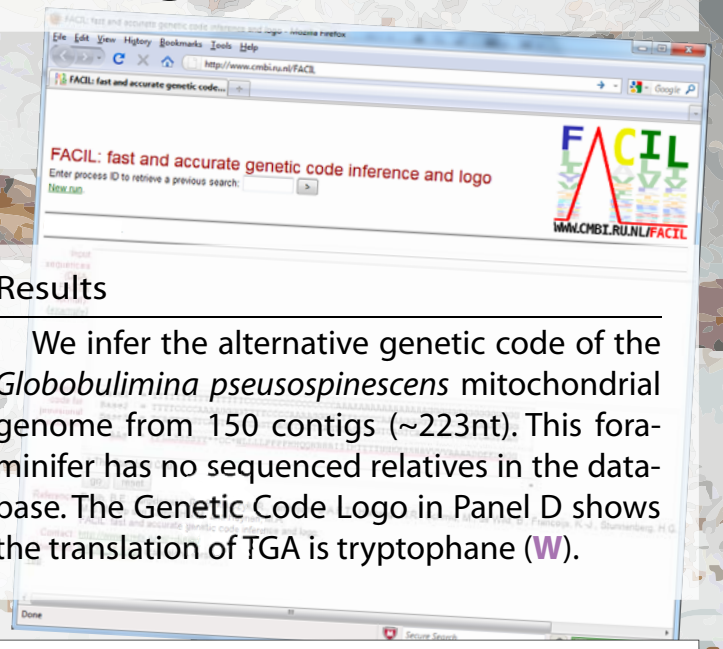
FACIL: fast and accurate genetic code inference and logo

B.E. Dutilh, R. Jurgelenaite, R. Szklarczyk, S.A.F.T. van Hijum, H.R. Harhangi, M. Schmid, B. de Wild, K.-J. François, H.G. Stunnenberg, M. Strous, M.S.M. Jetten, H.J.M. Op den Camp and M.A. Huynen



Summary

Alternative genetic codes (0.65% of Genbank) occur in organelles, bacteria, eukaryotes and their associated viruses. Default translation may confound downstream analyses.

FACIL evaluates DNA sequences or sequence fragments for their genetic code.

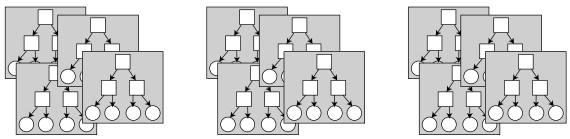
Results

We infer the alternative genetic code of the *Globobulimina pseusospinescens* mitochondrial genome from 150 contigs (~223nt). This foraminifer has no sequenced relatives in the database. The Genetic Code Logo in Panel D shows the translation of TGA is tryptophane (W).

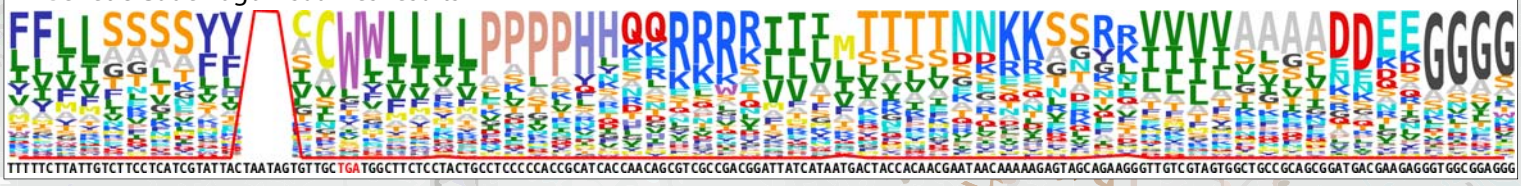
A Input DNA sequence
Provisional six-frame translation (user-defined code; stop→X)
Find protein domains (hmmsearch); align HMMs to codons

B Homology-based prediction for each codon
- coding codons: average AA frequencies across codon instances
- not coding codons: do not align to protein domains

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?



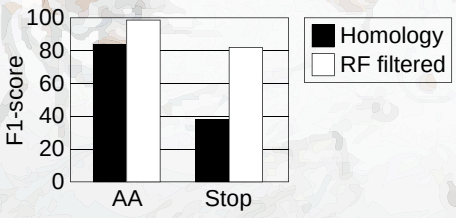
D Genetic Code Logo visualizes results



Methods

FACIL scans a provisional six-frame translation of the DNA for protein domains using sensitive HMM searches (Panel A).

The resulting homology-based prediction (B) for each codon is evaluated by three Random Forests (C), which improves the reliability.



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

Reference: B.E. Dutilh et al. (2011) "FACIL: fast and accurate genetic code inference and logo", *Bioinformatics*.