This is a draft of a chapter that has been accepted for publication by Oxford University Press in the book "*Codon Evolution: Mechanisms and Models*" edited by Gina M. Cannarozzi and Adrian Schneider, due for publication in 2012. Available from http://ukcatalogue.oup.com/product/9780199601165.do

---

15. Distance measures and machine learning approaches for codon usage analyses

Fran Supek and Tomislav Šmuc, Division of Electronics, Rudjer Boskovic Institute, Zagreb, Croatia.

Abstract: Unequal use of synonymous codons is a widespread phenomenon caused largely by directional mutation pressures and selection for genomic nucleotide composition, but also by selection for speed and/or accuracy of protein translation. Much effort was dedicated to investigate whether this 'translational selection' had an influence on codon choice of highly expressed genes in various genomes. In such analyses genes are typically represented as vectors of codon frequencies, and the data analyzed using multivariate techniques, commonly either (a) dimensionality reduction, e.g. correspondence analysis, or (b) distance measures in the codon frequency space, such as the codon adaptation index (CAI). Such representations of data can be challenging as genes are too short to allow precise estimation of codon frequencies, introducing noise and consequently leading to serious artifacts in some commonly used methods. A supervised machine learning approach, as embodied in the use of a classifier, provides an alternative more robust to noise and also more sensitive in detecting codon biases. We describe a Random Forest-based computational framework that enables control over confounding factors (here, the background nucleotide substitution patterns) while reliably detecting translational selection, demonstrated on a large set of prokaryotic genomes.

1. Causes of biased codon usage in genomes

Genetic code is a redundant, many-to-one mapping from codons to amino acids. Even though use of one or another synonymous codon does not lead to a different amino acid sequence, the synonymous codons are not used equally: there is a layer of biologically relevant information between genes at the level of coding sequences in the DNA, and the level of the protein sequence. The wide range of G+C content spanned by the known genomic sequences is illustrative of how nucleotide substitution patterns vary greatly across genomes, and these genome-wide biases in substitution patterns are the principal determinant of non-random codon usage in prokaryotes (Chen et al. 2004; Knight et al. 2001). On the other hand, there is also significant within-genome variation in direction and strength of nucleotide substitution patterns: along the prokaryotic chromosome, there is a general tendency toward A+T-enrichment near the replication terminus (Daubin and Perriere 2003), and vertebrate genomes exhibit so-called 'isochore' structure where the local G+C content varies along the chromosome in a periodic manner. Furthermore, genomes often display an asymmetry between the two DNA strands where the leading strand is enriched in G over C and T over A or 'GC-skewed', mostly consequent to deamination of cytosine in single-stranded DNA exposed during replication (Lobry and Sueoka 2002). The organizational features of prokaryotic genomes with respect to local sequence composition and gene distribution were reviewed by Rocha (2004).

Such biases in mutational processes appear to be an important contribution to the nucleotide substitution patterns, and may result from the nature of chemical changes to the nucleotides, but also from biases in errors of DNA replication and repair. In addition to being driven by the underlying mutational biases, the specific nucleotide compositions may have adaptive value and be selected for (Rocha and Danchin 2002); this also holds true for dinucleotides (Zeldovich et al. 2007). Two recent investigations have found that mutations in bacteria are generally biased toward A and T, while the high G+C content of some genomes would be explained by selective forces (Hildebrand et al. 2010; Hershberg and Petrov 2010). Possible explanations for selection on prokaryotic GC content are briefly reviewed in Rocha and Feil (2010).

Competing with the nucleotide substitution patterns, selection is acting on silent sites to make protein translation more accurate and also more 'efficient', in this context implying 'faster'.[1] Traditionally, the biases resulting from selection on codon usage were linked to abundances of tRNA isoacceptors for a particular codon (Kanaya et al. 1999). This correlation is consistent with a mechanistic model where the speed of translational elongation is limited by availability of charged tRNA molecules (Xia 1998). Biased codon use has also been shown to guard against missense and nonsense errors in proteins in the *Escherichia coli* genome; there, highly conserved sites and genes have a higher codon bias, but codon bias is also correlated to gene length and increases along the genes' length (Stoletzki and Eyre-Walker 2007). More recently, other more subtle translation-related determinants of codon usage have been observed, for instance the selective charging of tRNAs which promotes use of amino acid starvation-insensitive codons in amino acid biosynthetic pathways (Dittmar et al. 2005), and the 'load minimization' hypothesis which states that there should be a preference towards codons whose mutated forms cause less structural disruption to proteins (Najafabadi et al. 2007). Interestingly, a synonymous mutation in a human gene was shown to produce a phenotype via an altered protein structure (Kimchi-Sarfaty et al. 2007), an occurrence which may be related to the correlations between codon usage and protein structural features that have been observed previously by Oresic and Shalloway (1998). Regularities in codon order within genes were examined in multiple genomes and proposed to be related to translation dynamics in two different ways: (i) beginnings of genes exhibit a 30-50 codon 'ramp' where elongation is slowed down by use of specific codons (Tuller et al. 2010) and (ii) recurring instances of an amino acid in a protein tend to be encoded by codons read by the same tRNA more frequently than expected by chance (Cannarozzi, Schraudolph, Faty et al. 2010).

Only a small set of highly abundant or rapidly induced proteins – a typical representative are the ribosomal protein (RP) genes – is expected to be strongly affected by selection for translational efficiency and accuracy. The portion of a genome undergoing at least some degree of translational optimization may, however, be larger. Which genes are a part of this specific subset is an organism-specific characteristic which was speculated to be related to the organism's preferred environment (Carbone 2006). A variety of single-organism analyses of prokaryotic genomes have reported no influence of translational selection in the genomes, most notably for the insect endosymbionts *Buchnera* (Rispe et al. 2004), *Wigglesworthia* (Herbeck et al. 2003) and *Blochmannia floridanus* (Banerjee et al. 2004), or the slow-growing pathogens *Borrelia burgdorferi* (McInerney 1998) and

[1] Note that the use of the word 'efficiency' is in fact not fully appropriate in this context (Dethlefsen and Schmidt 2005). Still, in this text we opted not to use the more correct term 'translational power' in order to be more consistent with previous literature.

Helicobacter pylori (Lafay et al. 2000). Inconsistent with absence of translational selection, in *Buchnera* a correlation was found between codon composition of highly expressed genes and experimentally measured abundances of tRNAs (Charles et al. 2006). Three large multiple-genome analyses have not detected evidence of translational selection in approx. 75% (dos Reis et al. 2004) , 50% (Carbone et al. 2005) or 30% (Sharp et al. 2005) of the prokaryotes analyzed. The quite sizable differences in the extent of translational selection reported by these three studies can be explained by the use of a different mathematical apparatus in each analysis; it is unclear which of these approaches produced a more dependable result. Interestingly, on several occasions it has been found that several gene functions close in codon usage to RP genes in *E. coli* also have RP-like codon usage in some of the organisms which are supposed to lack translational selection, see e.g. the glycolysis genes in *H. pylori* (Carbone and Madden 2005) or respiration and ATP synthase genes in *B. floridanus* (Mrázek et al. 2006). If translational selection was not active in these organisms, it would be highly improbable to expect biased codon usage in any specific gene functional category, let alone a function that may be adaptive to a specific environmental conditions or a specific physiology. In summary, there are significant inconsistencies in the literature concerning the prevalence of translational selection among and within genomes; the relationship of translation-related codon biases to microbial ecology and physiology has also not been examined in a systematic, large-scale manner.

