Published online 18 January 2008

Nucleic Acids Research, 2008, Vol. 36, No. 3 ***e16***
doi:10.1093/nar/gkm1181

Codon choice in genes depends on flanking sequence information—implications for theoretical reverse translation

Karthikeyan Sivaraman[1], AswinSaiNarain Seshasayee[2], Patrick M. Tarwater[3] and Alexander M. Cole[1,*]

[1]Department of Molecular Biology and Microbiology, Burnett School of Biomedical Sciences, University of Central Florida, Orlando, FL, 32816, USA, [2]Genomics and Regulatory Systems Group, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [3]Department of Biostatistics, University of Texas School of Public Health, El Paso, TX, 79902, USA

Received September 26, 2007; Revised December 7, 2007; Accepted December 27, 2007

ABSTRACT

Algorithms for theoretical reverse translation have direct applications in degenerate PCR. The conventional practice is to create several degenerate primers each of which variably encode the peptide region of interest. In the current work, for each codon we have analyzed the flanking residues in proteins and determined their influence on codon choice. From this, we created a method for theoretical reverse translation that includes information from flanking residues of the protein in question. Our method, named the neighbor correlation method (NCM) and its enhancement, the consensus-NCM (c-NCM) performed significantly better than the conventional codon-usage statistic method (CSM). Using the methods NCM and c-NCM, we were able to increase the average sequence identity from 77% up to 81%. Furthermore, we revealed a significant increase in coverage, at 80% identity, from $< 20\%$ (CSM) to $> 75\%$ (c-NCM). The algorithms, their applications and implications are discussed herein.

INTRODUCTION

Word usage and codon usage in bacterial genomes has been extensively documented, both in the coding (1) and non-coding regions (2). These reports show that word usage in genomes is non-random and it serves as a biological signature of the organism in question. One such signature is codon usage in open reading frames (ORFs), and is reflected in measures such as the codon adaptation index (CAI) (3). Though CAI provides a convenient measure of codon bias, several reports show that codon usage is not a property of isolated codons and in several cases the bases immediately upstream or downstream affect the translation (4). Such neighboring base effects are well studied in case of stop codon read-through experiments where the flanking base or codon has been shown to affect the accuracy and magnitude of read-through (5). Apart from single bases, the effect of flanking codons has also been well studied in literature. Gutman and Hatfield (6) show that there is a strong first-order Markovian relationship between codons in a gene and this relation is seen even after translation, in proteins. Boycheva and colleagues extended this study to reveal that translation efficiency is strongly dependent on the dicodon pair that encodes for a given amino acid pair (7). They suggest that relative orientations of t-RNA in the ribosome may cause the observed differences in translation efficiency and subsequently certain dicodon pairs are selected evolutionarily. Moura and coworkers use a more recent and larger dataset for an analysis of dicodon usage patterns in both prokaryotes and eukaryotes. Their results suggest that the geometric constraints imposed by the translation machinery are driving forces in the evolution of gene sequences in bacteria (8). Collectively, these results suggest the existence of strong first-order Markovian relationships between codons in a gene. We hypothesized that information content of such correlations is carried over to the proteins, at least in part, when the gene is translated. This information manifests itself as a lack of randomness in the choice of codons and it is apparent when one attempts to theoretically reverse translate a protein sequence.

Reverse translation has been discussed earlier as an abstract logical flow of information from proteins to DNA (9). In this work, we consider the pragmatic problem of theoretical reverse translation itself, rather than that of information flow from proteins to DNA. Theoretical reverse translation of protein sequences has potential applications in primer design for degenerate

*To whom correspondence should be addressed. Tel: 407 823 3633; 407 823 3635; Email: acole@mail.ucf.edu

PCR and in design of synthetic genes (10). In degenerate PCR, several primers are designed, each representing a variant DNA sequence encoding the peptide region of interest. One of the best methods designed for degenerate PCR can, in the best case scenarios, still utilize up to 128 primers on one end (5′- or 3′-end) and one or more at the other end (11). Though no specific software is available for reverse translation, the conventional procedure is to substitute codons for residues based on the overall genomic codon usage probabilities which required different primers be designed for each ambiguous codon in the gene in the region of interest. In practice, it is common for almost all possibilities to be covered, increasing the number of required primers exponentially. Thus, improvements in reverse translation will help reduce the ambiguity in degenerate PCR.

Improvements in reverse translation can be brought about by studying the rules of codon usage in the genome, which is feasible due to availability of whole genome sequences. In this study, we created a framework for reverse translation of bacterial gene sequences and term it the neighbor correlation method (NCM), due to its use of neighboring (flanking) sequence information to predict codon usage. We provide evidence for the dependency of codon choice on the flanking amino acid residues and used this dependency to reverse-translate protein sequences from two model genomes. We confirmed that NCM was a substantial improvement over the conventional method (codon-usage statistic method—CSM). Furthermore, we introduced a modification to both CSM and NCM [consensus CSM (c-CSM) and consensus NCM (c-NCM)] to improve significantly the sensitivity of reverse translations by both CSM and NCM, and show that these observed differences in performance are statistically significant. Finally, using the protein sequences of *Salmonella typhi* CT18 and the probability matrix from *Escherichia coli* K12, we show that it is possible to reverse translate sequences from organisms for which a reverse translation matrix is not available, by using a matrix from a related organism.

