Bioinformatics, 31(20), 2015, 3254–3261
doi: 10.1093/bioinformatics/btv339
Advance Access Publication Date: 4 June 2015
Original Paper

OXFORD

Genome analysis

Quantitative frame analysis and the annotation of GC-rich (and other) prokaryotic genomes. An application to *Anaeromyxobacter dehalogenans*

Steve Oden[1,2] and Luciano Brocchieri[1,2,*]

[1]Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL 32610, USA and [2]Genetics Institute, University of Florida, Gainesville, FL 32610, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on January 9, 2015; revised on May 12, 2015; accepted on May 28, 2015

Abstract

Motivation: Graphical representations of contrasts in GC usage among codon frame positions (frame analysis) provide evidence of genes missing from the annotations of prokaryotic genomes of high GC content but the qualitative approach of visual frame analysis prevents its applicability on a genomic scale.

Results: We developed two quantitative methods for the identification and statistical characterization in sequence regions of three-base periodicity (hits) associated with open reading frame structures. The methods were implemented in the N-Profile Analysis Computational Tool (NPACT), which highlights in graphical representations inconsistencies between newly identified ORFs and pre-existing annotations of coding-regions. We applied the NPACT procedures to two recently annotated strains of the deltaproteobacterium *Anaeromyxobacter dehalogenans*, identifying in both genomes numerous conserved ORFs not included in the published annotation of coding regions.

Availability and implementation: NPACT is available as a web-based service and for download at http://genome.ufl.edu/npact.

Contact: lucianob@ufl.edu

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

1 Introduction

Accurate genome annotations are fundamental to a great variety of sequence analyses. Annotations of prokaryotic genomes are based on computational gene prediction methods that, using information from sequence composition (Delcher *et al.*, 2007; Hyatt *et al.*, 2010; Lukashin and Borodovsky, 1998; Salzberg *et al.*, 1998) and from sequence signals (e.g. ribosome binding site) (Hyatt *et al.*, 2010; Larsen and Krogh, 2003), have achieved remarkable levels of sensitivity (>99%). The availability of large databases of gene sequences also facilitates identification of coding regions using information from sequence similarity. Nevertheless, a multitude of ORFs conserved across distantly related genomes and not reported in published genome annotations have been recently identified (Warren *et al.*, 2010). Despite the high sensitivity of gene prediction methods, certain genes may still be difficult to predict, for example, genes of very short length or genes encoded by heterologous DNA. True genes may also be excluded from the annotation by over-conservative criteria of robustness of prediction across methods or of evolutionary conservation. For example, although Glimmer3.0 (Delcher *et al.*, 2007) is often considered to have lower specificity than other predictors (e.g. GeneMark HMM or Prodigal), many genes predicted by Glimmer3.0 and not included in genome annotations have been confirmed in a recent analysis, based on evolutionary conservation and functional characterization (Wood *et al.*, 2012). Bioinformatics tools (Aziz *et al.*, 2008; Kumar *et al.*, 2011; Lee *et al.*, 2013; Lewis *et al.*, 2002; Stewart *et al.*, 2009; Vallenet *et al.*, 2013; Yu *et al.*, 2008) and procedural standards (Angiuoli *et al.*, 2008; Madupu *et al.*, 2010) have been developed to facilitate and

standardize the annotation process integrating resources for sequence annotation.

Coding regions of high GC content (GC > 55–60%) in most cases can be identified by their three-base periodicity in GC content, using the method of frame-analysis (Bibb *et al.*, 1984). In frame analysis, the GC content of three subsequences composed of every third nucleotide starting from position 1, 2 or 3 of the original sequence is computed within a moving window and represented as three 'S-profiles' (Brocchieri *et al.*, 2005). Coding regions can then be visually recognized by the typical contrasts that they induce among profiles (Supplementary Fig. S1). Frame analysis has been implemented in the FramePlot software (Ishikawa and Hotta, 1999), a web-based tool for visualization of S-profiles, positions of stop codons, and potential start of translation codons, which has often been utilized for the characterization of specific gene families (e.g. Goranovic *et al.*, 2012; Huang *et al.*, 2011; Sherwood *et al.*, 2013).

Anecdotal observations in genomes of high GC content (about one-third of all sequenced genomes) indicate that many genes not listed in annotations can be identified by visual frame analysis. Genome-wide frame analysis has been made accessible with the release of recent versions of FramePlot and through the Artemis genome browser and annotation tool (Rutherford *et al.*, 2000). However, its applicability on a genomic scale is still impractical, due to its qualitative nature. To address these limitations, we developed procedures for the quantitative assessment and statistical characterization of sequence segments with three-base-periodicity ('quantitative frame analysis') and for their convenient genome-wide representation. We implemented these procedures in the newly developed N-Profile Analysis Computational Tool (NPACT), by which ORFs with significant periodicities are identified and compared with pre-existing annotations in graphical representations showing the positional relations of newly identified ORFs with three-base periodic sequence segments, compositional profiles, and pre-annotated coding regions. We evaluated the potential of quantitative frame analysis to identify genes missed from annotations, and analyzed with NPACT the genome sequences and annotations of two strains of *Anaeromyxobacter dehalogenans*, a deltaproteobacterium of interest for a variety of bio-remediation applications (He and Sanford, 2003; Marshall *et al.*, 2009; Sanford *et al.*, 2007). We identified in the two genomes several ORFs with compositional periodicity that were excluded from the published annotations. Upon further analysis, we found in most cases that their coding capacity was corroborated by evolutionary conservation, functional characterization and computational gene predictions.