Investigations of codon biases in genomes are interesting from a theoretical viewpoint, because these analyses (a) provide new knowledge of the rules by which DNA sequences evolve, but also (b) by using codon biases as a proxy for expression levels, offer new insight about organisms' adaptation to environment. In addition, there is a single quite significant application of codon bias analyses in biotechnology: optimization of gene sequences for heterologous expression (Welch et al. 2009b). For instance, when trying to overexpress a human gene in an *E. coli* host, if the gene contains many codons that the *E. coli* translational machinery translates slowly or inaccurately, the functional protein product may not be produced in great abundance; a prominent example are the Arg codons AGA and AGG. Common strategies to alleviate this problem include changing the gene sequence to consist exclusively of codons common in naturally highly expressed genes, or alternatively modifying the sequence so it mimics the codon frequencies of the host organism. Several software solutions exist to facilitate this process (Puigbo et al. 2007; Supek and Vlahovicek 2004). A recent investigation on foreign protein expression in *E. coli* employed libraries of silent site variants of proteins to conclude that the codons conducive to very high (above physiological levels) expression levels are different than the ones typically considered 'optimal' for the genome-encoded genes of the *E. coli* host (Welch et al. 2009a). As an alternative to the 'codon optimization' strategy, the problem of failed heterologous expression may be alleviated by using an expression host engineered with extra copies of certain tRNA genes (Tegel et al. 2010), such as the Agilent CodonPlus or the Novagen Rosetta strains of *E. coli*.

2. Methods for quantifying codon biases

A variety of specialized statistical approaches have been invented for the purpose of codon usage analyses, and the implementations of these methods are available in freely accessible software, such as INCA (Supek and Vlahovicek 2004). In the following section we aim to describe the prominent examples.

2.1. Unsupervised methods

Many methods commonly used for codon usage analyses are statistics that attempt to summarize genes' codon frequencies to a single numeric value – a measure of pairwise distance between vectors of codon frequencies. A typical example is the popular 'codon adaptation index' or CAI (Sharp and Li 1987) that measures the distance to the codon usage of a predefined set of highly expressed genes, and has been established as a surrogate for gene expression under optimal growth conditions of *E. coli* and *S. cerevisiae.* Other commonly used distance measures for vectors of codon frequencies include: (i) the "codon bias between gene groups" (CB) which is essentially a weighted Manhattan distance employed on many occasions employed by Karlin and Mrazek (2000) and colleagues for finding 'predicted highly expressed' genes in microbial genomes; and (ii) the "measure independent of length and composition" (MILC) (Supek and Vlahovicek 2005) which is a corrected χ^2-type statistic devised to resolve methodological deficiencies present in other widely used approaches, most notably the CB which strongly overestimates distances when comparing short genes (Fig. 1).

Here we describe how these distance measures are computed. Let G indicate a gene or group of genes with codon frequencies $g(x,y,z)$ for a codon (x,y,z) normalized such that $\Sigma\, g(x,y,z) = 1$, where the sum extends over all codons (x,y,z) translated to amino acid a. Let $f(x,y,z)$ indicate the codon frequencies of another gene or gene group F, again normalized to sum to 1 within each amino acid. Let $p_a(F)$ be the amino acid frequencies of the gene/genes in F, which sum to 1 over all amino acids.

The 'codon bias' measure (Karlin and Mrazek 2000) is computed as:

$$CB(F,G) = \sum_a p_a(F) \left[\sum_{(x,y,z)=a} \left| f(x,y,z) - g(x,y,z) \right| \right]$$

Distance in the codon frequency space between the genes or gene groups F and G is estimated by the MILC method (Supek and Vlahovicek 2005) as:

$$MILC(F,G) = 2\sum_a p_a(F) \left[\sum_{(x,y,z)=a} f(x,y,z) \cdot \ln \frac{f(x,y,z)}{g(x,y,z)} \right] - \frac{1}{L}\sum_a (r_a - 1) + \frac{1}{2}$$

Where r_a is the redundancy class for an amino acid (2 for two-fold degenerate amino acids, 3 for isoleucine, and so on), and L is the gene length. The sums should iterate only over the amino acids present in the gene F. Note that the formulae in the original paper (Supek and Vlahovicek 2005) contain errors in the equations; for more information, please see the formal correction (Supek and Vlahovicek 2010).

The codon adaptation index (Sharp and Li 1987) is computed as:

$$CAI(F,G) = \prod_a \left[\prod_{(x,y,z)=a} \left(\frac{g(x,y,z)}{g_{\max}((x,y,z)=a)} \right)^{f(x,y,z)} \right]^{p_a(F)}$$

where "$g_{max}((x,y,z)=a)$" denotes the maximum frequency of the codons coding for amino acid a in gene group G. The ratio g/g_{max} is called codon 'relative adaptiveness' by Sharp and Li (1987), and the

CAI is a geometric mean of these adaptiveness values of all codons in the gene *F*. In a typical usage scenario of the CAI, the group G would consist of a set of highly expressed genes, sometimes also called a 'reference set'; see Fig. 2A,B for an example of a visualization of such distances using CAI and MILC methods. In contrast to CB and MILC, CAI decreases with increasing distance, therefore it is technically a measure of similarity.

Correspondence analysis (CA), an unsupervised dimensionality reduction technique, has often been used in single-genome studies to detect dominant trends in codon usage patterns, reflected in the 'factors' the method provides as output. This approach has lead to qualitatively different results regarding presence or absence of translational selection depending on how the data was normalized, as demonstrated for *B. burgdorferi* (Perriere & Thioulouse 2002), and for a larger number of genomes using the related technique of principal component analysis (PCA) (Suzuki et al. 2005). In addition to the issues with data normalization prior to CA or PCA, use of these methods may also lead to erroneous conclusions for a different reason: While PCA/CA generally do well in summarizing the main trends in between-gene codon usage variability into first two or three factors which are then typically visualized (as in Fig. 2C) or used for clustering of genes, it is not at all guaranteed that the following (ignored) factors would contain no information useful for the task at hand – here, this implies checking if the factors correlate to gene expression levels, a signature of translational selection in the genome. A rule of thumb for use of PCA is to retain as many factors as necessary to describe at least 95% of variance from the original dataset; when analyzing codon frequencies of genes in a single genome, the first three factors of PCA typically explain far less than 95% of variance, and therefore using only them in visualization is bound to omit at least some informative ('non-noise') trends in the data. This issue is further aggravated by the inability of PCA/CA to deal with non-linearity in the data, meaning that a single trend might be 'split' between more than a single factor output from PCA/CA, or missed altogether due to inherently non-linear dependencies.

2.2. Supervised methods

The previously described methods – distance measures in codon frequency space, dimensionality reduction, visualization and clustering – are the tools traditionally employed in codon usage investigations, and all have one thing in common: they are examples of an unsupervised approach to data analysis. In other words, that what is sought for does not have a bearing on how the data is treated; the distance measure or the projection technique, for instance, does not take into account whether a gene is highly expressed or not. By exploiting this additional source of information, an analysis may gain in power. A supervised machine learning approach, embodied in the classifier paradigm, should therefore in principle be more desirable than unsupervised techniques (distance measures or dimensionality reduction) in discerning a specific class of genes – when looking for translational selection, a set of highly expressed genes. Since the composition of this class is an *a priori* assumption in codon usage studies, it is prudent to incorporate this information via a supervised formulation of the problem at hand. The classifier can then 'learn' from the data a function that most accurately maps the codon frequencies to the class labels without

resorting to an (inflexible) definition of distance in the codon frequency space such as the CAI, CB or MILC.[2]

Following this principle, and in contrast to previous approaches based on unsupervised techniques, Supek et al. (2010) have introduced a supervised machine learning-based framework for detecting the presence and the extent of translational selection in 461 prokaryotic genomes. Their method is based on the Random Forest (RF) classifier (Breiman 2001) which they have evaluated in the task of discriminating a group of genes affected by selection for translational efficiency and/or accuracy, using only codon frequencies; see Box 1 for a short description of the RF algorithm. The group of ribosomal protein (RP) genes was assumed to be highly expressed and therefore a representative subset of genes under such selective pressures. The RF method allows the internal structure of the classification model to be visualized by a projection of the genes' pairwise distance matrix; an example for *E. coli* genes is given in Fig. 2D to illustrate how the data is internally represented by the (supervised) RF algorithm, in comparison to the (unsupervised) use of codon distance measures or the PCA technique (Fig. 2A-C).

