Codon Evolution

Mechanisms and Models

EDITED BY

Gina M. Cannarozzi
University of Bern, Switzerland

Adrian Schneider
University of Utrecht, The Netherlands

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2012

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Library of Congress Control Number: 2011944051

Typeset by SPI Publisher Services, Pondicherry, India
Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

ISBN 978–0–19–960116–5

1 3 5 7 9 10 8 6 4 2

CHAPTER 13

Measuring codon usage bias

Alexander Roth, Maria Anisimova, and Gina M. Cannarozzi

13.1 Introduction

In protein-coding genes, the genetic code defines the translational mapping from nucleotide triplets (or codons) to amino acids. Synonymous codons translate to the same amino acid and are indistinguishable at the protein level. However, most genes and organisms do not use synonymous codon uniformly; certain synonymous codons are used preferentially, a phenomenon called codon usage bias (or shorter, codon bias). In this chapter we discuss the biological causes and the statistical measures of codon usage bias. Given the large number of existing codon bias measures, the specifics, utility, and comparative performance of different approaches may be often elusive, especially to a novice in the field.

Here we review, classify and compare most codon bias measures proposed to date.

13.2 Causes of codon usage bias

Most protein-coding DNA sequences use synonymous codons with very different frequencies. The first reports of non-uniform codon usage date to as early as four decades ago. Clarke (1970) and later Ikemura (1981a), and Akashi (1994), suggested that codon usage adapted to match an organism's tRNA pool. Observed differences in codon bias between species are a result of different evolutionary forces acting on the choice of codons (Ikemura, 1981a). Codon usage can differ widely not only between organisms, but also within a genome. For example, eukaryotic genomes are known to exhibit heterogeneous nucleotide content creating an isochore structure. Isochores are long DNA segments with relatively homogeneous GC content (Macaya *et al.*, 1976). Isochores are typically rich in protein-coding genes and consequently affect codon usage in genes within isochores.

While codon bias does not directly influence the protein sequence, it may have important impact on the protein product and cellular processes. However, the exact mechanisms driving synonymous variation are still not well understood. There exists a variety of hypotheses to explain mechanisms responsible for codon bias. But the relative importance and the possible interplay between the many explanations are poorly understood and the variation in synonymous codon usage continues to puzzle molecular and evolutionary biologists.

On a mechanistic level, codon usage is shaped by the balance between mutational biases and natural selection (see, for example: Duret, 2002, Hershberg and Petrov, 2008) but estimating the relative contributions of selection versus mutational biases can be difficult and varies between eukaryotes and prokaryotes. The influences of these two factors are discussed in the following sections.

13.2.1 Mutational biases affecting codon usage

Codon bias may result from mutational biases alone. Mutational biases are caused by underlying mechanisms that favour certain types of mutations, such as chemical decay of nucleotide bases Kaufmann and Paules (1996), non-uniform DNA repair, and non-random replication errors. The result is biased codon and amino acid usage (Knight *et al.*, 2001). Mutational biases are neutral (do not affect fitness) and typically act globally on all DNA sequences of a given organism.

For example, the dinucleotides TA and CG (known as TpA and CpG) occur at a lower frequency than expected based on the nucleotide frequencies (Kaufmann and Paules, 1996). In eukaryotes the cytosine in CG dinucleotides is easily methylated; the methylated form of cytosine then spontaneously deaminates into thymine. As the thymine is not detected by the DNA repair mechanisms, these errors are not corrected. In the human genome, the CG dinucleotide occurs at only 21% of the frequency expected by random chance given the frequencies of C and G (Lander *et al.*, 2001).

In most life forms TA nucleotides are also found less frequently than expected based on the nucleotide frequencies. This is thought to be due to the avoidance of the stop codons TAA and TAG, as well as the avoidance of UA in mRNA, which is susceptible to RNAse activity (Beutler *et al.*, 1989).