2 Quantitative frame analysis

We implemented quantitative frame analysis in the NPACT application. NPACT first identifies sequence segments with statistically significant compositional three-base periodicity that are associated with 'reading frames', i.e. with sequences of trinucleotides uninterrupted by 'stop codons'. To extend applicability of frame analysis to sequences of any composition, we extended the analysis of compositional three-periodicities from GC content to individual nucleotides. Periodicities are identified using two newly developed tests: the first test identifies three-base periodicities that are *expected* in coding sequences (Besemer and Borodovsky, 1999) of a given *local* composition; the second test identifies sequence regions with three-base periodicities of *any type*. ORFs of minimum length associated with segments of compositional periodicity are then identified when present, searching for the start-of-translation codon closest to the 5′-end of the region of periodicity. All ORFs identified by three-base periodicity are compared by NPACT to a set of predicted genes or to gene annotations provided by the user, distinguishing three-base periodicities that support the annotated genes from those that identify discrepancies and potential genes not included in the provided set of predictions. NPACT produces graphical representations that allow genome-wide uninterrupted visual comparison of compositional profiles, pre-annotated genes and sequence segments of three-base periodicity with 'Newly Identified ORFs', enabling frame analysis on a genomic scale.

2.1 The H-test: high-scoring segments of coding-like composition

To identify three-base periodicities associated with coding regions, we considered all 'reading frames' of the input sequence, i.e. continuous sequences of trinucleotides not interrupted by in-frame stop codons (TAA, TGA or TAG) that may contain coding regions. Each reading-frame was also used to identify the *local* composition of the sequence, based on which the nucleotide composition of the three codon positions can be predicted. Specifically, the expected frequency $E(\pi_{X_i})$ of nucleotide type X (= A, C, G, T) at codon position i (= 0, 1, 2) can be estimated by linear regression over its overall nucleotide frequency π_X from collections of known genes (Fig. 1A). To identify within a reading frame sequence segments of *any* length with the three-base periodicities expected in coding regions, we derived an implementation of the 'high-scoring segment' approach (Karlin and Altschul, 1990) based on cumulating, along the sequence, scores appropriate for identifying segments with expected phase-specific nucleotide frequencies, such as log-odds-ratio scores $S(X_i) = \ln [E(\pi_{X_i})/\pi_X]$, assigned to each nucleotide and codon position (Fig. 1B).

However, the absence of stop codons from reading frames generates dependencies between nucleotides in different codon positions (Supplementary Tables S3 and S4) and affects the probability with which nucleotides associate to form 'codons'. To account for these dependencies, we built instead log-odds-ratio scores for codons, based on their expected frequencies in 'pseudo-coding' and 'random' sequences, generated, respectively, drawing nucleotides with appropriate phase-specific probabilities $\{p_{X_i}\}$, or overall probabilities $\{p_X\}$ and removing stop codons. Given a specific reading frame, these probabilities can be calculated so that after removing stop codons, the expected nucleotide frequencies of the randomly generated sequences are the same as the overall frequencies observed in the reading frame, and in the pseudo-coding model, as the phase-specific frequencies expected in a corresponding coding region (Supplementary Methods). If $p^{(0)}_{stop}$ and $p^{(1)}_{stop}$ are the probabilities with which stop codons are generated in the random and pseudo-coding models, respectively, based on the ratio of expected frequencies, the log-odds-ratio score for codon XYZ is (Fig. 1C):

$$Score(XYZ|\pi) = \ln \frac{p_{X_1} p_{Y_2} p_{Z_3} (1 - p^{(0)}_{stop})}{p_X p_Y p_Z (1 - p^{(1)}_{stop})} \quad (1)$$

By summing codon scores along a reading-frame sequence, segments of any length with statistically significant cumulative score and expected phase-specific nucleotide composition can be identified, as described in Supplementary Figure S2. We evaluated cumulative-score values at different levels of statistical significance for 23 426 different compositional states, covering nucleotide frequencies from 0.0 to 1.0 at intervals $\Delta\pi_X = 0.02$, and sequence lengths doubled from 150 to 9600 nt, each from samples of 10 000

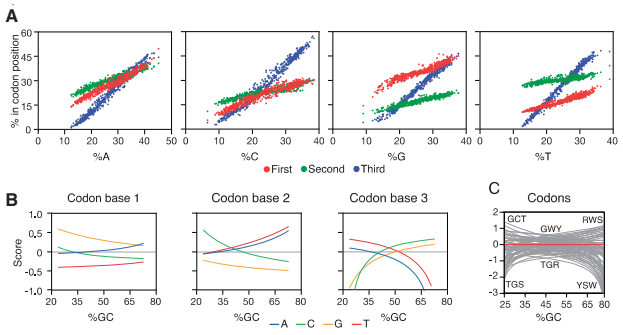

Fig. 1. Associations of nucleotide-type usages with codon position. (**A**) Base usages at the three codon positions of genes in relation to the overall usage of the same base. Each point represents the average usage in the corresponding codon position from the collection of all coding sequences annotated in one genome. (**B**) Log-odds-ratio scores associated to each base at the three codon base positions as a function of GC content. (**C**) Corresponding log-odds-ratio scores for the 61 codons with the highest- and lowest-scoring codon types indicated for low, intermediate and high GC content (W = AT, S = CG, R = AG, Y = CT)