MATERIALS AND METHODS

All sequences were obtained from the NCBI database. For the analyses, the genome and predicted ORF sequences of *E. coli* K12 (12), *B. subtilis* (13), and *S. typhi* CT18 (14), *Acidobacteria bacterium* (NC_008095), *Aquifex aeolicus* (15), *Bacteroides thetaiotaomicron* (16), *Bordetella pertussis* (17), *Campylobacter jejuni* (18), *Caulobacter crescentus* (19), *Chlamydia trachomatis* (20), *Clostridium acetobutylicum* (21), *Dehalococcoides ehtenogenes* (22), *Deinococcus radiodurans* (23), *Fusobacterium nucleatum* (24), *Lactobacillus acidophilus* (25), *Mesorhizobium loti* (26), *Methanococcus jannaschii* (27), *Methanopyrus kandleri* (28), *Mycobacterium bovis* (29), *Mycobacterium tuberculosis* (30), *Mycoplasma genitalium* (31), *Myxococcus xanthus* (32), *Nanoarchaeum equitans* (33), *Prochlorococcus marinus* (34), *Pseudomonas aeruginosa* (35), *Rickettsia prowazekii* (36), *Sulfolobus solfataricus* (37), *Synechococcus elongatus* (38), *Thermoplasma acidophilum* (39), *Ureaplasma urealyticum* (40) and *Magnetococcus* sp. (NC_008576) were used. We used *needle*, an implementation of the Needleman–Wünsch algorithm available in the EMBOSS package (41) for all sequence identity analyses. The algorithms discussed were implemented in PERL (script provided as Supplementary Data) on a Linux platform.

Analysis for non-random codon usage dependency on flanking amino acid residues

For codons of interest, random occurrence model was constructed based on codon usage and amino acid frequencies in a given genome. We used 10 000 such random sets to calculate the *z*-scores for each residue–codon–residue combination. From the *z*-scores, *P*-values were calculated and were multiply corrected for both codon occurrence and amino acid occurrence biases using *Bonferroni correction*. To identify those combinations that have a skewed occurrence, we used a stringent threshold of $P < 0.0001$.

Creation of the probability matrix for CSM

Codon usage in the genome interest was calculated using the CUSP program in the EMBOSS package (41), and a codon usage probability table was created based on that information. For each amino acid the segmented probability interval spans from 0.0 to 1.0, where each consecutive non-overlapping segment corresponds to probability of a unique codon (Figure 1A). This probability interval matrix had 64 individual data points under 21 categories (20 amino acids + stop codons).

Creation of the probability matrix for NCM

For each tripeptide A1–A2–A3 in the genome of interest, we calculated the usage probabilities of codons for A2 flanked by A1 and A3. Based of these probabilities, we created a probability interval matrix for all combinations of A1–C*–A3, where C* is the codon that encodes A2. The probability interval matrix thus created had 24 400 individual data points under 8000 categories ($20 \times 20 \times 20$ amino acid combinations). Creation of such a probability interval for the tripeptide S–A–S is illustrated in Figure 1B.

Reverse translation

In reverse translation using CSM, a random number *r* was generated where $0 \leq r \leq 1$, for each amino acid in the query protein sequence. The codon corresponding to the probability interval within which *r* falls was chosen for reverse translation. In NCM, overlapping tripeptides were used instead of single codons, and the codon was predicted for the second residue. However, when reverse translating with NCM, the first residue and stop codons were assigned based on probability alone. This procedure is also illustrated in Figure 1. c-NCM was created as an enhancement to NCM, in which reverse translation was performed *n* times using NCM for each protein sequence. The final DNA sequence was obtained by creating a consensus sequence from the *n* sequences created.

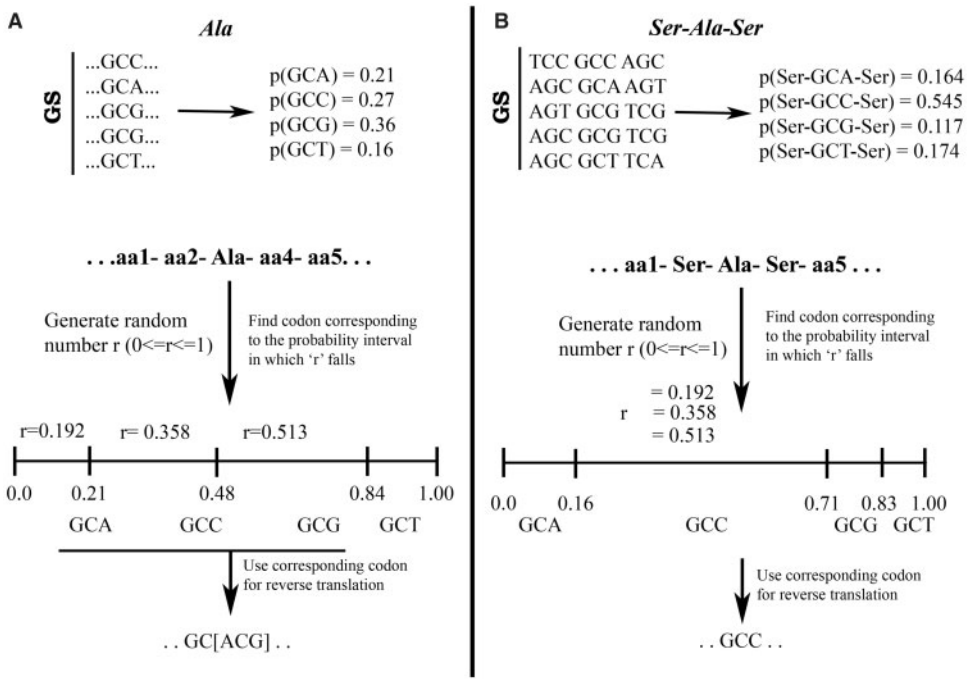

Figure 1. Illustration of reverse translation methods. (**A**) shows reverse translation of a protein sequence based on codon usage and (**B**) shows the reverse translation using NCM. GS represents ORF (gene) sequences from the genome of interest. The first part shows the creation of probability intervals for both panels. For NCM, Bayesian probabilities of codon usage were calculated given the flanking residues. Note that the codon usage profiles for alanine are distinct between the two methods. The second part depicts the reverse translation process, which is similar to both methods. A random number '*r*' was generated and the codon corresponding to the probability interval (the horizontal line spanning 0.0–1.0 in both panels) within which *r* fell was used for creation of the ORF. This codon was then used for reverse translation.