Many mutations originate from non-random mismatch repair following replication errors and methylation. Such strand-specific mutational biases result from different fidelities of replication of the leading and lagging strands. Such asymmetric mutation rates of the leading and lagging strands are found in both bacteria (Lobry, 1996; Fijalkowska *et al.*, 1998; McLean *et al.*, 1998) and eukaryotes (Pavlov *et al.*, 2003; Kunkel *et al.*, 2003).

Global species differences in codon usage are typically explained by mutational biases.

13.2.2 Selection affecting codon usage

In contrast to purely mutational mechanisms, selective forces may also influence synonymous codon usage. Codon bias caused by selection may be specific to genes or even codon positions, where it can induce more efficient or accurate translation or protein folding. These patterns can be observed by comparing coding and non-coding regions of DNA.

Selection acts upon the changes created by neutral mutational processes and may originate from many sources and vary in strength. For example, in some genes the synonymous codon usage is primarily shaped by translational selection, while in others, it may be shaped by mutational bias. Different types of selection act on different levels.

At the DNA level there are patterns that are avoided or preferred. These can be related to functional elements of DNA such as DNA packing of nucleosomes and other varying nucleotide distributions along the genome.

At the RNA level, selection for effective transcription (Xia, 1996) has been proposed, in which mRNA with more abundant nucleotides are transcribed more quickly. In these cases, the codons are enriched in common nucleotides. Selection can also take place at the mRNA level, where some patterns are avoided or preferred, to influence mRNA folding and decay. Codon bias also correlates well with mRNA levels. This is an indication that there is a global optimization of minimizing the time the ribosomes are engaged in translation of the mRNA. Codons evolving under positive selection have corresponding tRNAs in larger quantities and possibly bind to the mRNA at the ribosome more rapidly (Ran and Higgs, 2010).

At the translation level, an important factor determining protein yield is the initiation rate. Experiments in which the synonymous codons have been randomized, show that strong folding in the region around the ribosome-binding site inhibits the initiation of translation by making the binding site inaccessible to the ribosome (Kudla *et al.*, 2009). Such patterns are selected against in natural populations. Consequently, the most likely candidates for codons under selection are those that can influence mRNA folding by inducing strong secondary structures, in particular those close to the ribosome binding site (Kudla *et al.*, 2009). In addition, the splicing of mRNA requires specified nucleotide motifs. Synonymous mutations on such sites can introduce cryptic splice sites and have large effects on the phenotype (Pagani and Baralle, 2004). The use of preferred or rare codons affects the rate of translation and elongation, and consequently can influence the co-translational protein folding (Kimchi-Sarfaty *et al.*, 2007).

However, the main factor influencing codon usage is selection for optimal translation at the level of protein synthesis. Highly expressed genes are enriched in the most frequent ('optimal') codons. Genes that are less abundant, often show milder codon preference. In several organisms there is a significant correlation between codon usage bias and protein abundance. Codon choice is constrained by tRNA availability. Transfer RNA availability at elongation is an important factor contributing to the choice of codons. Codons corresponding to rare tRNA species can induce long waiting times and stall elongation at such positions.

There is a strong correlation between the codon bias and the gene copy number of the corresponding tRNA (Ikemura, 1981a). In addition to being translated quickly, fidelity of translation is also important, in particular for abundant proteins. Misincorporations can have dire consequences and cause protein misfolding (Drummond *et al*., 2006). A large error rate in the synthesis of crucial proteins, means that a large fraction of the proteins produced are non-functional and must be catabolized. This can carry a high cost for the organism.

Selection for optimal translation is most effective in organisms with large effective population sizes (Bulmer, 1987). Indeed, strong codon bias was reported in the genomes of *E. coli* and yeast, which have large population sizes. Mammalian genomes have a small effective population size and there is much less evidence for selection. The codon usage in mammals is correlated to the local chromosomal nucleotide content of flanking regions and introns and mutational bias appears to be dominating the evolution of codon usage. However, there is also evidence that synonymous codon usage in mammals is not neutral (Chamary *et al*., 2006; Kimchi-Sarfaty *et al*., 2007).