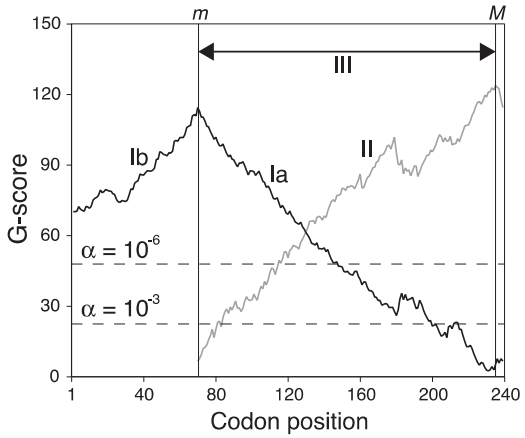

Fig. 2. An example of the procedure used to identify 'G-type' significant hits of any length within a reading frame of 240 trinucleotides containing a coding region starting at position *m*. Ia and Ib are G-scores assigned to codon positions *i*, with $1 \le i \le 240$, corresponding to G-values calculated over the intervals [*i*,240], with maximum G-score G_m at position *m*. II are G-scores assigned to codon positions *j* calculated over the intervals [*m*,*j*], with $m \le j \le 240$ and maximum G-score G_M at position *M*. III is the G-type hit identified by the two positions of maximum G-score. Threshold values of G_M corresponding to two significance levels α are indicated by dashed lines

randomly generated sequences (Supplementary Methods). We found that threshold-score values were linearly related to the log-log of the length of the sequence (Supplementary Fig. S3). Based on this relation, threshold scores for any length L and compositional state k, can be obtained from pre-calculated linear-regression coefficients (a_k,b_k) specific to each state k as $S*(L|k) = a_k \ln\ln L + b_k$ We refer to the sequence regions identified by this approach as 'H-type hits' or 'H-hits', and to the corresponding test as the 'H-test'.

Because high-scoring segments could be induced by genes encoded in alternative frames, we also scored each significant H-hit against all alternative frames substituting in log-odds-ratio scores the frequencies expected in a random model with those expected in alternative coding frames. If the frame of a high-scoring segment is the identity permutation $\mathcal{I}$, and alternative frames are identified by alternative permutations $\mathcal{P}$ of codon frame positions, scores relative to $\mathcal{P}$ are obtained as:

$$Score(XYZ|\pi,\mathcal{P}) = \ln \frac{p_{X_1} p_{Y_2} p_{Z_3} (1 - p^{(\mathcal{P})}_{stop})}{p_{X_{P(1)}} p_{Y_{P(2)}} p_{Z_{P(3)}} (1 - p^{(\mathcal{I})}_{stop})}, \qquad (2)$$

where $p^{(\mathcal{I})}_{stop}$ and $p^{(\mathcal{P})}_{stop}$ are the probabilities of stop codons under the pseudo-coding model for the corresponding permutations. Cumulating these scores along a H-hit sequence can result in a new, potentially shorter high-scoring segment. This is in turn re-scored relative to other permutations (Supplementary Fig. S2) until all five alternative permutations are tested. The final highest-scoring segment is retained only if it is still significant relative to the random sequence.

2.2 The G-test: significant nucleotide association with codon positions

To identify within a reading-frame three-base periodicities of any type, we propose a procedure based on the G statistics (Sokal and Rohlf, 1994), which is often used as an alternative to the χ^2 for testing associations. In our implementation, we used values of the G statistics (G-values) to score association of nucleotides with codon positions, as illustrated in Figure 2 and detailed in Supplementary Methods. Briefly, in a reading frame of L codons, a G-value is calculated over the interval $[i,L]$ $(1 \le i \le L)$ and is assigned to codon position i as its 'G-score'. The codon position $i = m$ of maximum G-score identifies a sequence interval $[m,L]$. Within this interval, G-values are calculated over intervals $[m,j]$ and assigned to codon positions j $(m \le j \le L)$. Position $j = M$ $(M \ge m)$ of maximum G-score G^*_M identifies with position m the sequence segment $[m,M]$, which is characterized as 'high-scoring' if G_M exceeds a threshold score G^*_M of significant probability. We identified thresholds of significance for the G_M statistics for a large number of compositions and different sequence lengths based on large samples of randomly generated sequences, as described for the H-test. As for the H statistics, we found that threshold values of G_M depend on sequence composition and are linearly related to the log-log-length of the sequence (Supplementary Fig. S4) and can similarly be inferred for any sequence length and composition. We refer to this test of association as the 'G-test', and to the high-scoring segments identified by the G-test as 'G-type hits' or 'G-hits'.

Because three-base periodicities identified as G-hits do not provide information on the reading frame and coding strand of a potential gene, G-hits were re-scored, as for the H-hits, by H-scores against coding in alternative frames (see earlier), and retained only if the G-value of the final high-scoring segment scored significantly by the G-test.

2.3 Three-base periodicities, low-complexity and newly identified ORFs

Using the newly developed tests, sequence segments with three-base periodicity can be identified within all reading frames of the direct and complementary strand of an input sequence. In our implementation, we first applied the H-test, and then applied the G-test to regions not covered by H-hits. We then evaluated if three-periodicities could be associated with regions of low complexity, which could generate significant cumulative scores if composed of trinucleotides with positive score. We calculated Shannon entropy H (Shannon, 1948) for each hit, based on its codon frequencies, and normalized it to the interval [0,1] by expressing it in log(61) units (61 being the

size of the 'codon alphabet'). We defined repetitive elements as sequence segments with $H < 0.4$, based on the distribution of H among genes of a large set of characterized coding regions (data not shown).