Statistical analyses of differences between various methods

In order to statistically test the difference in performance of the different methods, we used (i) either Kolmgorov–Smirnov (KS) or Mann–Whitney (MW) test for comparing distributions of nucleotide sequence identity and (ii) *F*-test followed by FDR to identify sequence identity range that is over-represented in one method over another. These tests were used to compare (i) c-NCM and NCM (ii) NCM and CSM and (iii) c-NCM and CSM. In case of KS and MW tests, we used the sequence identity data. For the *F*-test and subsequent FDR analysis, we used the number of sequences scoring within a given sequence identity interval (for example, 300 sequences scored between 80% and 85%). All tests were run in R (http://www.r-project.org). The complete statistical analysis and data are provided in Supplementary Data.

Statistical analyses of iteration threshold for c-NCM

The c-NCM was performed on a random set of 1000 sequences in the *E. coli* K12 genome. Various iterations were used, ranging from 5 to 100 in five steps. Resultant sequences were compared with reference gene sequences using *needle* and percentage identity calculated. The distribution of scores from 50 iterations was compared to (i) that of NCM for these 1000 sequences and (ii) the distribution of scores from 100 iterations. For the comparison, we used KS test with alternative hypothesis = greater. There was no significant difference between the scores of iterations 50 and 100 ($P = 0.2406$). However, there was a significant difference between NCM and the 50-iteration c-NCM (same test as above, $P < 2.2 \times 10^{-16}$), and hence we used 50 iterations as the threshold for c-NCM predictions. A similar approach was used to test the performance of c-CSM. The results of c-CSM were then compared with those of c-NCM.

RESULTS AND DISCUSSIONS

Reverse translation of protein sequences is necessary for the design of degenerate primers. In most cases, reverse translation uses the codon usage statistics of the complete genome or a representative set of genes for the organism of interest. While dictated by overall genomic preference, this method rests on the assumption that usage of a codon in a gene is essentially random. Until this study, there has been no comprehensive analysis on the statistics of reverse-translation using the classical method. In this work, we show that the choice of codons for reverse translation can be refined further by taking into account the residues flanking the residue of interest in a protein. Based on this observation, we have devised a method called the NCM that uses the correlation between codon usage and flanking residues in proteins. As a case study, we have analyzed the efficiency of reverse-translation using NCM performed on the set of predicted ORF of *E. coli* K12 and *B. subtilis*.

Table 1. Table showing strong distribution of the codon 'GGC' flanked by hydrophobic amino-acids (ILV)

Residue 1	codon	Residue 2	Occurrence (Occ)	p(R1-GGC-R2) (pX) = Occ/Total	pX/pRand
A	GGC	G	326	0.008 134	3.253
A	GGC	V	399	0.009 956	3.982
G	GGC	G	362	0.009 033	3.613
G	GGC	V	343	0.008 559	3.423
I	GGC	A	352	0.008 783	3.513
I	GGC	G	371	0.009 257	3.703
I	GGC	V	303	0.007 561	3.024
L	GGC	A	364	0.009 083	3.633
L	GGC	G	473	0.011 803	4.721
L	GGC	V	477	0.011 902	4.761
S	GGC	G	324	0.008 085	3.234
V	GGC	G	326	0.008 134	3.253

Occurrence in table denotes overall genomic occurrence of the combination. The pX denotes the occurrence probability of the combination X (occurrence/total occurrences of the codon GGC). The pRand denotes the random occurrence probability of the combination X (pRand = 1/400 = 0.0025). The pX/pRand denotes the ratio between observed and expected probabilities. These 12 combinations (out of 400) represent almost 12% of the total occurrences of GGC in the genome (expected = 3%) representing a skew in codon usage dependent on flanking residues.

Correlation between codon choice and the flanking amino acid residues in the *E. coli* K12 genome

We analyzed the codon usage in the genomes of both *E. coli* K12 and *B. subtilis* and observed that the codon usage was not random but was to some extent dependent on the flanking codons. This dependency on flanking codons was reflected as a dependency on the flanking residues in proteins. For example, the codon GGC (Gly) encodes for 40.5% of all glycine residues present in *E. coli* (Supplementary Data). In the NCM, there are 400 possible theoretical combinations for any given codon. If the distribution of GGC were to be random, each of the combinations would span 0.25% (random probability = 0.0025) of the probability space. However, we observed that GGC is often flanked by branched chain aliphatic amino acids and hydrophobic amino acids. The 12 combinations (3% of total possible combinations) shown in Table 1 contribute almost 12% of total GGC usage in the genome, yielding a usage that is as much as four times the expected random usage.

