N. R. McEwan · D. Gatherer

The mutational-response index and codon bias in genes from a *Frankia nif* operon

Received: 8 August 1997 / Accepted: 19 September 1997

Abstract The mutational response index (MRI) and measurements of codon bias were determined in three characterised genes, and two open reading frames of unknown function, from the *Frankia nif* operon, which encodes genes for nitrogen fixation. The merits of the different systems of measuring codon usage are discussed in the light of the results, as are the applicability of these techniques to the assessment of the translational function of putative open reading frames.

Key words *Frankia* · Codon usage · Codon bias · ORFs · Mutation pressure

Introduction

Frankia is a filamentous actinomycete which can fix nitrogen in symbiotic association with the roots of over 200 species of dicotyledonous plants, covering 25 genera and eight families. In keeping with other actinomycetes, it has a very high genomic GC content – normally in the order of 68–72%. In most organisms with either a very high or low GC genomic content there is, respectively, a high or low proportion of G or C nucleotides in the third position of open reading frames, resulting in a high level of codon bias. This is defined as codon bias due to mutational pressure (Sueoka 1988). Additionally, highly expressed genes may be under selection for codon-usage patterns which permit rapid translation. This is defined as codon bias due to translational selection (Gouy and Gautier 1982). Both of these pressures may be acting on genes simultaneously.

A number of different techniques have been developed for quantifying the relative usage of different codons including: optimal codon-anticodon energy (P2; Gouy and Gautier 1982), effective number of codons (Nc; Wright 1990), and the intrinsic codon deviation index (ICDI; Freire-Picos et al. 1994). In addition, the extent to which a gene responds to mutational pressure may be measured by the mutational response index (MRI; Gatherer and McEwan 1997). Details of these measurements are listed below in the Materials and methods section.

Bioinformatics and genome projects generate large amounts of sequence data which may contain open reading frames (ORFs). However, these ORFs may not necessarily represent expressed genes. Here we consider how codon-usage indices and MRI may assist in determining the likelihood that an ORF is an expressed gene.

Materials and methods

Sequences

All sequence information used in this work was obtained from the GenBank database (URL = http://www2.ncbi.nlm.nih.gov/cgi-bin/genbank). The accession number for the operon used was U53363 (Oh et al. 1996). This sequence contains the *nifV*, *nifH* and *nifD* genes and also contains two undesignated putative reading frames 5′ to the other three genes.

Computer programs

Sequence information was obtained from GenBank using Netscape Navigator. Analysis of sequences was performed using Microsoft

Communicated by Alföldi

Neil R. McEwan (✉)
Land Resources Department, Scottish Agricultural College,
Craibstone Estate, Aberdeen AB21 9TN, Scotland
E-mail: nrm@rri.sari.ac.uk

Derek Gatherer
School of Biomolecular Sciences, John Moores University,
Byrom Street, Liverpool L3 3AF, England

Word Version 6 and Microsoft Excel Version 5. Blast searches to identify similar sequences already lodged in databases were performed using the internet, (URL = http://genome.eerie.fr/bin/blast-guess.cgi).

Equations

The effective codon number (N_c) was calculated as in Wright (1990). Firstly F caret ($F^\wedge$) is calculated for each synonymous group:

$$F^\wedge = \left[\left(n_{aa} \sum_{j=1}^{j} p^2\right) - 1\right] \Big/ (n_{aa} - 1), \tag{1}$$

where p is the proportion of usage of a codon i within its synonymous group of size j, and n_{aa} the total usage of that synonymous group.

$F^{\wedge av.}$ the average $F^\wedge$ for synonymous groups of the same size (i.e. for 2, 4 and 6), and N_c is calculated as follows:

$$N_c = 2 + 9/F^{\wedge av2} + 1/F^{\wedge 3} + 5/F^{\wedge av4} + 3/F^{\wedge av6}. \tag{2}$$

N_c provides an alternative to chi-square-based measures of codon bias, and satisfies the objection of Rees (1994), which points out that where expected values are less than 5 chi-square tests are not always appropriate. This is often the case where codon bias is being considered.

Translational selection (P2) is calculated following Gouy and Gautier (1982):

$$P2 = (WWC + SSU)/(WWY + SSY), \tag{3}$$

where $W = A$ or U, $S = C$ or G and $Y = C$ or U.

P2 measures the efficiency of codon-anticodon interaction, and provides an indication of translational efficiency when information on the preferred codon set is unavailable.