There is a bias in the choice of pyrimidine bases at the third position of synonymous codons such that the codon–anticodon binding has an intermediate strength in the choice of pyrimidine bases at the third position of synonymous codons (Grosjean *et al*., 1978; Ikemura, 1981a). If the two first positions of a codon are strong nucleotides (S = G or C, three hydrogen bonds) then the third codon position has more often a weak nucleotide (W = A or T, two hydrogen bonds). The other scenario is also true, if weak nucleotides are found at the first two positions, then the strong nucleotides are more common at the third position. This bias is independent of amino acid composition. Selection for uniform binding properties of tRNA are likely to be beneficial for translation, by preventing stalling on strong codons and insufficient binding of weak codons.

Moreover, the choice of a codon in a new instance of a synonymous codon at a position downstream may be influenced by a previous occurrence, implicating the order of synonymous codons as a factor. It was found that use of a codon decoded by the same isoacceptor tRNA is preferred to other synonymous codons at subsequent occurrences of the same amino acid (Cannarozzi *et al*., 2010). As two tRNAs are simultaneously bound to the ribosome only briefly, tRNA reuse is possible at the +2 codon (Uemura *et al*., 2010). Codon bias in different gene regions appears to be under different selective constraints, due to the early phase of translation (Karlin *et al*., 1998). The first 30 to 50 codons are translated with low efficiency. In order to reduce traffic congestion of ribosomes, they form a 'ramp' to reduce the speed of translation in the early stage of the elongation cycle (Tuller *et al*., 2010). It is also possible that the codon usage acts as an extra level of regulation to fine-tune the levels of protein abundance (Begley *et al*., 2007; Parmley and Huynen, 2009), through the usage of regulatory codons. Further evidence for this is that the levels of protein abundance for orthologs among species are surprisingly more conserved than the mRNA levels (Weiss *et al*., 2010). Also, some metabolic genes are enriched in a subset of non-common codons. These codons are decoded by tRNAs that, upon amino acid starvation, are preferentially recharged over other isoacceptor tRNA (Elf *et al*., 2003).

Other constraints on the amino acid level may shape the codon composition. There is a relationship between codon choices and the secondary structure of proteins (Adzhubei *et al*., 1996). For example, membrane proteins have a much higher incidence of alpha helixes, which bias the choice of codons to G-ending codons (de Miranda *et al*., 2000). In eukaryotic repetitive elements, there is a small subset of codons being reiterated within homo-peptides (Faux *et al*., 2007). Synonymous codon usage biases may be associated with various other biological factors, such as: genome size (dos Reis *et al*., 2004), gene length (Duret and Mouchiroud, 1999), amino acid composition (D'Onofrio *et al*., 1999), local protein structure (Saunders and Deane, 2010), codon context, biased gene conversion (Harrison and Charlesworth, 2011), recombination rate (Zhou *et al*., 2005), gene translation initiation signal (Qin *et al*., 2004), and length of 3'-UTR. Global codon bias has been shown to correlate with GC content (Ikemura, 1981a; Kanaya *et al*., 2001; Knight *et al*., 2001), tRNA content (Kanaya *et al*., 2001), and organism growth temperature (Lao and Forsdyke, 2000), although the latter may influence selective forces on both mRNA structure (Lao and Forsdyke, 2000) and codon bias (Lynn *et al*., 2002).

It is very difficult to know how all the evolutionary and functional constraints interact and the causality is often difficult to infer. For example, while CG content may cause codon bias, codon bias may act also in the opposite direction, influencing the nucleotide composition. It is unclear how the two evolutionary processes of change in codon usage and nucleotide compositions interact. The causality of codon patterns continues to puzzle evolutionary and molecular biologists.

Hopefully, new experimental technology will help to disentangle some effects related to codon bias. Here we continue by reviewing the wealth of statistical measures that have been proposed to measure codon bias.