Though the analysis of GGX shows that codon usage is non random, the data discussed is specific for *E. coli*. Furthermore, glycine is encoded by only four codons and does not exhibit maximum degeneracy. In order to both test these observations in multiple genomes as well as to use a more degenerately encoded amino acid, we have analyzed the codon usage for the amino acid arginine in 30 genomes. Arginine is encoded by six codons and is amongst the most degenerately encoded amino acids along with leucine and serine. In our analysis, for each of the six codons (C), we generated 10 000 random distributions with flanking amino acid residues (R1–C–R2). Using these random distributions, *z*-scores and *P*-values for each observed combination were calculated. The calculated *P*-values were adjusted for both codon representation bias (for a given codon) and amino acid representation biases (across all codons for a given flanking pair) using the Bonferroni correction. The resultant values were screened using a stringent threshold of $P < 0.0001$. We observed that even after stringent corrections there were several combinations that had a non-random distribution. The results of these tests are given in Supplementary Data.

These tests prove that codon usage varies with a change in flanking amino acid residues. We therefore hypothesized that a method exploiting the flanking residue information will be more sensitive in detecting signals that are lost in the conventional method (CSM) for reverse translation.

Comparison and analysis of reverse translations using CSM and NCM

In order to compare the performance of CSM and NCM, we reverse translated all the proteins in two genomes, *E. coli* K12 and *B. subtilis*, using both methods. Identity of the reverse translated proteins with the reference (original) ORF was used to quantify sensitivity of the methods. First, the distribution of percentage identities of nucleotide sequences reverse translated via NCM is significantly greater than that for CSM ($P < 2.2 \times 10^{-16}$; one-tailed KS test). A second assessment of performance using ratios of sequence identities ($\%ID_{NCM}/\%ID_{CSM}$) revealed that there was a small yet statistically significant increase in average sequence identities ($P < 2.2 \times 10^{-16}$; one-tailed MW test, null-hypothesis: ratio = 1). The average increase in sequence identity for all the sequences was ~1%. We then grouped all sequence identities into bins of width 5% and tested which bins were significantly enriched in NCM over CSM. This revealed that NCM reverse translates a significantly large number of protein sequences to nucleotides withq high identities of >80–85% (*P*-value: 4.5×10^{-15}, Fisher's test and FDR correction; Figure 2A). At this sequence identity range, there are twice as many DNA sequences predicted by NCM (239 sequences) as are predicted by CSM (103 sequences).

With *B. subtilis* the overall numbers were lower, although we observed a similar trend. For NCM, the total number of sequences that had more than 75% identity was 3670 and the same for CSM was 3347 (Figure 2B). This represented an increase >10% in NCM predictions over CSM predictions. A 1% increase in the median sequence identity was seen for the *B. subtilis* data set as was in *E. coli*. As we increased the threshold to 80%, we observed 200 proteins that were reverse translated by NCM yet only 114 by CSM. Moreover, the average identity of sequences reverse translated by either method was lower in *B. subtilis* than in *E. coli* K12. On the whole, these results suggest more random choice of codons in *B. subtilis* than in *E. coli* K12. These collective results underscore an important and fundamental distinction between the two groups of bacteria tested: increased randomness in the gram positive genome (*B. subtilis*) may be an indicator of its earlier evolutionary origin as compared to the gram negative (*E. coli*) genome (42).

In order to identify the CAI range within which NCM was effective, we compared the distribution of CAI values of genes whose ID_{NCM}/ID_{CSM} ratio was >1.01 with those

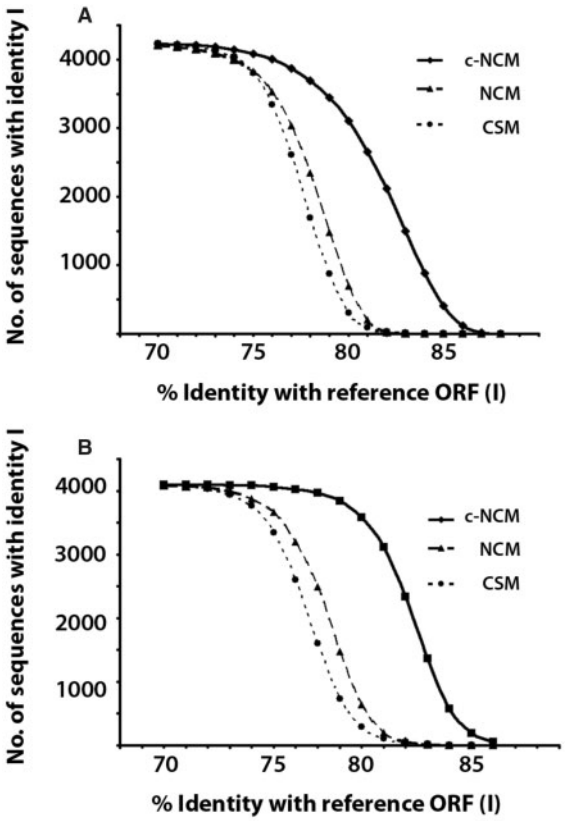

Figure 2. Percentage identity distribution of reverse translated ORF sequences in *E. coli* K12 and *B. subtilis* are shown in Panel **A** and **B**, respectively. This study compares the identities of reverse translations by CSM, NCM and c-NCM, revealing the distribution of percentage identities of reverse translated ORFs to the native ORFs. Two genomes, *E. coli* K12 and *B. subtilis*, are represented herein. Note that the NCM predictions are both qualitatively and quantitatively better and are also more numerous beyond 77% identity in both cases. This graph depicts the current limits of theoretical reverse-translation at ~85% for all the methods. The improvement of c-NCM over CSM and NCM, especially in regions of higher sequence identity, is clearly visible and significant (*F*-test with FDR correction; $P < 2.3 \times 10^{-44}$).

whose ID_{NCM}/ID_{CSM} ratio was <0.99 (KS test, alternative hypothesis = CAI distribution of NCM is lesser than that of CSM; $P < 1.0 \times 10^{-6}$). These results show that NCM performs significantly better than CSM in regions of low CAI.