The association of hits with ORF structures was determined looking for a potential start of translation codon upstream of the 5′-end of a hit. If no canonical start codon (AUG or GUG) was present, alternative (rare) start codons were searched for in the order of preference UUG, CUG and AUU. ORFs were ranked by the score of the associated hit, and each hit was also characterized by an E-value, accounting for multiple testing (Supplementary Methods).

2.4 NPACT and the identification of genes missed from annotations

We implemented the described procedures in the web-based bioinformatics tool NPACT. NPACT compares ORFs identified by three-base periodicity with annotated or predicted coding regions provided by the user, identifying all ORFs not included in the set of pre-annotated genes, and characterizing pre-annotated genes as confirmed by significant periodicities, non-supported (if no significant hits are identified in frame within the annotated gene), contradicted (if hits superimposed to the gene are associated with an ORF in an alternative frame) or modified (if periodicities suggest a different start of translation for the gene). Furthermore, NPACT provides a convenient graphical representation, in which all ORFs identified in annotated inter-genic regions (INTER) or superimposed to annotated genes (SUPER) not supported by periodicities, are highlighted in a separate track, and can be visually evaluated in comparison to profiles of frame-specific nucleotide-usage (frame-analysis), to regions of significant three-base periodicity ('Hits') and to the position of pre-annotated genes. NPACT also generates nucleotide and translated amino acid sequences for newly identified ORFs, for subsequent analyses.

Several features distinguish NPACT representations from other frame-analysis viewers and from a classical 'genome-browser' representation: (i) NPACT generates a continuous representation of the genome sequence and its features in a multi-line format enabling canvas interactivity under the KonvaJS HTML5 Canvas JavaScript framework. Compared with a traditional 'genome-browser' representation, canvas interactivity allows users to seamlessly navigate through an entire prokaryotic genome in a few seconds, facilitating discovery and analysis of inconsistencies between annotated genes and newly identified ORFs. (ii) Besides plotting by default phase-specific profiles of GC content (S-profiles) as in frame analysis, NPACT allows users to explore profiles of any alternative nucleotide combination ('N-profiles'). This feature can be useful for the analysis of sequences of any GC content. (iii) Visualization of all H-type and G-type hits provides information on the statistical significance, coding-frame support, and boundaries of the contrasts visualized by the compositional profiles, as well as on the presence of underlying reading-frame structures. (iv) By plotting pre-annotated genes with hits and profiles, inconsistencies between annotated genes and sequence periodicities can be visually identified and evaluated. (v) Highlighting newly identified ORFs in a separate track directs focus on three-base periodicities not explained by previously identified genes and facilitates their visual evaluation in comparison with hits, compositional profiles, and position relative to predicted neighboring genes.

Measured on a MacPro Quad 2.4 GHz personal desktop computer, genome sequences and annotations are processed and visualized by NPACT at a rate of ∼0.2 Mbp/s, practically constant over sequence lengths up to at least 90 Mbp.

3 Results

3.1 Power of the tests

We assessed the potential of coding regions of different length and composition to generate significant periodicities over sets of functionally characterized coding sequences (CDS) annotated in 1083 prokaryotic genomes from the NCBI Genome database. A subset of 2 452 743 CDS (here referred to as the 'Characterized' set) was created excluding all sequences described as hypothetical (from NCBI *.ffn files), which formed a second set of 1 094 230 'Hypothetical' sequences (Table 1; Supplementary Table S5 for compositional details). Three-base periodicities were identified at two significance levels ($\alpha=10^{-2}$ and $\alpha=10^{-3}$) in both sets (Table 1). Overall, hits of the H-type or G-type were identified in ∼85% of the sequences in the Characterized set, whereas hits of either type were found in > 90% of the sequences (84.5% at $\alpha=10^{-3}$). In comparison, a significantly lower fraction of hits were identified in the Hypothetical set (72.0% at significant level $\alpha=10^{-2}$, and 61.3% at $\alpha=10^{-3}$).

Partitioning the Characterized and Hypothetical datasets into classes of different length and GC-content (Supplementary Table S5), we found as expected that the frequency of genes with significant three-base periodicities increased with sequence length (Fig. 3A). Compositional periodicities were identified in the vast majority (>99%) of Characterized genes with length ≥ 600 codons, and in > 90% of the sequences of length ≥ 250 codons at the significance level $\alpha=10^{-2}$. Significant periodicities were still observed in the majority (60.6%) of the sequences in the length range 50–99 codons and in almost one-third (31.9%) of sequences shorter than 50 codons. Mirroring the overall result, we found a lower frequency of hits among Hypothetical sequences than among those Characterized (Fig. 3A), demonstrating that the lower frequency of periodic ORFs in the Hypothetical set did not depend on the average shorter length of hypothetical genes (Table 1).