Supek et al. (2010) have demonstrated RF to be a more accurate tool in discriminating the RP genes from the rest of the genes in the genome, when compared to three previous approaches based on pairwise distances (Sharp and Li 1987; Karlin and Mrazek 2000; Supek and Vlahovicek 2005); see Fig. 3 for a comparison on two selected genomes using the Receiver Operating Characteristic (ROC) curve methodology (Fawcett 2006). Additionally, the RF predictions correlated with experimental measurements of protein concentrations in *Escherichia coli* slightly better than previous methods did (Supek et al. 2010). Furthermore, this side-by-side evaluation of the methods has revealed the commonly used CAI method (Sharp & Li 1987) to be suboptimal for genomes with highly imbalanced G+C content, as was previously hinted by its author in an analysis of *Pseudomonas aeruginosa* genes (Grocock and Sharp 2002) . This would also explain the CAI's inability to predict gene expression levels in the A+T rich unicellular eukaryote *Plasmodium falciparum* (Supek and Vlahovicek 2005).

In addition to a classifier proving useful in examinations of codon biases in genome-encoded genes, supervised machine learning has also been employed to search for a relationship between codon usage and protein levels in heterologous gene expression (Supek and Šmuc 2010). In this specific instance, two non-linear regression methods – M5' decision trees and Support Vector Machines (SVM) – were used to show that codon frequencies play a role in determining protein expression in a library of 154 GFP variants that differed only in the genes' silent sites. In the original analysis of this 154 gene variant experiment (Kudla et al. 2009), this correlation was not found and 5' mRNA secondary structure was reported as the only determinant of protein expression; this was due to an overrepresentation of gene variants with very strong 5' structures in this dataset, coupled with the use of linear regression which did not capture the complex three-way relationship between 5' mRNA structure, codons and expression. This relationship became evident only with the M5' and SVM non-linear algorithms (Supek and Šmuc 2010). Note that both decision trees and the SVM can also be used as classifiers, i.e. for predicting categorical variables, instead of continuous ones (as in a regression setting).

---

[2] The dimensionality reduction techniques mentioned above rely on some notion of distance which is also blind to the class label.

3. Application to bacterial and archaeal genomes

We argued that introducing a supervised machine learning-based computational framework for codon usage analysis would lead to an increased sensitivity in detecting codon usage biases over commonly used unsupervised techniques, as well as better agreement to protein expression levels. Here we describe in detail how such an classifier-based approach was applied to prokaryotic genomes on a large scale, and what insights were gained from this analysis.

3.1. Rationale behind using classifiers to control for background nucleotide composition.

In addition to offering increased sensitivity over the unsupervised methods, there is another highly important issue in codon usage analyses that a classifier can help resolve: the need to control for a strong confounding factor – the background nucleotide composition – that shapes codon usage, but in a manner not related to protein translation. The background nucleotide composition results from an interplay of directional mutation pressures (mutational biases) and selection on genomic nucleotide composition, as we have discussed earlier.

To provide an illustrative example, let us imagine a genome with a strong 'GC skew', meaning the leading strand of DNA has an excess of G over C, which would cause the silent sites in codons of genes on the leading strand to prefer G over C even in the absence of translational selection, and vice versa on the lagging strand. Furthermore, let us suppose that optimal codons for that organism are commonly C-ending. If a specific gene on the lagging strand is found to be enriched with C-ending codons, to determine whether the gene was subject to translational selection, we would need to examine if the extent of the GC-skew (which can be quantified from non-coding DNA surrounding the gene) is sufficient to explain the bias towards C-ending codons observed in the gene. If not, the bias has to be due to a different reason; if the direction of the bias is towards the codon usage of a set of a highly expressed genes – such as the RPs – we can conclude this bias is related to translational selection. Note that another possible cause of codon bias in prokaryotic genes is horizontal transfer, where a gene retains the codon usage of a host genome. It is, however, unlikely that the codon usage of a gene unbiased in the genome of origin would match the codon preferences of highly expressed genes in the recipient genome, thus mimicking the effect of translational selection. Additionally, the horizontal transfer event should be evident in the composition of intergenic DNA, meaning it would be controlled for in the same manner as the GC-skew or the regional nucleotide composition biases.

In order to control for these factors that confound detection of translational selection, Supek et al. (2010) have harnessed a supervised machine learning-based computational methodology that couples the RF classifier to standard statistical tests. The classifiers' ability to learn mappings from one space to another allows for a convenient way to control for a strong confounder: the local/strand-specific nucleotide substitution patterns that affect codon usage, but not as a result of translational selection. By using a classifier both sets of data can be mapped to the same space (which here has a single dimension: the gene's probability of belonging to the highly expressed gene class), and the results than compared in this new space, thereby controlling for the confounder; see Fig. 5 for an illustration of the procedure. Using the crossvalidation technique to evaluate the RF models in multiple runs of computation, the framework has allowed the authors to obtain

statistically-backed calls about (a) in which genomes, (b) and on which genes in the genomes translational selection can be shown to act.

3.2. An example application of supervised machine learning in codon usage analysis.

Supek et al. (2010) employed the above strategy to test whether individual prokaryotic genomes are subject to translational selection. For each of the 461 genomes, the authors encoded the information about local nucleotide substitution patterns affecting each gene by computing mononucleotide and dinucleotide frequencies in the non-coding regions of DNA neighboring the translated part of the gene. Genes for functional RNA molecules such as tRNA and rRNA were also treated as coding and thus did not contribute toward mono- and di-nucleotide frequencies of intergenic DNA. The size of the neighborhood window was set to either 5, 10 or 20 kilobases upstream from the gene's start codon, and 5, 10 or 20 kilobases downstream from the stop codon. The window size of 10 kb upstream + 10 kb downstream guaranteed that in 99% of the genomes (457 out of 461), 99% of the genes have at least 142 non-coding nucleotides available for estimation of mono- and di-nucleotide frequencies. The analysis described below was re-run for all three values of the 'window size' parameter, and the results combined into a consensus set at a later stage.

To detect if selection for translational efficiency acts on a genome, Supek et al. (2010) employed the following procedure (depicted in Fig. 4): The RF classifier was first trained to distinguish ribosomal protein genes ('positive class') based solely on the mono- and di-nucleotide frequencies of genes' neighboring non-coding DNA within a given window size. Fifty runs of four-fold crossvalidation were used to estimate the accuracy of the classifier using the area-under-ROC-curve (AUC) score (Fawcett 2006), and the AUC for each of the 50 runs of crossvalidation was recorded.