Comparison of CSM and NCM on various phyla in the bacterial kingdom

In the previous section, we discussed and compared the results of reverse translation using CSM and NCM in two divergent bacterial species. Despite the phylogenetic distance between the two species, the both were eubacteria with moderate GC content. In order to show that the differences are real, we tested and compared the methods 28 different bacterial genomes, each representing a major clade in the bacterial kingdom as listed in KEGG (43). This list included one each of various groups like Archaebacteria, alpha-, beta-, gamma- and delta-proteobacteria, firmicutes, mollicutes, actinomyces, halo- and acido-bacteria, green sulfur and non-green sulfur bacteria, and cyanobacteria. The exhaustive list of organisms and a comparison of CSM and NCM in these genomes is tabulated in Table 2, and Supplementary Data lists the minima, maxima, median, first and third quartiles for these methods for all the genomes. From Table 2, it is evident that NCM outperforms CSM not only in genomes with moderate GC content but also in all major bacterial clades.

Improving the performance of NCM: the c-NCM

While NCM offered a better method to reverse-translate protein sequences, the overall improvement over CSM was apparent only at a higher sequence identity cut-off and for only a small fraction of the sequences. In order to improve the sensitivity of NCM, we developed a technique known as c-NCM, where the same protein was reverse-translated *n* times and a consensus was derived from the resultant sequence set. Our tests with a random set of 1000 sequences derived from *E. coli* K12 genome (Figure 3) demonstrated a drastic improvement from 1 cycle to 20 cycles. After 25 cycles, there was only a small improvement in prediction efficiency, which became insignificant beyond 50 cycles as compared to 100 cycles (KS test, alternative hypothesis = two-tailed: *P*-value = 0.2406). Moreover, our tests with a sample set of 100 sequences show that there is no significant improvement in sequence identity between 100 and 1000 cycles (data not shown). Hence we chose to use 50 cycles for subsequent c-NCM-based studies. The results of c-NCM are summarized along with those for CSM and NCM, in Figure 2A and 2B for *E. coli* K12 and *B. subtilis*, respectively. The average identity of reverse translated sequences increased by 4% with c-NCM when compared to the results from NCM. In summary, c-NCM reverse translated >75% of sequences with 80% identity or more while the percentage of sequences scoring the same with NCM was <20% in both genomes. This difference is highly significant ($P = 0$ for >85–90% ID and 2.3×10^{-44} for >80–85% ID, Fisher's test and FDR correction). These results revealed that c-NCM is an effective method for reverse translation of protein sequences based on genomic usage matrices, and also indicate that the performance of c-NCM was significantly better than both NCM and CSM. As was the case for CSM and NCM, we tested c-NCM on all the 30 genomes (Table 2). It could be seen that the performance of c-NCM was significantly different between both NCM and CSM for all phyla.

Apart from testing c-NCM on different genomes, we were also interested in analyzing the effects of consensus improvisation on the CSM method. Differences from the normal trend, if any, would allow us to discern genomes that have increased, or decreased randomness in their codon usage. On the same set of 30 genomes, we performed c-CSM (50 cycles) and compared the results with that of c-NCM using a Wilcoxon Rank Sum test. The results in Table 2 show that in 70% (21 of 30) of the tested genomes, c-NCM had a better performance than c-CSM. In five cases, the difference between the two methods was insignificant. These genomes were, *Aquifex aeolicus*

Table 2. Table comparing performance of CSM, NCM, c-CSM and c-NCM in 30 different clades of bacterial kingdom

Clade	Organism	Genome ID	CSM–cNCM	NCM–cNCM	cCSM–cNCM
Hyperthermophiles	*Aquifex_aeolicus*	NC_000117	2.2xE-16	2.2xE-16	0.2643
Bacteroides	*Bacteroides thetaiotaomicron*	NC_000908	2.2xE-16	2.2xE-16	2.2xE-16
Beta-proteobacteria	*Bordetella pertussis*	NC_000909	2.2xE-16	2.2xE-16	2.2xE-16
Delta-proteobacteria	*Myxococcus xanthus*	NC_000918	2.2xE-16	2.2xE-16	2.2xE-16
Epsilon-proteobacteria	*Campylobacter jejuni*	NC_000919	2.2xE-16	2.2xE-16	2.2xE-16
Alpha-proteobacteria	*Caulobacter crescentus*	NC_000962	2.2xE-16	2.2xE-16	2.2xE-16
Chlamydia	*Chlamydia trachomatis*	NC_000963	2.2xE-16	2.2xE-16	2.2xE-16[a]
Clostridia	*Clostridium acetobutylicum*	NC_001263	2.2xE-16	2.2xE-16	2.2xE-16
Green-nonsulfur	*Dehalococcoides ethenogenes*	NC_002162	2.2xE-16	2.2xE-16	5.168xE-08
Deinococcus	*Deinococcus radiodurans*	NC_002163	2.2xE-16	2.2xE-16	2.2xE-16
Fusobacteria	*Fusobacterium nucleatum*	NC_002516	2.2xE-16	2.2xE-16	1.567xE-08
Lactobacillales	*Lactobacillus acidophilus*	NC_002578	2.2xE-16	2.2xE-16	2.2xE-16
Alpha-rhizobacteria	*Mesorhizobium loti*	NC_002678	2.2xE-16	2.2xE-16	2.2xE-16
Euryarchaeota	*Methanococcus jannaschii*	NC_002696	2.2xE-16	2.2xE-16	0.2876
Euryarchaeota	*Methanopyrus kandleri*	NC_002754	2.2xE-16	2.2xE-16	2.936xE-10
Actinobacteria	*Mycobacterium bovis*	NC_002929	2.2xE-16	2.2xE-16	2.2xE-16
Actinobacteria	*Mycobacterium tuberculosis*	NC_002936	2.2xE-16	2.2xE-16	2.2xE-16
Mollicutes	*Mycoplasma genitalium*	NC_002945	2.2xE-16	2.2xE-16	2.2xE-16
Nanoarchaeota	*Nanoarchaeum equitans*	NC_003030	2.2xE-16	2.2xE-16	2.2xE-16[a]
Cyanobacteria	*Prochlorococcus marinus*	NC_003454	2.2xE-16	2.2xE-16	2.2xE-16[a]
Gamma-proteobacteria	*Pseudomonas aeruginosa*	NC_003551	2.2xE-16	2.2xE-16	0.01897
Alpha/Rickettsiae	*Rickettsia prowazekii*	NC_004663	2.2xE-16	2.2xE-16	2.2xE-16
Crenarchaeota	*Sulfolobus solfataricus*	NC_005072	2.2xE-16	2.2xE-16	0.761
Cyanobacteria	*Synechococcus elongatus*	NC_005213	2.2xE-16	2.2xE-16	0.0011
Euryarchaeota	*Thermoplasma acidophilum*	NC_006576	2.2xE-16	2.2xE-16	2.2xE-16
Spirochete	*Treponema pallidum*	NC_006814	2.2xE-16	2.2xE-16	2.2xE-16 [a]
Mollicutes	*Ureaplasma urealyticum*	NC_008009	2.2xE-16	2.2xE-16	0.0455
Acidobacteria	*Acidobacteria bacterium*	NC_008095	2.2xE-16	2.2xE-16	5.471xE-05
Magnetococcus	*Magnetococcus MC-1*	NC_008576	2.2xE-16	2.2xE-16	2.2xE-16

