomically divergent from known species, provided that homologous domains are found. For each codon, FACIL examines which AAs are most frequently associated to it among the aligned protein domains, taking into account the frequency distribution of AAs per position as defined by the domain HMMs. Because we use protein domains as the search unit and the speed of the HMMER 3.0 hmmsearch algorithm, FACIL is extremely fast and insensitive to fragmented DNA, frameshifts due to e.g. sequencing or assembly errors, introns and split gene sequences, and does not require gene annotation.

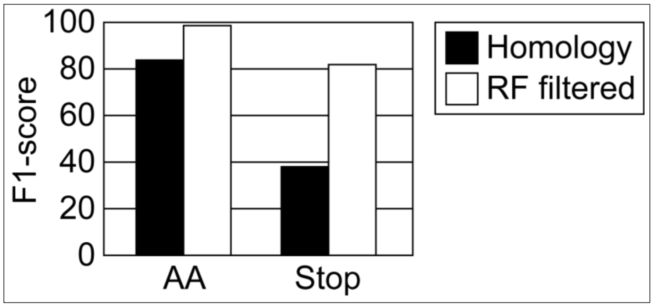


Fig. 2. F1-score for predicting coding (AA) and stop codons by homology alone and after RF filtering. Values are based on the predictions for all codons from bacterial, archaeal and organellar genomes (see Supplementary Table 2).

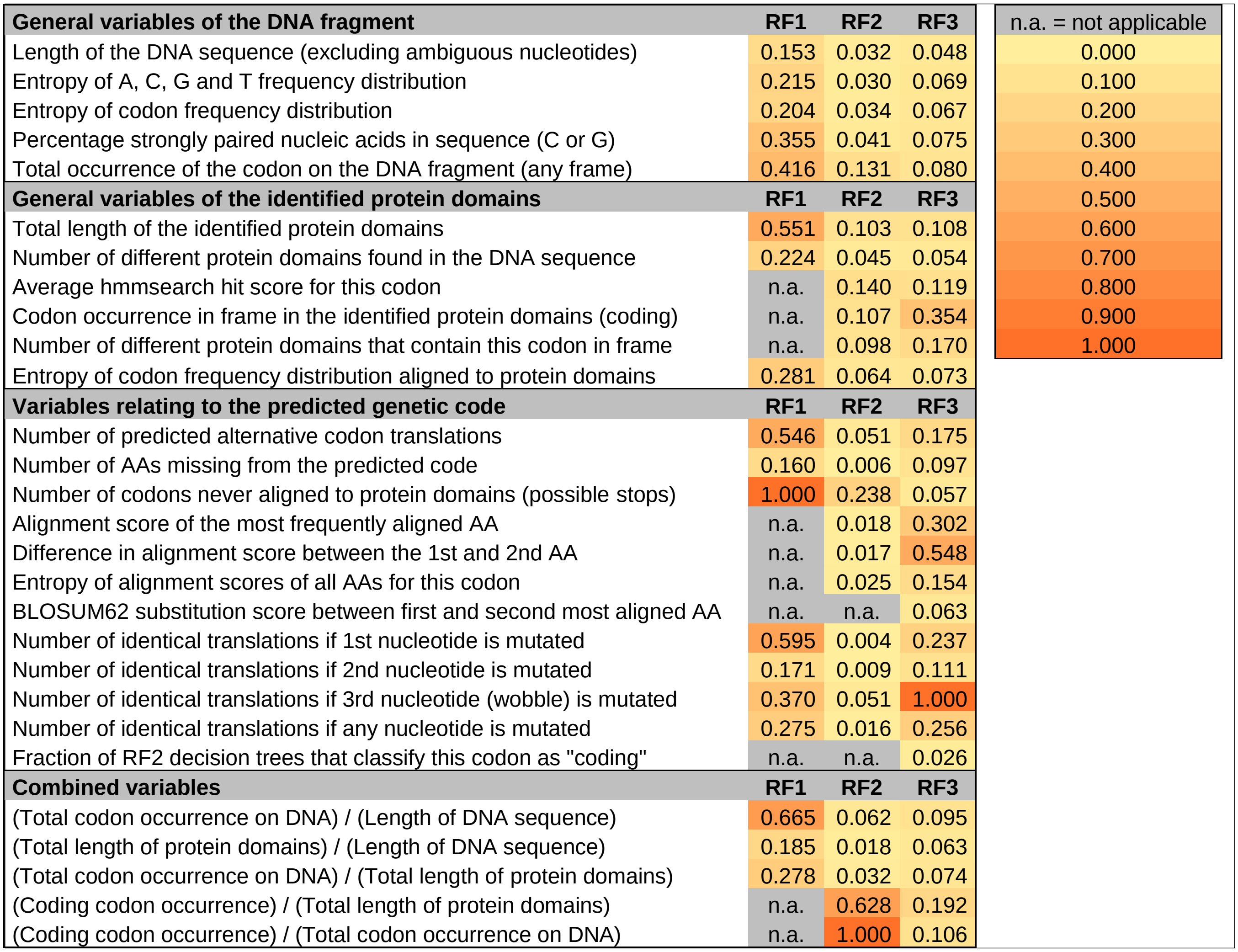
3.2 Random forest-based evaluation of homology-based prediction