This procedure was then repeated for a second time, however now the codon frequencies were also included in the dataset for the RF classifier training, in addition to the description of the intergenic regions. The AUC scores were again recorded for each of the 50 runs of crossvalidation. To determine if selection for translational efficiency acts on the genome, the sign test (McDonald 2009) was used to compare the 50 AUC scores obtained without codon frequencies to those obtained with them, for each genome. If the AUC score exhibited a consistent increase over 50 runs of crossvalidation, this meant that the introduction of codon frequencies improved the ability to discriminate ribosomal protein genes, providing evidence that translational selection acted on that specific genome. The weakest result was observed in the bacterium *Saccharophagus degradans 2-40*, the only genome where the change in AUC scores was not statistically significant at $p<10^{-3}$. The majority of examined prokaryotes (457 of 461) had sign test $p<10^{-9}$. The results were qualitatively equivalent for window sizes 5k and 20k, with 460 and 459 (of 461) genomes, respectively, having sign test $p<10^{-9}$. The *Saccharophagus degradans* 2-40 genome was previously found to exhibit a mosaic structure with respect to the G+C content (Weiner et al. 2008), probably due to large amounts of recently horizontally transferred (HT) DNA. The fact that translational selection has gone unnoticed might be explained if we hypothesize that not enough evolutionary time has passed for these HT segments to 'ameliorate' (Lawrence and Ochman 1997) to match the new host's translational apparatus; of course, it is also possible this organism genuinely does not exhibit translationally selected codon usage.

Since the codon frequencies consistently facilitated classification over the baseline, Supek et al. (2010) concluded that translational selection is, in fact, universal among prokaryotes, in contrast

to previous large-scale studies (Reis et al. 2004; Carbone et al. 2005; Sharp et al. 2005) which failed to find evidence of translational selection in many genomes (30-75%, depending on the study).

3.3. Proportion of genomes subject to translational selection and correlations with gene functional categories

Expectedly, the accuracy of codon-trained RF classifiers generally reflected the intensity of codon biases within a genome, but the accuracy is also bound to be related to the proportion of genes within a genome that are affected by translational selection - Supek et al. (2010) showed a sizeable portion of each prokaryotic genome to have an above-baseline codon usage similarity to the RP genes. It is easier to understand how such calls may be made on a per-gene level if we consider the classifiers' predictions for each gene as a measure of similarity of each gene's codon usage to the ribosomal protein genes, an approach in concept similar to the codon distance measures CB, CAI or MILC. Supek *et al.* (2010) proposed the following naming convention: a gene is declared to have optimized codon usage (OCU) if this similarity increases more frequently than expected by chance after codon frequencies are introduced to the classifier; statistical significance is determined by a conservative criterion of sign test $p<10^{-15}$ across 50 runs of crossvalidation. The authors found that genomes contained on average 13.2% of OCU genes, ranging from 5.4% to 33.0%.

Roughly speaking, the estimate of the extent of translational selection within bacterial and archaeal genomes expressed as % OCU genes (5-33%) is similar to the quantity arrived at in a study of eukaryotes (Resch et al. 2007) that reported purifying selection at synonymous sites in ~28% of the analyzed rat-mouse orthologous pairs. The findings from the two studies, when accepted together, suggest that it cannot be assumed that synonymous sites of protein coding genes in any genome evolve neutrally, regardless of the phyletic subdivision examined.

By subdividing the genes into OCU and non-OCU groups, Supek et al. (2010) did not imply that there was a clearcut boundary between the codon frequencies of the two groups. Instead, a gradient of codon usages exists in genomes (as seen for example in Fig. 2), where the genes labeled as OCU are those above the detection threshold of the RF-based method. This subdivision is conceptually similar to the approach formulated by Karlin and Mrazek (2000; Karlin et al. 2005; Mrázek et al. 2006) where a set of genes in the genome is labeled "PHX" (predicted highly expressed) by codon usage similarity to a set of RP and other translation-related protein genes. There are, however, three important features of OCU assignments that distinguish them from the PHX approach: (a) OCU labels are based on a RF classifier that outperformed the CB distance measure used for PHX assignments, (b) OCU is separated from non-OCU by a significance call of a statistical test instead of relying on an arbitrary threshold, and (c) OCU assignments are made using a control for local nucleotide substitution patterns which confound codon usage analyses. Use of a continuous variable instead of a categorical one, such as the OCU/non-OCU status, would be more descriptive with respect to gene expression levels. However, it remains to be seen to what degree of accuracy a quantitative estimate of expression can be reliably derived from codon biases alone, especially for genomes where translational selection is weak (e.g. slow-growing bacteria).

3.4. Distribution of codon-optimized genes within specific gene functional categories and relationship to microbial lifestyle

Generally, %OCU was smaller in larger genomes, and this anticorrelation could be largely explained by the change of relative proportions of certain gene functional categories with genome size: Smaller genomes have a larger proportion of genes that typically exhibit translationally selected codon usage, such as genes for protein production and energy metabolism (all steps leading to ATP production); such genes are mostly indispensable even in a very reduced genome. Larger genomes have an increased proportion of genes dealing with regulation, transport, sensing and signaling which are rarely codon-optimized. Several papers dealt with examining codon biases in specific gene functions in bacterial genomes using different methods (Carbone and Madden 2005; Karlin and Mrazek 2000) and were mostly consistent in their conclusions; equivalent general trends of correlation of gene function and codon bias were recognized also in Archaea (Supek et al. 2010).

On several occasions, researchers have touched on the issue how use of translationally optimal codons (consequent to elevated gene expression) in certain pathways might be related to the organism's environment. One would expect that, for instance, in obligate aerobes the citric acid cycle or the oxidative phosphorylation proteins would be more abundant than in facultative aerobes, and this should be reflected in codon usage biases of these functional categories. Precisely this example of aerobic vs. anaerobic microbial lifestyle was discussed in Bacteria (Carbone and Madden 2005; Karlin et al. 2005) and also in nine fungal genomes (Man and Pilpel 2007). Building on this concept, Supek et al. (2010) performed a systematic analysis of the distributions of translationally optimized genes along combinations of gene functional categories and different environments or phenotypes of organisms. They defined an 'adaptome' as a set of gene functions with expression levels elevated specifically in organisms living in a specific environment, where codon biases can serve as a proxy for expression levels in any fully sequenced genome enabling large-scale studies. As an example, the 'adaptome' of thermophilic bacteria and archaea was analyzed, and putative mechanisms for the protection of DNA and proteins against thermal denaturation were proposed (Supek et al. 2010). A number of other possible 'adaptomes' (related e.g. to use of certain modes of DNA repair, or to pathogenicity) were tested for statistical significance; the results are available from the authors' website at http://www.adaptome.org/.

The presence of translationally selected codon usage in mammalian genomes has been researched extensively (Urrutia and Hurst 2003) but is still an unresolved issue, as discussed in a recent analysis (Parmley and Huynen 2009). The RF classifier-based method Supek et al. (2010) applied to prokaryotes could potentially be fruitfully applied also to mammalian genomes where local variation in GC content (isochores) greatly complicates analysis, while non-coding DNA is plentiful, thus allowing for an dependable control for the local GC content variation.