[a]In these cases, c-CSM performed significantly better than c-NCM.

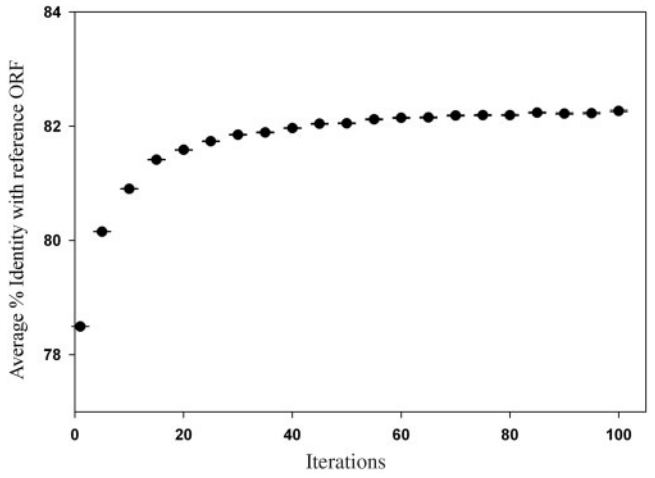

Figure 3. Standardization of iteration values for c-NCM. This figure illustrates the improvement in sensitivity as the number of iterations is increased in NCM. We performed c-NCM-based reverse translations for 1000 randomly chosen proteins using various iterations (5–100, in five steps) and compared the results with (**A**) predictions from NCM and (**B**) predictions from 100 iterations of c-NCM. It can be seen that the largest difference is between iteration values of 1 (NCM) and 50 (KS test, alternative = greater; $P = 2.2 \times 10^{-16}$). However, there is a small increase of sensitivity as the iterations are increased. The sensitivity difference was tested to 100 cycles and since there was no significant difference between 50 and 100 cycles (KS test, alternative = greater; $P = 0.2406$), we chose the threshold for c-NCM at 50 cycles.

(hyperthermophile), *Bordetella pertussis* (beta-proteobacteria), *Methanopyrus kandleri* (euryarchaeota), *Prochlorococcus marinus* (cyanobacteria) and *Acidobacteria bacterium* (acidobacteria). In four other cases, c-CSM performed significantly better than c-NCM: they were *Rickettsia prowazekii* (alpha-proteobacteria/Rickettsiae), *Clostridium acetobutylicum* (Clostridia), *Fusobacterium nucleatum* (Fusobacteria), and *Lactobacillus acidophilus* (Lactobacillales). These results, at least for *P. marinus* and *M. kandleri*, show that in archaeal and cyanobacterial genomes very little of tricodon usage information is carried over to the protein level.

Application of reverse translation to an external genome: *Salmonella typhi* CT18

In the previous sections, we demonstrated that the improvised method (c-NCM) performed significantly better than CSM and NCM. We hypothesized that NCM matrices created from a genome can be used for reverse translating protein sequences from a related genome. *S. typhi* CT18 is 67% identical to *E. coli* K12 genome at the DNA level, and hence was a good model system to test our hypothesis. Results from these comparisons showed significant differences between the prediction quality between CSM and NCM. Again, as was seen in 21 other genomes, the use of c-NCM improved prediction quality, with average identity beyond 80%. There was a very small difference in the average identities and the distribution between *S. typhi* CT18 (Figure 4) and *E. coli* K12 (Figure 2A). These observations confirmed