The homology-based prediction creates a matrix of 64 codons by 20 AAs for each DNA molecule (Figure 1b), where we consider the AA that most frequently aligns to a codon within the protein domains as its most likely translation. 83.3% of the AAs thus predicted are correct (Figure 2), but stop codons (that do not align to the protein domains) may be over-predicted for short sequences with strong codon bias, causing a low precision of 24.0% for stop codons. Also, AAs with similar properties may align to a codon with almost equal frequencies, due to neutral evolution at the protein level (see ATA in Figure 1d). All in all, this leads to relatively low precision and sensitivity scores (see Supplementary Table 2: precision: 83.3% and 24.0%, sensitivity: 83.3% and 89.6%, for AAs and stop codons, respectively). We expected that these errors can be identified by inspecting the variables relating to the homology-based prediction (Table 1). To quantify the reliability of the predictions and assess which parameters are important to achieve a reliable prediction, we implemented a Random Forest approach (RF; Figure 1c). RF is a non-parametric classification algorithm capable of integrating many variables, yet difficult to overtrain due to the use of many classification trees in parallel that each are trained with a subset of the training data (Breiman, 2001). RFs can even capture sub-classes in the training data: clusters of instances with a specific variable importance.

For every DNA sequence, we evaluate a range of variables for each of the 64 codons to estimate the confidence of the homology-based prediction (Table 1). Firstly, we include general variables of the DNA fragment including sequence length and the total occurrence of the codon in the sequence. Secondly, we include variables of the identified protein domains like the average hmmsearch hit score. Thirdly, we include variables relating to the predicted genetic code (e.g. the number of AAs missing from the predicted genetic code, the number of codons never aligned to protein domains), variables that represent the confidence of the homology-based prediction (e.g. the similarity between the two top-scoring AAs as defined by their BLOSUM62 substitution score (Henikoff and Henikoff, 1992)) and variables that relate to the robustness of the predicted genetic code (number of single mutation codons translated to the same AA). Finally, we include several combined parameters, including the fraction of codons occurring in frame within the protein domains over their occurrence in the entire DNA sequence.

We trained three RFs of 100 trees, each specialized to answer a specific question. RF1 (91.03% accuracy) and RF2 (99.95% accuracy) were designed to discern stops from coding codons among those codons that do not and do align to protein domains, respectively. RF3 (95.08% accuracy) predicts whether the AA that most frequently aligns to a codon is indeed its correct translation. The assessment of the homology-based predictions by these RFs increased the precision and sensitivity scores (see Supplementary Table 2: precision: 97.1% and 99.3%, sensitivity: 88.1% and 75.8%, for AAs and stop codons, respectively). The decrease in sensitivity for stop codons is mainly due to the many training cases where not all codons are present in the protein domain alignments, e.g. for short input sequences. Thus, RF1 is strict in accepting them as true stop codons. We recommend to be critical of potentially novel alternative "rare coding codons", especially when analyzing short input sequences. Note that these are cases where FACIL will not predict an AA translation, as the codon is not aligned to any protein domain.