13.3 Applications for indices of codon usage bias

Codon usage indices are generated by a dedicated function that maps some aspect of codon usage, often the codon frequencies, to a single number. Codon usage indices have found a number of applications, for example, several indices were originally developed to assess the likelihood of being in a certain protein-coding reading frame, i.e. to recognize protein-coding genes. Open reading frames (ORFs) containing a high incidence of rare codons are unlikely to encode a protein, even weakly expressed genes tend to have far fewer rare codons than expected from the genomic frequencies. This phenomenon has been used both to identify pseudo-genes and to detect DNA sequencing errors resulting in the insertion or deletion of bases within a coding sequence (Gribskov *et al*., 1984), as well as to identify spurious ORFs (long sequences that may be coding but have occurred by chance), as they tend to have codon usage different from that of verified ORFs (Ghaemmaghami *et al*., 2003).

Codon usage bias often differs significantly among organisms. Hence, indices can be used for detecting lateral gene transfer (Carbone *et al*., 2003; Sugaya *et al*., 2004; Cortez *et al*., 2005; Tsirigos and Rigoutsos, 2005; Bodilis and Barray, 2006) and for the comparison of codon usage in different organisms to study functional conservation of gene expression across organisms (Lithwick and Margalit, 2005).

As codon usage frequency has been shown to correlate with protein and mRNA abundance in many organisms, indices are also commonly used to predict and optimize protein expression levels, either in the native organism or for heterologous expression of genes in foreign hosts. Codon optimized genes are important both for biotechnological production and for DNA vaccines (Ruiz *et al*., 2006). Verification that the codon usage of the heterologous protein is similar to that of the host organism is critical, since rare codons can have a detrimental effect on protein yield. In addition to avoiding rare codons, there are several other factors that must be taken into account for the optimization of protein yield such as, translation initiation regions and mRNA structural elements.

13.4 Previous studies of codon usage indices

The correlation between several different indices and experimental data, such as mRNA expression levels or protein concentration, has been examined in many studies (Comeron and Aguadé, 1998; Coghlan and Wolfe, 2000; Goetz and Fuglsang, 2005; Supek and Vlahovicek, 2005; Tuller *et al*., 2007; Suzuki *et al*., 2008), since the prediction of expression levels is the aim of many researchers. As most studies agree that highly expressed genes are associated with codon usage biased towards usage of the most frequent codons, many indices are built on codon usage frequencies. The most commonly used index is the Codon Adaptation Index (CAI by Sharp and Li, 1987), which has consistently found use as a predictor of gene expression levels. Since the introduction of the CAI, many new measures, which often compare favourably with the CAI, have been developed and are described in the following sections.

These studies usually examine the correlation of the indices to absolute concentrations of mRNA and protein, but not to the protein synthesis rate. It can be argued that the underlying evolutionary pressure for high expression has also forced mRNA and protein levels to correlate with the protein synthesis rate. Correlating indices to the rate of protein synthesis is an alternative. Unfortunately, very few whole-genome datasets of protein half-times, necessary for the prediction of synthesis rate, are available (Belle *et al*., 2006).

13.5 Measures of codon bias

A large number of indices for measuring codon usage bias have been proposed; some of the most relevant and non-redundant ones are discussed here.

There are several ways of classifying codon bias indices. For example, one group of indices measures departure from the expected codon distribution (based on nucleotide frequencies). Another group measures closeness to a hypothetical optimal state of codons (or genes) and usually compares the codon usage of a gene to the preferred codon usage of a group of reference genes. It is possible to further classify different groups in the reference class. References have been made to optimal codons, highly expressed genes, a defined gene class, or all genes in the genome. Not all indices are easily classifiable; for example, several of the indices based on deviation from expected value can be modified to allow comparison to a reference set of highly expressed genes. Herein, we have chosen to classify indices based on historical and methodological similarities.