The power of the tests was also significantly affected by the GC content of the sequences (Fig. 3B). Significant hits ($\alpha=10^{-2}$) were identified in > 98% of all characterized sequences of high GC content (GC ≥ 0.60). Among sequences with a GC content in the range [0.40–0.45), corresponding to the least intense contrasts in frame-specific GC content (Supplementary Fig. S1), significant hits were found in ∼79% of the sequences. Consistently with the overall results, fewer significant hits were identified among hypothetical genes within each GC-class, particularly among genes of low GC content. Similar trends were identified for individual tests (Supplementary Table S6) and using a higher threshold of significance (Supplementary Table S7). The joint effect of length and GC content on the power of the combined tests is represented in Figure 3C

Table 1. Three-base periodicities in sets of characterized and hypothetical coding sequences

Sets	Size	Mean GC%	Mean Len/nt	%H	%G	%H∪G	%H∩G
Ch	2 452 743	53.3	1 067	84.71	85.45	90.61	79.55
				77.95	77.96	84.50	71.41
Hy	1 094 230	50.9	683	62.37	64.55	71.96	54.96
				52.22	54.12	61.28	45.06
Tot	3 546 973	52.5	948	77.80	79.00	84.84	71.95
				70.00	70.59	77.32	63.27

Sets: Characterized (Ch) and Hypothetical (Hy). Size: number of coding sequences in the set. %H: Percent of coding sequences with H-type hits. %G: Percent of coding sequences with G-type hits. %H∪G: Percent of coding sequences with hits of either type. %H∩G: Percent of coding regions with both types of hits.

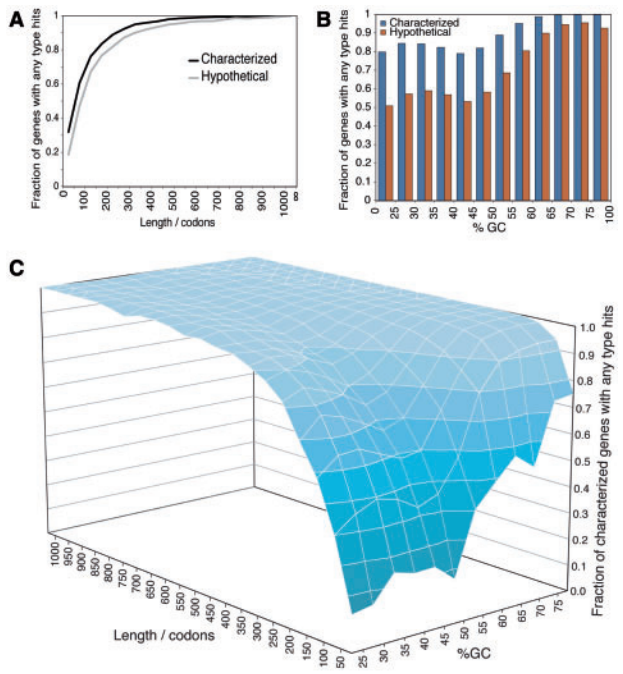

Fig. 3. Frequency of genes with H-type or G-type hits ($p \leq 10^{-2}$) in sets of genes annotated in about 1000 published prokaryotic genome sequences. (**A**) Frequency within the sets of functionally described genes ('Characterized') and of 'Hypothetical' genes for different classes of sequence length. (**B**) Frequency by GC content within the Characterized and Hypothetical sets of genes. (**C**) Frequency within the Characterized set of genes, partitioned by sequence length and GC content

($\alpha = 10^{-2}$, Characterized set), showing very high power for sequences of all compositions of sufficient length (>0.90 for all classes of length > 400–450 codons). Among sequences of high (≥60%) GC content, power reached 1.00 and remained consistently above 0.90 even among sequences as short as 50–100 codons. Significant periodicities were also observed in > 60% of the sequences shorter than 50 codons and with GC content ≥ 70% (Supplementary Table S6 for similar relations for individual tests). A higher stringency in significance level ($\alpha = 10^{-3}$) affected power mostly among short sequences (<100 codons) (Supplementary Fig. S5 and Supplementary Table S7). These results reflected the high power of frame analysis in its application to GC-rich sequences, for which the sensitivity of our methods compared to, or even exceeded, that of probabilistic approaches based on high-order Markov models (Supplementary Table S8). It also showed that three-base periodicities can provide useful information on the presence of coding regions also in sequences of lower GC content.

3.2 Newly identified ORFs with significant periodicity

We analyzed with NPACT the set of more than 1000 genomes used for the sets of Characterized and Hypothetical annotated genes. We identified in INTER regions 167 611 ORFs of significant periodicity, corresponding to an average of 44.6 ORFs/Mbp (Supplementary Table S9). Despite differences in power of the tests, inter-genic ORFs with three-base periodicity were identified in sequences of all GC contents, with frequencies ranging from 33 ORFs/Mbp in replicons of 40–45% GC content, to 60 ORFs/Mbp in those of 55–60% GC content (Supplementary Table S9). To seek further support to the coding potential of these ORFs, we evaluated their evolutionary conservation in sequence and in length (Supplementary Material), and we

Table 2. Conservation and prediction of ORFs newly identified in 1 000 genomes

		Number of predictors					Total Pred
		0	1	2	3	4	
INTER	N	84 165	11 016	6 052	5 190	6 364	28 622
	F	2 492	1 115	1 418	2 082	8 791	13 406
	C	3 778	1 902	2 307	4 435	26 504	35 148
	Tot	90 435	14 033	9 777	11 707	41 659	77 176
SUPER	N	140 243	3 664	835	522	527	5 548
	F	3 506	182	84	95	100	461
	C	5 448	369	238	249	1 120	1 976
	Tot	149 197	4 215	1 157	866	1 747	7 985