3.5. mRNA expression levels and codon preferences of genes subject to translational selection

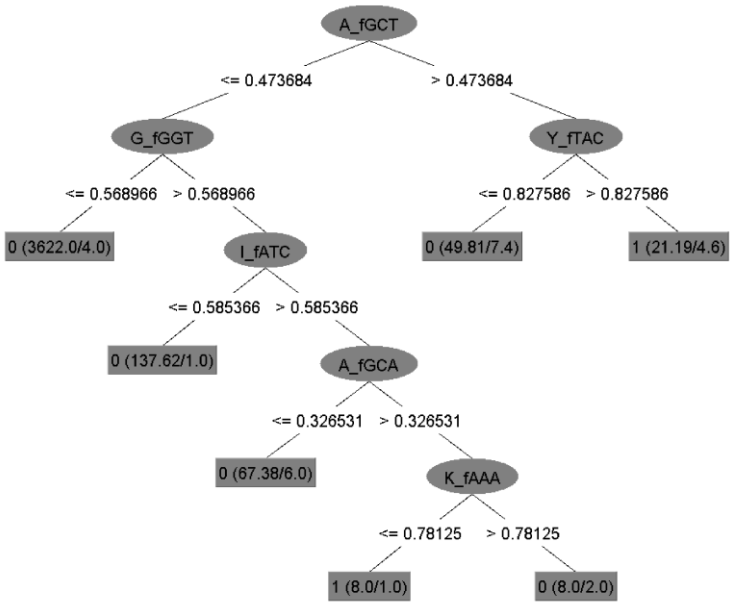
Supek et al. (2010) examined the correlation of OCU/non-OCU assignments to gene expression data from 19 phylogenetically diverse bacteria and found that OCU genes record microarray signal intensities on average 2.4-fold higher than non-OCU genes; for comparison, the RP – representing the most highly expressed genes – weigh in at 6.0x the average measurement in the same data. When quantifying the magnitude of these codon-expression agreements, one must be

aware that that the estimates of correlation may be conservative, as gene expression may match codon usage more or less strongly, depending on the conditions when expression was measured. The best match would be obtained under the conditions that were dominant during the periods of rapid competitive growth in the evolutionary history of the organism (Wagner 2000). A case in point are the microarray measurements taken under conditions of stress or nutritional deficiency, for instance in a minimal medium, which have been shown to correlate less well with codon usage in *Bacillus subtilis* and *Escherichia coli* (Supek and Vlahovicek 2005), when compared to conditions more conducive to rapid growth (rich media).

An early investigation of codon usage in yeast genes (Bennetzen and Hall 1982) proposed that, in cases where the genome encodes only a single tRNA species for an amino acid, the translationally optimal codons would be the ones that match the tRNA anticodon by canonical Watson-Crick base pairing (without wobble). When multiple tRNA species are available for the amino acid, the anticodons on very abundant tRNAs would correspond to optimal codons – in this context, the genomic tRNA gene count is frequently used as a proxy for tRNA abundance (dos Reis et al. 2004; Tuller et al. 2010).

In prokaryotes, the presence or absence of tRNA genes with specific anticodons given in the GtRNAdb database (Chan and Lowe 2009) indicates that the optimal codons for two-fold degenerate amino acids are almost always either C/A-ending, or undefined if the genome contains tRNA genes with both anticodons for an amino acid. Supek et al. (2010) found that the OCU genes prefer such putatively optimal codons 6.2x more frequently than the suboptimal ones in Bacteria and 3.1x more frequently in Archaea. The examples of amino acids and genomes where this regularity is reversed have lead the authors to speculate that these exceptions stem from chemical modifications of nucleosides in the tRNA that modulate the codon-anticodon interaction (Agris 2004; Agris et al. 2007). Unfortunately, the tRNA nucleoside modifications are currently fully known only for few organisms (Grosjean et al. 2009); future experimental data will shed some light on how the modifications relate to the OCU-preferred codons. Hershberg and Petrov (2009) discussed the choice of optimal codons in genomes and concluded that the main determinant of codon optimality was the direction of nucleotide substitution patterns evident in the overall genomic G+C content. Moreover, they also found that the gene set with the most highly biased codon usage in a genome was also typically enriched with RPs and translation elongations factors, and that this codon usage bias cannot be explained only by differences in composition of non-coding DNA. This study, together with the results of Supek et al. (2010), provides sufficient evidence for the notion of translationally selected codon usage as a prevalent phenomenon among prokaryotic genomes.

BOX 1



The Random Forest classification model is essentially a collection of decision tree classifiers, where each decision tree is constructed by recursively subdividing the data by attribute value tests (into 'nodes') in order to reduce the entropy of the class label within the resulting subdivisions ('branches'). An example of a decision tree that discriminates between *E. coli* ribosomal protein genes and all other *E. coli* genes is given in this box – in this example, classes are "1" for ribosomal, "0" for other genes. RF is a particular kind of an ensemble classifier (modified bootstrap aggregation method), different from a simple collection of decision trees in two ways: (i) each of the trees is constructed not from the full dataset, but from a bootstrap sample of the dataset; (ii) the choice of attributes at each node is artificially limited to reduce correlation between the individual trees, which has been found to help the predictive performance of the entire model. The final per-class probabilites of a RF model are obtained by averaging the prediction of individual trees ('voting') where each tree has equal weight. In addition to classification, the RF algorithm has some extra features such as: quantifying attribute importance, computing pairwise distances of all instances (see Fig. 2D), and providing a computationally efficient estimate of crossvalidation error called the out-of-bag error; see Breiman (2001) or http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm for more information.

References

Agris, P.F., 2004. Decoding the genome: a modified view. Nucleic Acids Research, 32(1), 223-238.

Agris, P.F., Vendeix, F.A. & Graham, W.D., 2007. tRNA's Wobble Decoding of the Genome: 40 Years of Modification. Journal of Molecular Biology, 366(1), 1-13.

Angellotti, M.C., Bhuiyan, S.B., Chen, G. & Wan, X.-F., 2007. CodonO: codon usage bias analysis within and across genomes. Nucleic Acids Research, 35(suppl_2), W132-136.

Banerjee, T., Basak, S., Gupta, S.K. & Ghosh, T.C., 2004. Evolutionary forces in shaping the codon and amino acid usages in Blochmannia floridanus. Journal of Biomolecular Structure & Dynamics, 22(1), 13-23.

Bennetzen, J.L. & Hall, B.D., 1982. Codon selection in yeast. The Journal of Biological Chemistry, 257(6), 3026-3031.

Breiman, L., 2001. Random Forests. Machine Learning, 45(1), 5-32.

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet,P., Gonnet, G. & Barral, Y., 2010. A Role for Codon Order in Translation Dynamics. Cell, 141(2), 355-367.

Carbone, A., Képès, F. & Zinovyev, A., 2005. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. Molecular Biology and Evolution, 22(3), 547-561.

Carbone, A., 2006. Computational prediction of genomic functional cores specific to different microbes. Journal of Molecular Evolution, 63(6), 733-746.

Carbone, A. & Madden, R., 2005. Insights on the evolution of metabolic networks of unicellular translationally biased organisms from transcriptomic data and sequence analysis. Journal of Molecular Evolution, 61(4), 456-469.

Chan, P.P. & Lowe, T.M., 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Research, 37(Database issue), D93-97.

Charles, H., Calevro, F., Vinuelas, J., Fayard, J.-M. & Rahbe, Y., 2006. Codon usage bias and tRNA over-expression in Buchnera aphidicola after aromatic amino acid nutritional stress on its host Acyrthosiphon pisum. Nucleic Acids Research, 34(16), 4583-4592.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L. & McAdams, H.H., 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proceedings of the National Academy of Sciences of the United States of America, 101(10), 3480-3485.

Daubin, V. & Perriere, G., 2003. G+C3 Structuring Along the Genome: A Common Feature in Prokaryotes. Molecular Biology and Evolution, 20(4), 471-483.

Dethlefsen, L. & Schmidt, T., 2005. Differences in codon bias cannot explain differences in translational power among microbes. BMC Bioinformatics, 6(1), 3.

Dittmar, K.A., Sørensen, M.A., Elf, J., Ehrenberg, M. & Pan, T., 2005. Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO reports, 6(2), 151-157.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

Grocock, R.J. & Sharp, P.M., 2002. Synonymous codon usage in Pseudomonas aeruginosa PA01. Gene, 289(1-2), 131-139.

Grosjean, H., de Crécy-Lagard, V. & Marck, C., 2009. Deciphering synonymous codons in the three domains of

life: Co-evolution with specific tRNA modification enzymes. FEBS Letters, 584(2): 252-264.

Herbeck, J.T., Wall, D.P. & Wernegreen, J.J., 2003. Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont Wigglesworthia. Microbiology, 149(Pt 9), 2585-2596.