The intrinsic codon deviation index (ICDI) is calculated according to Freire-Picos et al. (1994), but first requires calculation of the relative synonymous codon usage value (RSCU) value, as determined by Sharp and Li (1987):

$$RSCU_{ij} = X_{ij} \Big/ \left(1/n_i \sum_{j=1}^{n_i} X_{ij}\right), \tag{4}$$

where x_{ij} is the number of occurrences of the j^{th} codon for the i^{th} amino acid, and n_i is the size of the synonymous group for the i^{th} amino acid (i.e. 2, 3, 4 or 6). ICDI is calculated as follows. Firstly S_k is calculated for each synonymous group as:

$$S_k = \sum_{i=1}^{k} (RSCU_i - 1)^2/k(k-1), \tag{5}$$

where $RSCU_i$ is calculated for the i^{th} codon and k is either 2, 3, 4 or 6 depending on the degeneracy of the synonymous group. Then the S_k values are combined as follows:

$$ICDI = \left(\sum S_2 + S_3 + \sum S_4 + \sum S_6\right) \Big/ 18. \tag{6}$$

The mutational response index (MRI) was calculated according to Gatherer and McEwan (1997) and is the difference between scaled chi-square (SCS) values and corrected scaled chi-square values (CSCS). Both SCS and CSCS are calculated based on the standard chi-square formula:

$$\text{standard } \chi^2 = \sum_{i=1}^{18} \sum_{j=1}^{f_i} [e_i - o_{ij})^2/e_i], \tag{7}$$

where o_{ij} is the number of occurrences of the j^{th} codon for the i^{th} amino acid, e_i is the expected usage of the j^{th} codon under conditions of equal synonymity, f_i is the degeneracy of codons for the i^{th} amino acid. Those codons with no capacity for degeneracy, Met and Trp, were excluded from the chi-square calculations, as were stop codons, since only one can occur per reading frame.

Scaled chi-square (Shields and Sharp, 1987) is calculated as follows:

$$\text{scaled } \chi^2 = \left\{\sum_{i=1}^{18} \sum_{j=1}^{f_i} [e_i - o_{ij})^2/e_i]/(2 \times N_{aa})\right\}, \tag{8}$$

where the variables are as defined in (7) above with the addition of N_{aa}, the total number of amino acids (excluding Met and Trp residues). Because chi-square is additive, a slight bias exists when comparing genes of different lengths. Using a scaled chi-square removes this bias.

CSCS (Mather and Tuli 1991) is calculated as above (equation 8), except that an adjustment to e_i is made as follows:

$$e_i = \left(\sum_{j=1}^{f_i} O_{ij}\right) \times \left(N_j \Big/ \sum_{j=1}^{f_i} N_j\right) \times F_6, \tag{9}$$

where o_{ij} is defined as above (equation 7), and N_j is the number of j^{th} nucleotides at position 3 (i.e. the frequency of A, C, G or T at XXN). F_6 is 1 unless the codon degeneracy is 6 (i.e. for Arg, Leu and Ser). For Arg and Leu F_6 is calculated as follows:

$$F_6 = \left(\sum_{j=1}^{g_i} N_j\right) \Big/ \left(\sum_{j=1}^{6} N_j\right), \tag{10}$$

where g_j is 2 for the doublets and 4 for the quartets. For Ser, F_6 is as above but the value of j changes to the frequency of UCX and AGX in the gene. CSCS calculates the component of codon bias which is independent of the nucleotide content of the gene.

Results and discussion

The values of intrinsic codon deviation index (ICDI), translational selection (P2), effective codon number (N_c), mutational response index (MRI), scaled Chi square (SCS) and corrected scaled Chi square (CSCS) are shown in Table 1, together with the *GC3s*, *GC2s* and *GC1s* values (frequency of either G or C in the third, second or first nucleotide position of a codon). Derivation of values for ORF B is complicated by the fact that it is only 76 codons in length and its putative translated product lacks Glu, Phe, Asn and Lys. N_c is derivable for any ORF of greater than 61 codons in length, but is not recommended for those of less than 100 codons (Wright 1990). ICDI can be calculated in two ways in instances where amino acids are unrepresented. However, it is not clear which is the most appropriate (see footnotes to Table 1). Likewise, SCS and the other chi-square-based calculations do not apply to this ORF owing to the difficulty in estimating expected values.

ICDI, SCS and CSCS increase with increasing levels of deviation from synonymous codon usage whereas N_c, as a measure of effective codon number, decreases with greater codon bias. P2 reveals bias in favour of optimal codon-anticodon energy, and by inference translational selection, when its value is greater than 0.5. By this criterion, *nifH* is subject to weak translational selection.