Number of ORFs identified in INTER or SUPER regions that are not conserved across genera (N), conserved fragments (F), or conserved (C), and are predicted by 0 to 4 computational gene predictors. 'Total Pred' is the total number of genes predicted by one or more predictors.

compared the newly identified ORFs to collections of predicted genes automatically generated by the four popular gene-prediction methods Prodigal (Hyatt *et al.*, 2010), Glimmer3.0 (Delcher *et al.*, 2007), GeneMarkHMM (Lukashin and Borodovsky, 1998) and GeneMark2.5 (Borodovsky and McIninch, 1993), available at the NCBI bacterial-genome website. About one-quarter of all newly identified ORFs were conserved across genera or phyla (Table 2), with no clear trend across classes of GC content (Supplementary Table S9). Surprisingly, we also found that almost half of all 'inter-genic' ORFs were predicted by one or more of the gene predictors we tested, among which were ~90% of the conserved ORFs (Table 2 and Supplementary Table S9). A significant finding was that the majority of the ORFs predicted by at least one method were in fact predicted by all, and more than two-thirds by at least three methods (Supplementary Table S10), indicating that the exclusion of these ORFs from the annotations did not depend on limitations in the earlier gene predictors GeneMark2.5 and GeneMarkHMM. Furthermore, conserved and predicted ORFs were relatively long (693 nt on average, Supplementary Table S11), indicating that size was not a factor in their exclusion.

We also identified, across replicons of any GC content, 157 182 ORFs (41.8 ORFs/Mbp) in SUPER regions (Supplementary Table S12). In contrast to inter-genic ORFs, the coding potential of > 90% of these ORFs was neither supported by conservation nor by any gene predictor (Supplementary Table S12). In combination with their shorter length (160 nt on average, Supplementary Table S13), lack of conservation and lack of support from gene predictors suggest that the 'gene-like' periodicity of these ORFs was more likely induced by compositional fluctuations in the superimposed gene. Nevertheless, also among superimposed ORFs we identified 7424 longer ORFs (398 nt long on average) that were evolutionarily conserved in sequence and in length. Of these conserved ORFs, ~25% were also identified by some or by all prediction methods (Supplementary Tables S12 and S14).

3.3 The coding potential of *Anaeromyxobacter dehalogenans*

We analyzed the GC-rich (~74.8% GC content) genomes of strains 2CP-1 (Accession Number NC_011891) and 2CP-C (NC_007760) of the deltaproteobacterium *Anaeromyxobacter dehalogenans* using the NPACT tool as part of an annotation pipeline, by comparing collections of genes resulting from previous annotation efforts to ORFs with significant three-base periodicity. Selected examples of

Table 3. Annotated genes and significant compositional periodicities in two strains of *Anaeromyxobacter dehalogenans*

Features	2CP-1	2CP-C
Genome size	5 029 329	5 013 479
Annotated genes (CDS)	4 473	4 346
Total hits	6 235	6 114
Confirmed CDS	4 318	4 128
Contradicted CDS	10	37
New intergenic ORFs	125	209
New superimposed ORFs	6	34

CDS is coding regions annotated with RefSeq genome. ORFs are identified by regions of three-base periodicity. Confirmed and Contradicted is in reference to identified regions of three-base periodicity. Inter-genic or Superimposed is in reference to annotated CDS.

newly identified ORFs as visualized by NPACT are shown for strain 2CP-C in Supplementary Figure S6. Complete graphical results (significance level 0.01) for both strains are available as Supplementary Figures S7 and S8, respectively. The NPACT graphical comparison of S-profiles and annotated coding regions of the two *A. dehalogenans* genomes showed in the vast majority of cases a strong correspondence of sharp contrasts in S-profiles with annotated genes. The statistical significance of these contrasts and their association with annotated genes were in most cases confirmed by the identification of corresponding hits. However, NPACT also identified several ORFs with significant compositional periodicity not corresponding to any of the annotated coding regions (Table 3). Most of these ORFs occupied 'inter-genic' regions separating annotated genes (Supplementary Fig. S6A for examples).

In a few cases, NPACT identified compositional periodicities consistent with coding regions superimposed in a different frame to annotated genes (Supplementary Fig. S6B). Overall, considerably more missed or mis-annotated genes could be identified in *A.d.* 2CP-C than in *A.d.* 2CP-1. Among newly identified ORFs, the majority (55%) were conserved between the two strains (Supplementary Table S15), as determined by > 30% similarity between the corresponding proteins. In the majority of cases, we found homologs of these newly identified ORFs among coding regions annotated in the other strain (92.4% of newly identify ORFs from 2CP-C and 83.9% of those from 2CP-1) whereas only 19 homologous-pairs were newly identified in both genomes. Inclusion of the newly identified ORFs resulted in similar number of genes encoded by the two genomes.

As in the analyses of 1000 genomes, we characterized each newly identified ORF for conservation, among *Anaeromyxobacter* species/strains, across species belonging to different genera of deltaproteobacteria, or across different phyla, and recorded if it was among genes (identified by same stop-codon position) predicted by Prodigal, Glimmer3.0, GeneMarkHMM, and GeneMark2.5. We also assigned to each ORF a 'Quality' score based on visual examination of S-profiles (do S-profiles have the expected relations along the ORF?), good correspondence between ORF and hit (does the hit cover most of the ORF?), and positional relation with neighboring genes (does the ORF fit within or between operons?). Results for all ORFs are reported with available functional characterizations in Supplementary Tables S16–S23, listed by strain and class of conservation.