Hershberg, R. & Petrov, D.A., 2009. General Rules for Optimal Codon Choice. PLoS Genetics, 5(7), e1000556.

Hershberg, R. & Petrov, D.A., 2010. Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLoS Genetics*, 6, e1001115.

Hildebrand, F., Meyer, A. & Eyre-Walker, A., 2010. Evidence of Selection upon Genomic GC-Content in Bacteria. *PLoS Genetics*, 6, e1001107.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kerner, M. & Frishman, D., 2008. Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics, 9(1), 102.

Kanaya, S., Yamada, Y., Kudo, Y. & Ikemura, T., 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene, 238(1), 143-155.

Karlin, S., Brocchieri, L., Campbell, A., Cyert, M. & Mrázek, J., 2005. Genomic and proteomic comparisons between bacterial and archaeal genomes and related comparisons with the yeast and fly genomes. Proceedings of the National Academy of Sciences of the United States of America, 102(20), 7309-7314.

Karlin, S. & Mrazek, J., 2000. Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes. Journal of Bacteriology, 182(18), 5238-5250.

Karlin, S., Mrazek, J., Ma, J. & Brocchieri, L., 2005. Predicted highly expressed genes in archaeal genomes. Proceedings of the National Academy of Sciences of the United States of America, 102(20), 7303-7308.

Kimchi-Sarfaty, C., Oh, J.M., Kim, I.-W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V. & Gottesman, M.M., 2007. A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. Science, 315(5811), 525-528.

Knight, R.D., Freeland, S.J. & Landweber, L.F., 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biology, 2(4), RESEARCH0010.

Kudla, G., Murray, A.W., Tollervey, D. & Plotkin, J.B., 2009. Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science, 324(5924), 255-258.

Lafay, B., Atherton, J.C. & Sharp, P.M., 2000. Absence of translationally selected synonymous codon usage bias in Helicobacter pylori. Microbiology, 146(4), 851-860.

Lawrence, J.G. & Ochman, H., 1997. Amelioration of Bacterial Genomes: Rates of Change and Exchange. Journal of Molecular Evolution, 44(4), 383-397.

Lobry, J.R. & Sueoka, N., 2002. Asymmetric directional mutation pressures in bacteria. Genome Biology, 3(10), RESEARCH0058.

Man, O. & Pilpel, Y., 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nature Genetics, 39(3), 415-421.

McDonald, J., 2009. Sign Test. In Handbook of Biological Statistics. Sparky House Publishing, pp. 202-6.

McInerney, J.O., 1998. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proceedings of the National Academy of Sciences of the United States of America, 95(18), 10698-

10703.

Mrázek, J., Spormann, A.M. & Karlin, S., 2006. Genomic comparisons among gamma-proteobacteria. Environmental Microbiology, 8(2), 273-288.

Najafabadi, H.S., Lehmann, J. & Omidi, M., 2007. Error minimization explains the codon usage of highly expressed genes in Escherichia coli. Gene, 387(1-2), 150-155.

Oresic, M. & Shalloway, D., 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. Journal of Molecular Biology, 281(1), 31-48.

Parmley, J.L. & Huynen, M.A., 2009. Clustering of Codons with Rare Cognate tRNAs in Human Genes Suggests an Extra Level of Expression Regulation. PLoS Genetics, 5(7), e1000548.

Perriere, G. & Thioulouse, J., 2002. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Research, 30(20), 4548-4555.

Puigbo, P., Guzman, E., Romeu, A. & Garcia-Vallve, S., 2007. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Research, 35(suppl_2), W126-131.

dos Reis, M., Savva, R. & Wernisch, L., 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Research, 32(17), 5036-5044.

Resch, A.M., Carmel, L., Marino-Ramirez, L., Ogurtsov, A.Y., Shabalina, S.A., Rogozin, I.B. & Koonin, E.V. et al., 2007. Widespread Positive Selection in Synonymous Sites of Mammalian Genes. Molecular Biology and Evolution, msm100.

Rispe, C., Delmotte, F., van Ham, R.C.H.J. & Moya, A., 2004. Mutational and Selective Pressures on Codon and Amino Acid Usage in Buchnera, Endosymbiotic Bacteria of Aphids. Genome Research, 14(1), 44-53.

Rocha, E.P.C., 2004. The replication-related organization of bacterial genomes. Microbiology, 150(Pt 6), 1609-1627.

Rocha, E.P.C. & Danchin, A., 2002. Base composition bias might result from competition for metabolic resources. Trends in Genetics, 18(6), 291-294.

Rocha, E.P.C. & Feil, E.J., 2010. Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria? *PLoS Genetics*, 6, e1001104.

Sharp, P.M. & Li, W.H., 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research, 15(3), 1281–1295.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F. & Sockett, R.E., 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Research, 33(4), 1141-1153.

Stoletzki, N. & Eyre-Walker, A., 2007. Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. Molecular Biology and Evolution, 24(2), 374-381.

Supek, F., Škunca, N., Repar, J., Vlahoviček, K. & Šmuc, T., 2010. Translational selection is ubiquitous in prokaryotes. PLoS Genetics, 6(6): e1001004.

Supek, F. & Šmuc, T., 2010. On relevance of codon usage to expression of synthetic and natural genes in Escherichia coli. Genetics, 185(3), 1129-1134.

Supek, F. & Vlahovicek, K., 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics, 6(1), 182.

Supek, F. & Vlahovicek, K., 2010. Correction: Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics, 11, 463.

Supek, F. & Vlahovicek, K., 2004. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics, 20(14), 2329-2330.

Suzuki, H., Saito, R. & Tomita, M., 2005. A problem in multivariate analysis of codon usage data and a possible solution. FEBS Letters, 579(28), 6499-6504.

Tegel, H., Tourle, S., Ottosson, J. & Persson, A., 2010. Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). Protein Expression and Purification, 69(2), 159-167.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I. & Pilpel, Y., 2010. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell, 141(2), 344-354.

Urrutia, A.O. & Hurst, L.D., 2003. The Signature of Selection Mediated by Expression on Human Genes. Genome Research, 13(10), 2260-2264.

Wagner, A., 2000. Inferring Lifestyle from Gene Expression Patterns. Molecular Biology and Evolution, 17(12), 1985-1987.

Weiner, R.M., Taylor, L.E., Henrissat, B., Hauser, L., Land, M., Coutinho, P.M., Rancurel, C., Saunders, E.H., Longmire, A.G., Zhang, H. et al., 2008. Complete genome sequence of the complex carbohydrate-degrading marine bacterium, Saccharophagus degradans strain 2-40 T. PLoS Genetics, 4(5), e1000087.

Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J., Gustafsson, C., 2009a. Design Parameters to Control Synthetic Gene Expression in Escherichia coli. PLoS ONE, 4(9), e7002.

Welch, M., Villalobos, A., Gustafsson, C., Minshull, J., 2009b. You're one in a googol: optimizing genes for protein expression. Journal of the Royal Society Interface, 6 Suppl 4, S467-476.

Wright, F., 1990. The 'effective number of codons' used in a gene. Gene, 87(1), 23-29.

Xia, X., 1998. How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics, 149(1), 37-44.

Zeldovich, K.B., Berezovsky, I.N. & Shakhnovich, E.I., 2007. Protein and DNA Sequence Determinants of Thermophilic Adaptation. PLoS Computational Biology, 3(1), e5.