Freire-Picos et al. (1994) defined moderately expressed genes as having ICDI levels between 0.3 and

Table 1 ICDI, P2, N_c, MRI, CSCS, SCS, *GC3s*, *GC2s* and *GC1s* values for genes or ORFs in this operon

Gene/ORF	P2	N_c	ICDI	MRI	CSCS	SCS	*GC3s*	*GC2s*	*GC1s*
nif V	0.328	30.1	0.314	0.338	0.092	0.430	0.881	0.520	0.736
nif H	0.505	25.6	0.345	0.457	0.196	0.653	0.955	0.446	0.634
nif D	0.441	26.9	0.345	0.472	0.116	0.588	0.962	0.515	0.564
ORF A	0.358	44.5	0.262	0.096	0.065	0.161	0.717	0.600	0.711
ORF B	0.348	39.2[a]	0.329[b]	N/A[c]	N/A[c]	N/A[c]	0.737	0.697	0.697

[a] N_c is adjusted following Wright (1990), to take account of the fact that four amino acids are not represented in ORF B

[b] Unlike N_c, there is no adjustment recommended for ICDI. The figure given here is based on an assumption that those amino acids which are not represented have a value of zero. However, strict adherence to the calculation of Freire-Picos et al. (1994, See Materials and methods) gives a value of 1 to each codon of unrepresented amino acids. This has the effect of inflating ICDI to around 0.5

[c] SCS and CSCS cannot be calculated for ORF B as the expected value of the codons in the unrepresented amino acids is zero. This places zero in the denominator of the chi-square calculation and gives an answer of infinity. MRI consequently cannot be calculated as it requires values for both SCS and CSCS

0.5. All of the genes and ORFs studied here fall into the lower half of this range and ORF A has ICDI of less than 0.3. Therefore, there appears to be little reason for expecting high levels of translational selection in the *nif* operon, despite its inducibility under conditions of nitrogen-starvation.

Muto and Osawa (1987) reviewed the correspondence between genome GC content and positional nucleotide content in 11 prokaryotes, including the actinomycete *Streptomyces vanaceus*, and predicted that a genome of approximately 70% GC content should display 85–90% GC in the 3rd position, 70% in the 1st position, and 45% in the 2nd position. This is a reflection of the fact that the third nucleotide position has greater scope for variability, other nucleotides within a codon being subjected to greater conservational pressure due to the limitations of synonymity. Of the above genes and ORFs, *nifV* shows the closest correspondence to Muto and Osawa's predictions. In both *nifH* and *nifD GC3s* is elevated above predictions, and these two genes also display lower than expected values of *GC1s*. These deviations are reflected in their greater codon bias as measured by SCS and ICDI. However, the *GC3s* values of the two undefined ORFs are much lower than for the other three genes, being closer to the genomic GC content. These patterns are well outside of Muto and Osawa's predicted ranges for position-specific nucleotide content, and strongly suggest that ORF A and ORF B are unexpressed.

MRI, which defines that component of codon bias which is produced by mutational pressure, is also elevated in the *nif* genes compared to the ORFs. Expressed sequences appear to 'over-respond' to mutation pressure with elevated *GC3s*, while maintaining low levels of *GC2s*. In species where genomic GC content is high, such as *Frankia*, ORFs with elevated MRI are good candidates for expressed sequences.

Blast searches were performed to check ORFs A and B for similarity to any other sequences in the various databases. At the DNA level, both ORFs failed to show significant similarities (even when the cut-off value was set to $P = 0.1$) to any other DNA sequences. As an additional check, a blast search was performed at the protein level. This time at a significance level of $P = 0.05$, ORF B continued to show no significance to any other protein described. ORF A did show similarity to one other protein ($P = 0.0005$). This corresponded to a 33% homology to a 53-residue sequence of a particularly glycine-rich stretch of a human *twist*-related protein. It is primarily the large number of glycine residues which generated this homology.

References

Freire-Picos MA, Gonzalez-Siso MI, Rodriguez-Belmonte E, Rodriguez-Torres AM, Ramil E, Cerdan ME (1994) Codon usage in *Kluyveromyces lactis* and in yeast cytochrome *c*-encoding genes. Gene 139 : 43–49

Gatherer D, McEwan NR (1997) Small regions of preferential codon usage and their effect on overall codon bias – the case of the *plp* gene. Bioch Mol Biol Int 43 : 107–114

Gouy M, Gautier C (1982) Codon usage in bacteria – correlation with gene expressivity. Nucleic Acids Res 10 : 7055–7074

Mather M, Tuli R (1991) Analysis of codon usage in genes for nitrogen fixation from phylogentically diverse diazotrophs. J Mol Evol 32 : 364–373

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84 : 166–169

Oh B, Twigg P, Hong J, Mullin B, An CS (1996) *nifV* is contiguous to *nifHDK* in *Frankia* strain FaC1. GenBank database direct submission U53363

Rees DC (1994) Essential statistics for medical practice. Chapman and Hall, London

Sharp PM, 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

Shields DC, Sharp PM (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational pressure. Nucleic Acids Res 15 : 8023–8040

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85 : 2653–2657

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87 : 23–29