We found different variables to be important to each of these questions. For codons that are never aligned to protein domains, the most important variable to distinguish true stop codons from rare coding codons (RF1) is how many of the 64 possible codons did not align to any protein domain. If the sequence contains many codons that never align to a protein domain, this is likely a result of a combination of the low number of identified protein domains, the short length or the low complexity of the query sequence, although individually, those parameters were less important to RF1. Among codons that do align to protein domains, coding codons can be distinguished from spuriously aligned stop codons (RF2) by their occurrence ratio in-frame within protein domains and in the entire sequence. This includes off-frame occurrence in the coding region, where it has been hypothesized that stop codons are abundant to terminate frame-shifted translation (Seligmann and Pollock, 2004). To determine if the AA with the highest alignment score is indeed the correct coding translation for a codon (RF3), the difference between the first and second best alignment score is an important variable. Interestingly, however, the most important variable in RF3 turns out to be a basic characteristic of the genetic code, i.e. the translation redundancy at the third nucleotide of the codon. The genetic code is characterized by a low impact of wobble base-pairing of tRNAs at the third nucleotide and apparently, spuriously high scoring AAs can be recognized as being in violation with this



General variables of the DNA fragment	RF1	RF2	RF3
Length of the DNA sequence (excluding ambiguous nucleotides)	0.153	0.032	0.048
Entropy of A, C, G and T frequency distribution	0.215	0.030	0.069
Entropy of codon frequency distribution	0.204	0.034	0.067
Percentage strongly paired nucleic acids in sequence (C or G)	0.355	0.041	0.075
Total occurrence of the codon on the DNA fragment (any frame)	0.416	0.131	0.080
General variables of the identified protein domains	**RF1**	**RF2**	**RF3**
Total length of the identified protein domains	0.551	0.103	0.108
Number of different protein domains found in the DNA sequence	0.224	0.045	0.054
Average hmmsearch hit score for this codon	n.a.	0.140	0.119
Codon occurrence in frame in the identified protein domains (coding)	n.a.	0.107	0.354
Number of different protein domains that contain this codon in frame	n.a.	0.098	0.170
Entropy of codon frequency distribution aligned to protein domains	0.281	0.064	0.073
Variables relating to the predicted genetic code	**RF1**	**RF2**	**RF3**
Number of predicted alternative codon translations	0.546	0.051	0.175
Number of AAs missing from the predicted code	0.160	0.006	0.097
Number of codons never aligned to protein domains (possible stops)	1.000	0.238	0.057
Alignment score of the most frequently aligned AA	n.a.	0.018	0.302
Difference in alignment score between the 1st and 2nd AA	n.a.	0.017	0.548
Entropy of alignment scores of all AAs for this codon	n.a.	0.025	0.154
BLOSUM62 substitution score between first and second most aligned AA	n.a.	n.a.	0.063
Number of identical translations if 1st nucleotide is mutated	0.595	0.004	0.237
Number of identical translations if 2nd nucleotide is mutated	0.171	0.009	0.111
Number of identical translations if 3rd nucleotide (wobble) is mutated	0.370	0.051	1.000
Number of identical translations if any nucleotide is mutated	0.275	0.016	0.256
Fraction of RF2 decision trees that classify this codon as "coding"	n.a.	n.a.	0.026
Combined variables	**RF1**	**RF2**	**RF3**
(Total codon occurrence on DNA) / (Length of DNA sequence)	0.665	0.062	0.095
(Total length of protein domains) / (Length of DNA sequence)	0.185	0.018	0.063
(Total codon occurrence on DNA) / (Total length of protein domains)	0.278	0.032	0.074
(Coding codon occurrence) / (Total length of protein domains)	n.a.	0.628	0.192
(Coding codon occurrence) / (Total codon occurrence on DNA)	n.a.	1.000	0.106

n.a. = not applicable
0.000
0.100
0.200
0.300
0.400
0.500
0.600
0.700
0.800
0.900
1.000

Table 1. Variables used in each RF and their normalized importance (calculated as MeanDecreaseGini / max MeanDecreaseGini; see Supplementary Table 1).

rule. Note that, e.g., the amount of protein-coding sequence falling into Pfam-domains is not an important distinguishing variable in any of the RFs. Its major contribution is in RF1, which is in accordance with the most important RF1 variable, i.e. the number of codons that did not align to any protein domain.