Summary statistics on conservation of newly identified ORF are reported in Table 4 in comparison to pre-annotated genes. Most pre-annotated genes were conserved across genera (88.5%) or phyla (83.1%), and virtually all were conserved among *Anaeromyxobacter* species or strains. Among newly identified ORFs, 35.9% of those identified in 2CP-1 and 27.6% of those from 2CP-C

Table 4. ORF conservation across taxonomic units

Strain	2CP-1		2CP-C	
Conservation	# CDS (%)	# ORF (%)	# CDS (%)	# ORF (%)
Phyla	3 667 (82.0)	39 (29.8)	3 631 (83.5)	45 (18.5)
Genera	243 (5.4)	3 (2.3)	224 (5.2)	17 (7.0)
Anaeromyxo.	560 (12.5)	9 (6.9)	491 (11.3)	83 (34.2)
Total cons.	4 470 (99.9)	51 (38.9)	4 346 (100.0)	145 (59.7)
Non-cons.	3 (0.1)	80 (61.1)	0 (0.0)	98 (40.3)
Total	4 473	131	4 346	243

CDS is coding regions annotated in the genome. ORF includes potential coding regions identified by three-base periodicity and not included in published genome annotations. Numbers (percentages) refer to queries conserved only up to the indicated taxonomic level. For example, 'Genera' corresponds to the number (percentage) of queries with homologs identified across different genera of deltaproteobacteria (phylum) but not in other phyla (E-value < 1.0E-6).

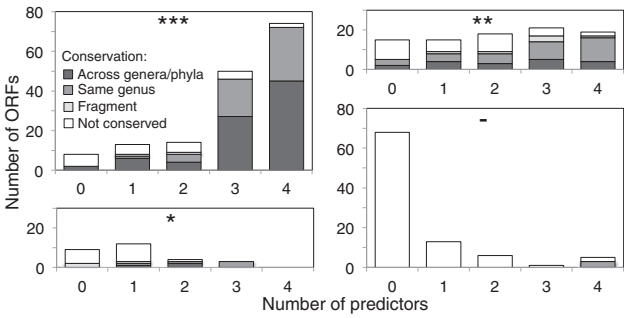

Fig. 4. Conservation of newly identified ORFs of four levels of 'quality' determined based on visual analysis of S-profiles, hits, and of ORF positional relation with neighboring genes (best: ***; good: **; possible: *; dubious: -). ORFs are characterized by the 'Number of predictors' (0 to 4) supporting their coding potential

were conserved at least across genera and mostly across phyla (Table 4). Many of the conserved ORFs newly identified in one strain were homologs of genes annotated in the second strain (48 ORFs newly identified in 2CP-C and 25 in 2CP-1) and most were functionally characterized (Supplementary Tables S16 and S17). Some were identified as encoding widely conserved protein families, such as, in 2CP-C, the ribosomal protein L28 and the molecular chaperone DnaJ, and in 2CP-1 chemotaxis protein CheY. Furthermore, virtually all the newly predicted ORFs conserved across genera found in 2CP-C, and the majority of those found in 2CP-1, were predicted by most or all gene-predictors. Eighty-three ORFs from 2CP-C (Supplementary Table S18) and nine ORFs from 2CP-1 (Supplementary Table S19) were conserved only in *Anaeromyxobacter* species/strains. In contrast to the ORFs conserved across genera and phyla, the vast majority of these poorly conserved ORFs matched non-characterized hypothetical genes. However, in most cases these functionally uncharacterized ORFs were also supported by most gene prediction methods. We did not identify any conserved homologs in the NCBI nr database for 166 ORFs (Supplementary Tables S20 and S21), most of which were not or poorly supported by gene-prediction methods. Finally, 11 newly identified ORFs appeared to be gene fragments (Supplementary Tables S22 and S23).

Virtually all ORFs conserved across genera and phyla, as well as those conserved between *Anaeromyxobacter* species or strains, showed best (***) or good (**) visual quality, and among these most were confirmed by the majority of gene predictors (Fig. 4).

These results suggested that visual assessments of quality based on compositional contrast and position are reliable indicators of the coding capacity of an ORF, as supported by conservation and/or computational prediction, and thus of the level of confidence in the prediction. Furthermore, positional and compositional quality as well as gene-prediction, mutually supported many potential genes that could otherwise be excluded based on their lack of or limited conservation (Fig. 4 and Supplementary Tables S18–S21).

4 Conclusions

Accuracy in genome annotation depends on successful integration of the information provided by computational gene-prediction methods, by pre-existing annotations of closely related genomes, by gene conservation, and by evidence of functionality and expression (Koonin and Galperin, 2003; Madupu *et al.*, 2010; Pati *et al.*, 2010). Given the inherent uncertainty in computational gene prediction and in identification of homology, as well as the complexity of whole-genome annotation, it is perhaps not surprising that studies on conservation and functional characterization of genomic ORFs (Warren *et al.*, 2010), or of sets of genes automatically predicted by Glimmer3.0 (Wood *et al.*, 2012) showed evidence of several conserved genes missing from genome annotations. We showed how genes missed in prokaryotic genome annotations can be detected by quantitative and visual frame analysis using intrinsic information from three-base periodicity. Extending frame analysis from phase-specificity in GC content to phase-specificity in any nucleotide type, our procedures are applicable to sequences of variable composition. Although NPACT is most effective on sequences of high GC content (55% or higher), we showed that three-base periodicities are useful to identify conserved coding regions missed by the annotations also in genomes of lower GC content. Furthermore, because our method only requires a few base pairs to determine the *local* composition of the sequence, it can be used to identify potential coding regions in sequences too short or compositionally too heterogeneous to derive detailed sequence-specific gene models. These features make NPACT a useful tool for the analysis of collections of DNA or mRNA sequence fragments, such as those that can be obtained from metagenomic sequencing projects.