All amino acids with more than one codon can show a bias. In order to create an index, the contribution of each amino acid has to be combined in a sensible way; for example, weighing each amino acid contribution according to their frequency in the gene. The degree of codon degeneracy (one, two, three, four, or six codons per amino acid) must be considered and the one-codon amino acids (Met and Trp) excluded. Start codons should be considered separately, since these are often read by a special initiator tRNA and are often excluded by many measures. Stop codons are also often excluded for the same reasons and should not be considered part of the coding sequence. Many indices have difficulties in computing accurate values for short sequences; therefore it is recommended to avoid or be very cautious with sequences shorter than 80–100 codons.

In this text, the notation of several indices has been changed from the original publication, in order to create a uniform notation and to better see and understand the relationships between indices.

Table 13.1 Frequently used symbols

C	entire set of codons
A	set of amino acids
c	index for codons
a	index for amino acids
C_a	codons used by amino acid a
o_{ac}	count of codon c for amino acid a
k_a	number of synonymous codons of amino acid a (codon degeneracy)
L	length of the sequence
F_a	frequency of amino acid a
f_{ac}	frequency of codon c encoding amino acid a
r_{ac}	relative synonymous codon usage (RSCU) for codon c and amino acid a

First, rather than using i for indexing codon and j for amino acids, the subscripts c is used for the synonymous codons of an amino acid, while a is used for the amino acid. For example, o_{ac} is the observed count of synonymous codon c of the amino acid a. When indexing all 64 codons, c is again used. A single index c points to any of the 64 codons. The observed number of codons can be denoted by the vector $\mathbf{o} = [o_1, \ldots, o_{64}]$ of length 64, where the elements are of the number of occurrences of the codons. The number of codon occurrences o is also indexed by the codon names; for example, o_{NNG} is the number of G-ending codons in a sequence. The entire set of codons in an analysis is denoted by C. The subset of synonymous codons used by amino acid a is denoted by C_a. The number of synonymous codons of an amino acid is k_a, also referred to as codon degeneracy or codon redundancy. The length of the sequence in number of amino acids is L. The set of amino acids used by an index is denoted by A, e.g. A_1 is Alanine, etc. The usage of hats for estimates (e.g. $\hat{F}$) is avoided, since it is clear from the context when the measures are estimates.

For an objective way of quantifying the performance of indices, a framework for incorporating all aspects of protein synthesis from many sources is desirable. Implementations of indices exist in various packages and as stand-alone programs. The program suite CodonW (Peden, 2000) has implemented and documented some of the existing indices. Unfortunately, it appears that there is little momentum in the development of CodonW, even though it is an open source project. There are several other programs for computing codon indices: INCA by Supek and Vlahovicek (2004) and GCUA by McInerney (1998). Implementations also exist in libraries of common programming languages, such as, BioPerl by Brenner *et al*. (2002), seqinR by Charif *et al*. (2005), and EMBOSS by Rice *et al*. (2000). Easy access under one framework to the bulk of codon usage bias indices would facilitate comparison, benchmark studies and performance analysis.

13.5.1 Relative codon frequencies

Many indices often require the codon counts to be normalized into codon frequencies to remove the dependence on gene length. Frequencies can be computed in a number of ways. The simplest way is to normalize by the sum of all codons in the vector:

$$g_c = \frac{o_c}{\sum\limits_{c \in C} o_c}, \tag{13.1}$$

where g denotes a global frequency.

This can also be expressed in the double-index notation:

$$g_{ac} = \frac{o_{ac}}{\sum\limits_{a \in A} \sum\limits_{c \in C_a} o_{ac}}. \tag{13.2}$$

This normalization has the problem that it will overweight frequent amino acids. It is therefore usually better to normalize within each amino acid separately to avoid the confounding influence of amino acid content:

$$f_{ac} = \frac{o_{ac}}{\sum\limits_{c \in C_a} o_{ac}} \tag{13.3}$$

where f denotes the frequency within amino acid a.