To assess potentially conflicting variable importance between standard and alternative codons, we did an additional experiment where we trained the RFs for each of these groups separately (see tabs in Supplementary Table 1). While the variable importances for the RFs trained with only standard codons were very comparable to the complete set, the RFs trained with alternatively encoded codons gave a different picture. For alternative codons that are never aligned to protein domains, the most important variable to distinguish true stop codons from rare coding codons (RF1) is the percentage of strongly paired nucleic acids (GC content). This reflects the difficulty in predicting alternative genetic codes for genomes with a high GC-skew. For alternative codons that do align to protein domains, RF2 assigns the highest importance to the difference in alignment score between the first and second most frequently aligned amino acid. To determine if the AA assigned by homology is indeed the correct coding translation (RF3), the translation redundancy at the first nucleotide of the codon is the most important distinguishing variable. This analysis pinpoints some of the specific variables that are important for the prediction of alternative codons.

3.3 Genetic Code Logo and web server

The output of FACIL, i.e. a predicted translation for each codon along with confidence values based on the supporting fraction of decision trees in the RFs, is visualized in a Genetic Code Logo (Figure 1d). We implemented FACIL into a web server (www.cmbi.ru.nl/FACIL/) that enables the user to easily obtain a code prediction with details and a Genetic Code Logo for any sequenced genome or set of contigs. This site also contains a downloadable stand-alone version of the software. Both the web server and the stand-alone version of FACIL take FASTA formatted DNA sequences as input and allow the user to specify the genetic code used for the provisional translation. Mitochondrial genetic code of *Globobulimina pseudospinescens*

Alternative genetic codes are perhaps most abundant in mitochondrial genomes. To illustrate the use of our method, we set out to decipher the genetic code of the mitochondrial genome of the foraminifer *Globobulimina pseudospinescens*. *Foraminifera* belong to the *Rhizaria*, a kingdom with only very few protein sequences in the databases, none of which are derived from a mitochondrial genome. This means that no close relatives of this species are represented in the Pfam-fs protein domains used in FACIL. Moreover, the data are particularly challenging, as the genome sequence is highly fragmented and incomplete (150 contigs with an average length of 223nt). Nevertheless, we obtain strong support that the *G. pseudospinescens* mitochondrial genome uses the "Protozoan Mitochondrial Code" (NCBI translation table 4, see Figure 1d and Supplementary Figure 1), with all 62 high-confidence AA translations correctly classified by RF3, including the alternative translation of TGA into tryptophan (W). Both stop codons (TAA and TAG) were identified by RF1 (red line). BlastX searches and manual curation are consistent with these results (Supplementary Dataset 1 and Supplementary Table 4). Running the 8,306 remaining contigs (average length 168 nt) through FACIL predicted the "Standard Code" (NCBI translation table 1) as the most similar code, with 63 of the 64 codons predicted correctly. In the homology step, ATG was more often aligned to leucine (L) than to methionine (M; see Supplementary Figure 2), but that translation was considered unreliable by RF3 and filtered out. TGA was identified as a stop codon. This analysis exemplifies the value of our method for the reliable discovery of code variants, even in fragmented DNA from taxonomically divergent organisms.

3.4 Performance

As explained above, a FACIL query consists of two main steps. The first is a homology search where the six-frame translation of the input sequences are queried for known protein domains by hmmsearch (Figure 1a), the second is an evaluation of the alignment-based predictions by three specialized RFs (Figure 1c). For the 150 *G. pseudospinescens* sequences (length ~223nt) presented as an example, these steps take approximately four and one minutes, respectively, on our current web server (3GHz, 32Gb memory). This brings the total run time for prediction of the genetic code to five minutes, only a fraction of the 50 minutes required for a BlastX search against the proteins in the NCBI Refseq protein database on the same machine (E-value ≤0.01; Supplementary Dataset 1). Indeed, the main performance gain of FACIL comes from the difference in database size that it queries. FACIL only needs to go through 9,318 Pfam-fs profiles, whereas the BlastX-based analysis queries at least the Refseq database (9,004,816 proteins in the December 2010 version we used) and preferably even NR, especially for less well-characterized organisms. Moreover, the BlastX results need to be parsed by a custom script and simply selecting the most often aligned AA for each codon may lead to errors, e.g. the two stop codons are occasionally aligned in BlastX hits (see Supplementary Table 4). The RFs in FACIL filter out these cases. For large-scale data sets, great improvement may be expected by running hmmsearch in parallel on a high-speed hardware accelerator.