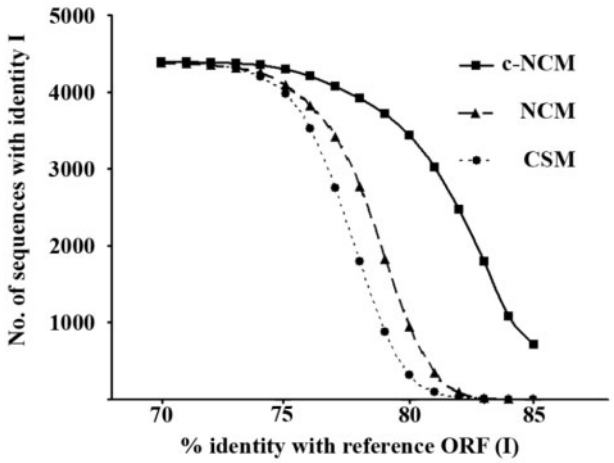

Figure 4. Reverse translation of *S. typhi* proteins using *E. coli* K12 matrices. *S. typhi* CT18 proteins were reverse translated using codon usage and NCM matrices of *E. coli* K12 genome. Analyses of identities with reference ORFs show that predictions using c-NCM are both qualitatively and quantitatively better than those using CSM (KS test: alternative = greater; $P < 2.2 \times 10^{-16}$). These results prove the applicability of c-NCM in cases where genome sequence data and NCM matrices are not available for the organism of interest.

that our method can be successfully applied to related genomes, suggesting increased fidelity in the design of degenerate primers for an organism whose gene sequence information is meager or non-existent. In such cases, the use of (c-)NCM matrices from a related organism is a viable alternative.

Throughout this work, we have concentrated on the applications of reverse translation in design of degenerate PCR. However, these studies also reveal the underlying logic of codon usage in genes in general, and such knowledge will be imperative in the design of synthetic genes to be used in artificial genetic systems and can also be used to adapt recombinant genes in a host specific manner.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work has been supported by grants from the National Institutes of Health: AI052017, AI065430 and AI060753 (to A.M.C.). Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health grant AI060753.

Conflict of interest statement. None declared.

REFERENCES

1. Burge,C., Campbell,A.M. and Karlin,S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.
2. Sivaraman,K., Seshasayee,A.S., Swaminathan,K., Muthukumaran,G. and Pennathur,G. (2005) Promoter addresses: revelations from oligonucleotide profiling applied to the *Escherichia coli* genome. *Theor. Biol. Med. Model*, **2**, 20.
3. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
4. Fedorov,A., Saxonov,S. and Gilbert,W. (2002) Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Res.*, **30**, 1192–1197.
5. Tork,S., Hatin,I., Rousset,J.P. and Fabret,C. (2004) The major 5′ determinant in stop codon read-through involves two adjacent adenines. *Nucleic Acids Res.*, **32**, 415–421.
6. Gutman,G.A. and Hatfield,G.W. (1989) Nonrandom utilization of codon pairs in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **86**, 3699–3703.
7. Boycheva,S., Chkodrov,G. and Ivanov,I. (2003) Codon pairs in the genome of Escherichia coli. *Bioinformatics*, **19**, 987–998.
8. Moura,G., Pinheiro,M., Arrais,J., Gomes,A.C., Carreto,L., Freitas,A., Oliveira,J.L. and Santos,M.A. (2007) Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS ONE*, **2**, e847.
9. Biro,J.C. (2004) Seven fundamental, unsolved questions in molecular biology. Cooperative storage and bi-directional transfer of biological information by nucleic acids and proteins: an alternative to ''central dogma''. *Med. Hypotheses*, **63**, 951–962.
10. Presnell,S.R. and Benner,S.A. (1988) The design of synthetic genes. *Nucleic Acids Res.*, **16**, 1693–1702.
11. Laging,M., Fartmann,B. and Kramer,W. (2001) Isolation of segments of homologous genes with only one conserved amino acid region via PCR. *Nucleic Acids Res.*, **29**, E8.
12. Blattner,F.R., Plunkett,G., 3rd, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of Escherichia coli K-12. *Science*, **277**, 1453–1474.
13. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al.* (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
14. Parkhill,J., Dougan,G., James,K.D., Thomson,N.R., Pickard,D., Wain,J., Churcher,C., Mungall,K.L., Bentley,S.D., Holden,M.T. *et al.* (2001) Complete genome sequence of a multiple drug resistant *Salmonella enterica serovar* Typhi CT18. *Nature*, **413**, 848–852.
15. Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M. *et al.* (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature*, **392**, 353–358.
16. Xu,J., Bjursell,M.K., Himrod,J., Deng,S., Carmichael,L.K., Chiang,H.C., Hooper,L.V. and Gordon,J.I. (2003) A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science*, **299**, 2074–2076.
17. Sebaihia,M., Preston,A., Maskell,D.J., Kuzmiak,H., Connell,T.D., King,N.D., Orndorff,P.E., Miyamoto,D.M., Thomson,N.R., Harris,D. *et al.* (2006) Comparison of the genome sequence of the poultry pathogen *Bordetella avium* with those of *B. bronchiseptica*, *B. pertussis*, and *B. parapertussis* reveals extensive diversity in surface structures associated with host interaction. *J. Bacteriol.*, **188**, 6002–6015.
18. Parkhill,J., Wren,B.W., Mungall,K., Ketley,J.M., Churcher,C., Basham,D., Chillingworth,T., Davies,R.M., Feltwell,T., Holroyd,S. *et al.* (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665–668.
19. Nierman,W.C., Feldblyum,T.V., Laub,M.T., Paulsen,I.T., Nelson,K.E., Eisen,J.A., Heidelberg,J.F., Alley,M.R., Ohta,N., Maddock,J.R. *et al.* (2001) Complete genome sequence of *Caulobacter crescentus*. *Proc. Natl Acad. Sci. USA*, **98**, 4136–4141.
20. Stephens,R.S., Kalman,S., Lammel,C., Fan,J., Marathe,R., Aravind,L., Mitchell,W., Olinger,L., Tatusov,R.L., Zhao,Q. *et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science*, **282**, 754–759.
21. Nolling,J., Breton,G., Omelchenko,M.V., Makarova,K.S., Zeng,Q., Gibson,R., Lee,H.M., Dubois,J., Qiu,D., Hitti,J. *et al.* (2001) Genome sequence and comparative analysis of the