We were surprised to discover that many three-periodic ORFs not included in annotations were also predicted by most or all popular gene prediction methods, indicating that the slightly lower sensitivity of the methods available and commonly used by the earliest annotators—GeneMark2.5, published in 1993—does not explain the exclusion of these ORFs from genome annotations. Lack of conservation or small size of an ORF could be criteria for excluding predicted genes as likely false positives. However, among the excluded genes predicted by most methods, many were also relatively long, were conserved in sequence and in length across distantly related genomes, and were functionally characterized. These findings suggest that the sensitivity and specificity of gene predictors is not the only factor, and maybe not the dominant factor, limiting accuracy of prokaryotic gene annotation. Bioinformatics tools that facilitate comparison and visualization of information for the evaluation of predicted genes across genome sequences, may have a greater impact on the quality of gene annotations than could be achieved from improvements in computational gene-prediction algorithms. Along these lines, we believe the representation of sequence features provided by NPACT can significantly contribute to the amelioration of the annotations of high GC content (and other) genomes facilitating the identification of true novel or predicted genes.

Acknowledgements

We thank the many colleagues with whom we discussed our work and an anonymous reviewer for his constructive criticisms.

Funding

National Institutes of Health (NIH) Grant 5R01GM087485 to L.B.

Conflict of Interest: none declared.

References

Angiuoli,S.V. *et al.* (2008) Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation. *OMICS*, **12**, 137–141.

Aziz,R.K. *et al.* (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

Besemer,J. and Borodovsky,M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.*, **27**, 3911–3920.

Bibb,M.J. *et al.* (1984) The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene*, **30**, 157–166.

Borodovsky,M. and McIninch,J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.*, **17**, 123–133.

Brocchieri,L. *et al.* (2005) Predicting coding potential from genome sequence: application to betaherpesviruses infecting rats and mice. *J. Virol.*, **79**, 7570–7596.

Delcher,A.L. *et al.* (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.

Goranovic,D. *et al.* (2012) FK506 biosynthesis is regulated by two positive regulatory elements in *Streptomyces tsukubaensis*. *BMC Microbiol.*, **12**, 238.

He,Q. and Sanford,R.A. (2003) Characterization of Fe(III) reduction by chlororespiring *Anaeromyxobacter dehalogenans*. *Appl. Environ. Microbiol.*, **69**, 2712–2718.

Huang,T. *et al.* (2011) Identification and characterization of the pyridomycin biosynthetic gene cluster of *Streptomyces pyridomyceticus* NRRL B-2517. *J. Biol. Chem.*, **286**, 20648–20657.

Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Ishikawa,J. and Hotta,K. (1999) FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. *FEMS Microbiol. Lett.*, **174**, 251–253.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, **87**, 2264–2268.

Koonin,E.V. and Galperin,M.Y. (2003) Genome annotation and analysis. In: *Sequence—Evolution—Function. Computational Approaches in Comparative Genomics* Kluwer Academic, Boston.

Kumar,K. *et al.* (2011) AGeS: a software system for microbial genome sequence annotation. *PLoS One*, **6**, e17469.

Larsen,T.S. and Krogh,A. (2003) EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, 21.

Lee, E. *et al.* (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, **14**, R93.

Lewis,S.E. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, RESEARCH0082.

Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.

Madupu,R. *et al.* (2010) Meeting report: a workshop on best practices in genome annotation. *Database*, **2010**, baq001.

Marshall,M.J. *et al.* (2009) Electron donor-dependent radionuclide reduction and nanoparticle formation by *Anaeromyxobacter dehalogenans* strain 2CP-C. *Environ. Microbiol.*, **11**, 534–543.

Pati,A. *et al.* (2010) GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat. Methods.*, **7**, 455–457.

Rutherford,K. *et al.* (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

Salzberg,S.L., *et al*. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.

Sanford,R.A. *et al*. (2007) Hexavalent uranium supports growth of *Anaeromyxobacter dehalogenans* and *Geobacter* spp. with lower than predicted biomass yields. *Environ. Microbiol.*, **9**, 2885–2893.

Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 623–656.

Sherwood,E.J. *et al*. (2013) Cloning and analysis of the planosporicin lantibiotic biosynthetic gene cluster of *Planomonospora alba*. *J. Bacteriol.*, **195**, 2309–2321.

Sokal,R.R. and Rohlf,F.J. (1994) *Biometry: The principles and practices of statistics in biological research*. Freeman and Co., New York.

Stewart,A.C. *et al*. (2009) DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinformatics*, **25**, 962–963.

Vallenet,D. *et al*. (2013) MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.*, **41**, D636–D647.

Warren,A.S. *et al*. (2010) Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics*, **11**, 131.

Wood,D.E. *et al*. (2012) Thousands of missed genes found in bacterial genomes and their analysis with COMBREX. *Biol. Direct*, **7**, 37.

Yu,C. *et al*. (2008) The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. *BMC Bioinformatics*, **9**, 52.