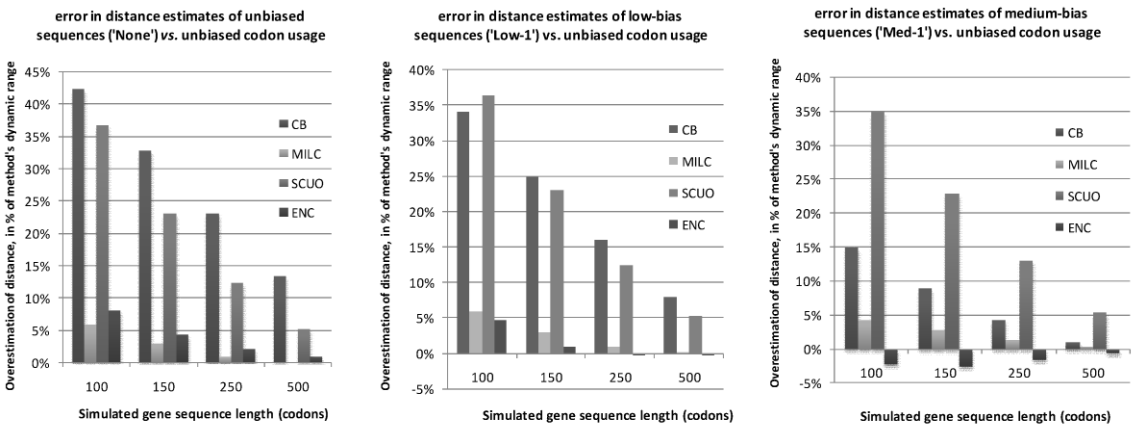


Figure 1. Mis-estimation of distances in codon usage for various methods. The values of a distance measure at varying simulated sequence lengths (on the x axis) were subtracted from the values calculated for 2500 codons which are used as a gold standard – it was assumed there is no length-related bias in very long genes. The overestimation of distances (shown on y axis) were computed by calculating the mean distance for 10000 randomly generated sequences per method per length, and are expressed as percentages of the measures' dynamic ranges; see Supek & Vlahovicek (2005) for a description of the methods, and the codon frequencies used for the three cases 'None', 'Low-1' and 'Med-1'. The CB and MILC measures are as described in Section 2.1 of this chapter. SCUO is a normalized measure of Shannon entropy of a gene's codon usage (Angellotti et al. 2007). ENC is the widely used method called 'effective number of codons' (Wright 1990). Unlike the CAI, MILC or CB, SCUO and ENC are special cases of codon distance measure which can only be used to compare the codon usage of a gene to the unbiased codon frequencies, and not to any arbitrary set of codon frequencies.

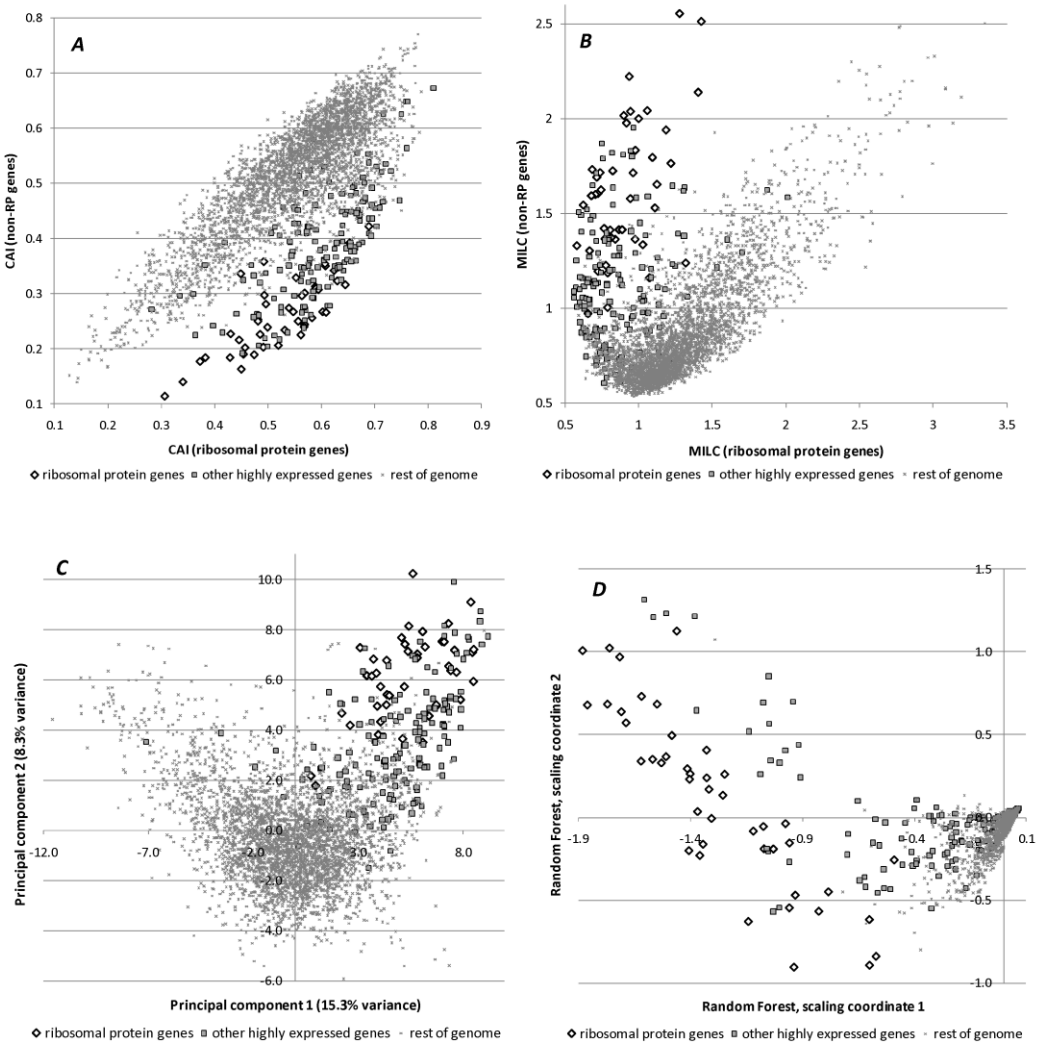


Figure 2. Visualizations of codon usage of all genes in the *E. coli* K12 genome; only genes at least 80 codons long are shown. The "other highly expressed genes" class consists of genes coding for the 200 most abundant *E. coli* cytoplasmic proteins according to Ishihama et al. (2008), excluding the ribosomal protein genes. (**A**, **B**) x axis shows the distance of a gene's codon usage from the ribosomal protein genes, y axis shows the distance from the average codon frequencies of all non-ribosomal protein genes. Panel A uses the 'codon adaptation index' similarity measure, while panel B uses the MILC distance measure. (**C**) A principal components plot of the genes' codon frequencies. (**D**) A principal components plot of the Random Forest algorithm's internal representation of the data as a matrix of gene-gene distances derived from the structure of the decision trees that constitute the RF classification model; see Breiman (2001) for more information.

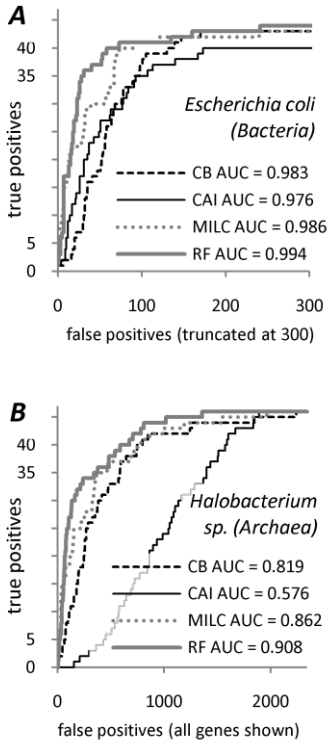


Figure 3. Receiver operating characteristic (ROC) curves showing the accuracy of various mathematical methods in discerning the ribosomal protein (RP) genes from all other protein coding genes by their codon frequencies. The area under ROC curve (AUC) statistic is higher in more accurate classifiers, ranging from 0.5 (random guessing) to 1.0 (perfectly accurate). All results are from ten-fold crossvalidation. On a set of 461 prokaryotic genomes, RF outperformed the distance measures: CAI in 459 genomes, CB in 440 genomes and MILC in 351 genomes (Supek et al. 2010).