solvent-producing bacterium *Clostridium acetobutylicum*. *J. Bacteriol.*, **183**, 4823–4838.
22. Nonaka,H., Keresztes,G., Shinoda,Y., Ikenaga,Y., Abe,M., Naito,K., Inatomi,K., Furukawa,K., Inui,M. and Yukawa,H. (2006) Complete genome sequence of the dehalorespiring bacterium *Desulfitobacterium hafniense* Y51 and comparison with *Dehalococcoides ethenogenes* 195. *J. Bacteriol.*, **188**, 2262–2274.
23. White,O., Eisen,J.A., Heidelberg,J.F., Hickey,E.K., Peterson,J.D., Dodson,R.J., Haft,D.H., Gwinn,M.L., Nelson,W.C., Richardson,D.L. *et al.* (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science*, **286**, 1571–1577.
24. Kapatral,V., Anderson,I., Ivanova,N., Reznik,G., Los,T., Lykidis,A., Bhattacharyya,A., Bartman,A., Gardner,W., Grechkin,G. *et al.* (2002) Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586. *J. Bacteriol.*, **184**, 2005–2018.
25. Altermann,E., Russell,W.M., Azcarate-Peril,M.A., Barrangou,R., Buck,B.L., McAuliffe,O., Souther,N., Dobson,A., Duong,T., Callanan,M. *et al.* (2005) Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM. *Proc. Natl Acad. Sci. USA*, **102**, 3906–3912.
26. Kaneko,T., Nakamura,Y., Sato,S., Asamizu,E., Kato,T., Sasamoto,S., Watanabe,A., Idesawa,K., Ishikawa,A., Kawashima,K. *et al.* (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Res.*, **7**, 331–338.
27. Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D. *et al.* (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1073.
28. Slesarev,A.I., Mezhevaya,K.V., Makarova,K.S., Polushin,N.N., Shcherbinina,O.V., Shakhova,V.V., Belova,G.I., Aravind,L., Natale,D.A., Rogozin,I.B. *et al.* (2002) The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proc. Natl Acad. Sci. USA*, **99**, 4644–4649.
29. Garnier,T., Eiglmeier,K., Camus,J.C., Medina,N., Mansoor,H., Pryor,M., Duthoy,S., Grondin,S., Lacroix,C., Monsempe,C. *et al.* (2003) The complete genome sequence of *Mycobacterium bovis*. *Proc. Natl Acad. Sci. USA*, **100**, 7877–7882.
30. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E., 3rd *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.
31. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
32. Goldman,B.S., Nierman,W.C., Kaiser,D., Slater,S.C., Durkin,A.S., Eisen,J.A., Ronning,C.M., Barbazuk,W.B., Blanchard,M., Field,C. *et al.* (2006) Evolution of sensory complexity recorded in a myxobacterial genome. *Proc. Natl Acad. Sci. USA*, **103**, 15200–15205.
33. Waters,E., Hohn,M.J., Ahel,I., Graham,D.E., Adams,M.D., Barnstead,M., Beeson,K.Y., Bibbs,L., Bolanos,R., Keller,M. *et al.* (2003) The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proc. Natl Acad. Sci. USA*, **100**, 12984–12988.
34. Dufresne,A., Salanoubat,M., Partensky,F., Artiguenave,F., Axmann,I.M., Barbe,V., Duprat,S., Galperin,M.Y., Koonin,E.V., Le Gall,F. *et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl Acad. Sci. USA*, **100**, 10020–10025.
35. Stover,C.K., Pham,X.Q., Erwin,A.L., Mizoguchi,S.D., Warrener,P., Hickey,M.J., Brinkman,F.S., Hufnagle,W.O., Kowalik,D.J., Lagrou,M. *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**, 959–964.
36. Andersson,S.G., Zomorodipour,A., Andersson,J.O., Sicheritz-Ponten,T., Alsmark,U.C., Podowski,R.M., Naslund,A.K., Eriksson,A.S., Winkler,H.H. and Kurland,C.G. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**, 133–140.
37. She,Q., Singh,R.K., Confalonieri,F., Zivanovic,Y., Allard,G., Awayez,M.J., Chan-Weiher,C.C., Clausen,I.G., Curtis,B.A., De Moors,A. *et al.* (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl Acad. Sci. USA*, **98**, 7835–7840.
38. Sugita,C., Ogata,K., Shikata,M., Jikuya,H., Takano,J., Furumichi,M., Kanehisa,M., Omata,T., Sugiura,M. and Sugita,M. (2007) Complete nucleotide sequence of the freshwater unicellular cyanobacterium *Synechococcus elongatus* PCC 6301 chromosome: gene content and organization. *Photosynth. Res.*, **93**, 55–67.
39. Ruepp,A., Graml,W., Santos-Martinez,M.L., Koretke,K.K., Volker,C., Mewes,H.W., Frishman,D., Stocker,S., Lupas,A.N. and Baumeister,W. (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature*, **407**, 508–513.
40. Glass,J.I., Lefkowitz,E.J., Glass,J.S., Heiner,C.R., Chen,E.Y. and Cassell,G.H. (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature*, **407**, 757–762.
41. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
42. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
43. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–357.