4 DISCUSSION

Currently, there is no standard available for inference of the genetic code of an unannotated DNA sequence, and a range of *ad hoc* methods that lack quality control and reported reliability scores obscure this research area (the notable exception being Gendecoder (Abascal, et al., 2006)). With FACIL, we present an easy, fast and reliable tool to predict the genetic code for nucleic acid sequences that does not depend on any a priori gene annotation. FACIL detects alternative genetic codes even in species distantly related to known organisms.

Previously, genetic code prediction has explicitly (Gendecoder (Abascal, et al., 2006)) or implicitly (BlastX searches) benefited from phylogenetic relatedness for reliable predictions. With FACIL, we chose to rely on general protein domains for two reasons. Firstly, this eliminates the requirement to know the taxonomic placement of the organism from which the DNA was derived, which may in particular be difficult for early-branching organisms. Secondly, this greatly improves its speed (see Section 3.5). Nevertheless, genetic code prediction may benefit from a more phylogenetically balanced selection of reference sequences, a proper model of evolution and a probabilistic phylogenetic method that considers the amino acids at neighboring nodes of the tree and uses branch lengths to calculate the probability of amino acids at an unknown state node.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their useful suggestions. *Funding*: B.E.D. was supported by the Netherlands Organisation for Scientific Research (NWO) Horizon grant (050-71-058) and by NWO Veni grant (016.111.075). R.J. was supported through the Virgo consortium (BSIK 03012), an Innovative Cluster approved by the Netherlands Genomics Initiative. R.S. was supported by NWO Horizon grant (050-71-053). M.Sc. was supported by the Darwin Center for Biogeosciences (project 142.16.1091). M.St. was supported by ERC starting grant MASEM (242635). This work is part of the programme of BiG Grid, the Dutch e-Science Grid, which is financially supported by NWO.

REFERENCES

Abascal, F., Zardoya, R. and Posada, D. (2006) GenDecoder: genetic code prediction for metazoan mitochondria, *Nucleic Acids Res*, **34**, W389-393.

Barrell, B.G., Bankier, A.T. and Drouin, J. (1979) A different genetic code in human mitochondria, *Nature*, **282**, 189-194.

Bove, J.M. (1993) Molecular features of mollicutes, *Clin Infect Dis*, **17 Suppl 1**, S10-31.

Breiman, L. (2001) Random forests, *Machine Learning*, **45**, 5-32.

Camacho, C., *et al.* (2009) BLAST+: architecture and applications, *BMC Bioinformatics*, **10**, 421.

Finn, R.D., *et al.* (2010) The Pfam protein families database, *Nucleic Acids Res*, **38**, D211-222.

De Grey, A.D.N.J (2005) Forces maintaining organellar genomes: is any as strong as genetic code disparity or hydrophobicity? *Bioessays*, **27**:436-446.

Helftenbein, E. (1985) Nucleotide sequence of a macronuclear DNA molecule coding for alpha-tubulin from the ciliate Stylonychia lemnae. Special codon usage: TAA is not a translation termination codon, *Nucleic Acids Res*, **13**, 415-433.

Henikoff, S. and Henikoff, J.G. (1992) Amino-Acid Substitution Matrices from Protein Blocks, *P Natl Acad Sci USA*, **89**, 10915-10919.

Hernandez, D., *et al.* (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer, *Genome Res*, **18**, 802-809.

Koonin, E.V. and Novozhilov, A.S. (2009) Origin and evolution of the genetic code: the universal enigma, *IUBMB Life*, **61**, 99-111.

Meyer, F., *et al.* (1991) UGA is translated as cysteine in pheromone 3 of Euplotes octocarinatus, *Proc Natl Acad Sci U S A*, **88**, 3758-3761.

Risgaard-Petersen, N., *et al.* (2006) Evidence for complete denitrification in a benthic foraminifer, *Nature*, **443**, 93-96.

Seligmann, H. and Pollock, D.D. (2004) The ambush hypothesis: hidden stop codons prevent off-frame gene reading, *DNA Cell Biol*, **23**, 701-705.

Shackelton, L.A. and Holmes, E.C. (2008) The role of alternative genetic codes in viral evolution and emergence, *J Theor Biol*, **254**, 128-134.

Wu, D., *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea, *Nature*, **462**, 1056-1060.