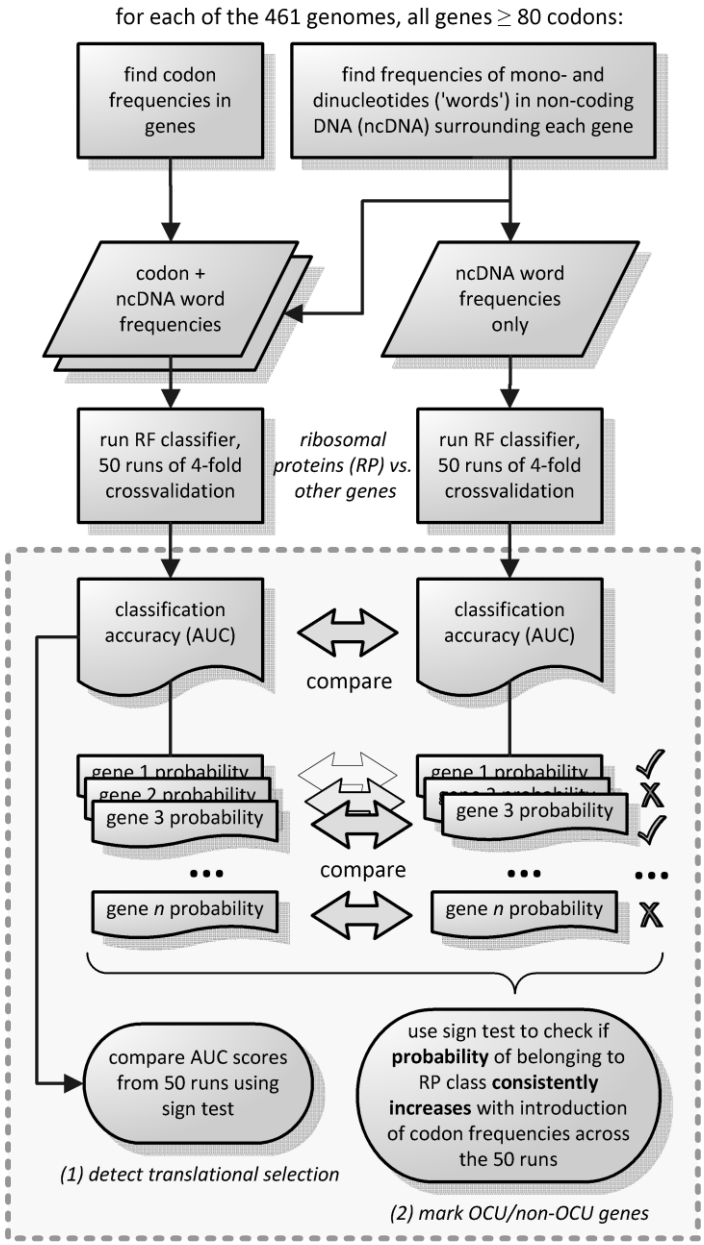


Figure 4. A flowchart representation of the Random Forest classifier-based computational framework (Supek et al. 2010) used to detect whether translational selection acts in a specific genome, and if so, which genes it affects.

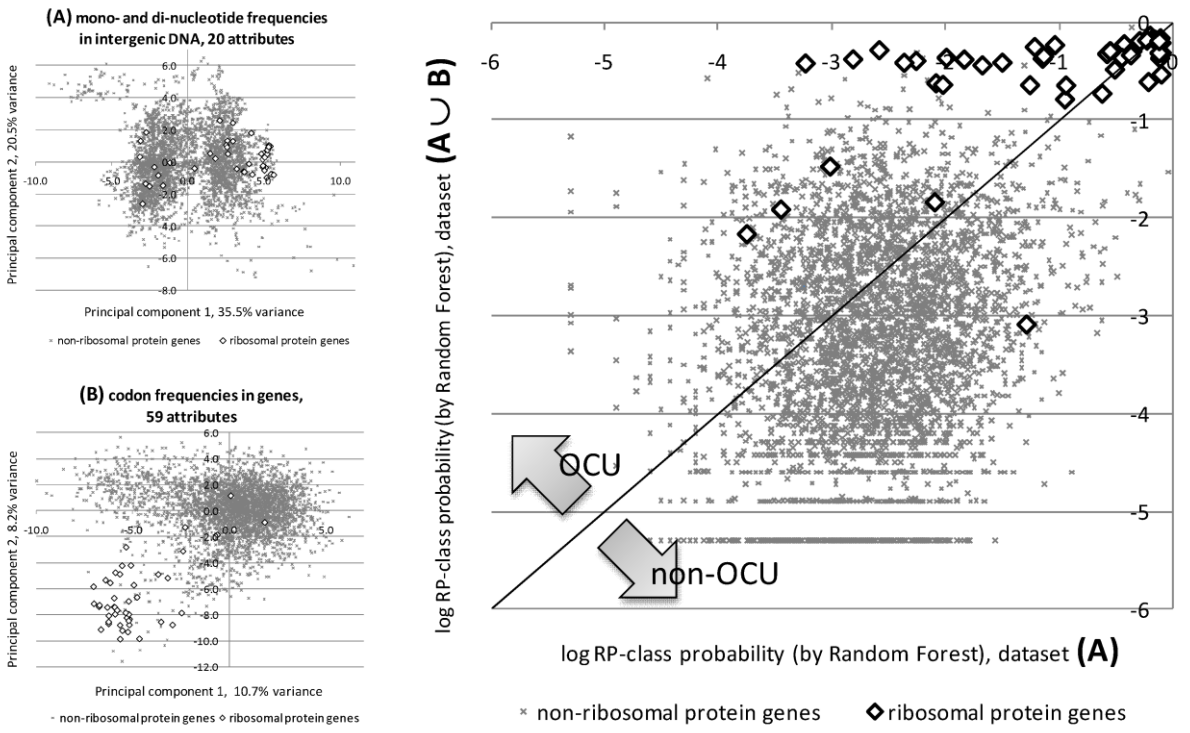


Figure 5. An illustration of the reasoning behind using a classifier to control for a confounding factor; in this instance, the Random Forest classifier is used to decide whether a gene's codon usage similarity to the RP genes can be explained by the similarity in the local (or strand-specific) nucleotide substitution patterns that are evident from the composition of intergenic DNA. Figures show actual data from the *B. subtilis* genome (all genes ≥ 80 codons). Subpanels **A** and **B** show a principal components plot of *B. subtilis* genes represented by: (**A**) the composition of intergenic DNA surrounding genes within a 10kb window, or (**B**) codon frequencies. The main panel (right-hand side) is a scatter plot of the Random Forest probabilities without (*x*-axis) against those including (*y*-axis) codon frequencies. A gene is assigned the OCU ("optimized codon usage") label if it increases in the RP-class probability after addition of codon frequencies. The diagonal line that divides OCU from non-OCU is for illustrative purposes only: The actual procedure to determine OCU (Fig. 4) involves checking for a consistent increase in Random Forest probability over 50 runs of crossvalidation, whereas the points in this Figure show only averages of probabilities over the 50 runs, while between-run variability is not depicted in the plot for clarity's sake. In effect, fewer genes in *B. subtilis* will be declared OCU than are above the diagonal line in the plot because of this variability.