The relative synonymous codon usage (RSCU) compensates for both the different number of synonymous codons for the various amino acids, as well as for the differing amino acid frequencies:

$$r_{ac} = \frac{o_{ac}}{\frac{1}{k_a} \sum\limits_{c \in C_a} o_{ac}} = \frac{o_{ac}}{\overline{o}_a}, \tag{13.4}$$

where k_a is the number of synonymous codons and r denotes a relative synonymous frequency (Sharp *et al*., 1986).

The RSCU values express the relationship between the observed number of codons and the number of times the codon would be observed if the synonymous codon usage was completely random (no codon usage bias). For average synonymous codon usage (no codon bias) the RSCU is 1. For codon usage more infrequent than the average codon usage, the RSCU is less than one, and for more frequent usage than the average for the amino acid, the RSCU is greater than 1.

Another way of normalizing the data is to use the relative adaptiveness, w, in which the frequency of each synonymous codons is normalized by the frequency of the most frequent codon. Thus the most frequent codon will have a relative adaptiveness of 1, while the others will have a relative adaptiveness of less than one. The relative adaptiveness is:

$$w_{ac} = \frac{o_{ac}}{\max\limits_{c \in C_a} o_{ac}}. \tag{13.5}$$

Amino acids decoded by one codon (Trp and Met for the standard genetic code) also have a relative adaptiveness of 1 and are often neglected, as they do not contribute additional information. Stop codons are also often disregarded, since their occurrence is rare compared to other codons and usually strongly biased toward one codon.

13.5.2 Measures based on reference

Many indices compare the query gene to a reference set of genes with some desirable quality. The idea is that certain profiles of codon usage are optimal. Assignment of optimal codons requires strong assumptions, since the factors shaping the codon usage may differ among genes and genomes. The reference set can be defined from either first principles (e.g. Fop) or using a reference set of highly expressed genes (e.g. CAI). Highly expressed genes are under stronger translational selection and the synonymous codons are under stronger selective constraints.

13.5.2.1 Frequency of optimal codons (Fop)

The frequency of optimal codons (Fop), the ratio of the number of optimal codons used to the total number of synonymous codons, was one of the first codon usage measures proposed (Ikemura, 1981b).

The optimal codons can be defined according to nucleotide chemistry, codon usage bias, or tRNA availability. In short: (1) pyrimidine two-codon amino acids prefer A-ending codons over G-ending; (2) purine two-codon amino acids prefer C-ending codons over U-ending; (3) if there exists a tRNA with inosine, the wobble position prefer U- and C-ending codons over those with A-endings; (4) codons with higher tRNA abundance are preferred; and (5) codons that are decoded by more than one different tRNA isoacceptors. The constraint of tRNA abundance is probably the most important constraint (Ikemura, 1985). Therefore, a convenient way to define translationally optimal codons is those codons that are cognate to the most abundant tRNA isoacceptor in each codon family. The tRNA abundances can be inferred from the tRNA gene copy number of genome data. Since tRNA abundance and codon usage are highly correlated, optimal codons can be alternatively defined as those that are the most common.

The frequency of optimal codons is the ratio of the number of optimal codons to the total number of codons:

$$\mathrm{Fop} = \frac{o_{\mathrm{opt}}}{o_{\mathrm{tot}}}. \tag{13.6}$$

The number of optimal codons is:

$$o_{\mathrm{opt}} = \sum_{c \in C_{\mathrm{opt}}} o_c. \tag{13.7}$$

The subset of optimal codons, C_{opt}, is defined according to the above criteria, from all the codons C that are included in the analysis. Amino acids with one codon do not contribute any information and are omitted. Amino acids with one isoacceptor are often excluded when the optimal codon can not be determined. The total number of codons in a sequence o_{tot} is the total number of codons included in the analysis.

13.5.2.2 *Codon bias index (CBI)*

The codon bias index also measures the extent to which preferred codons are used in a gene (Bennetzen and Hall, 1982). The preferred codons are defined as codons frequent in highly expressed genes and codons cognate to the major tRNA species. It is similar to Fop, but uses the expected usage as a scaling factor and thus is normalized between −1 and 1. A value of 1 means only preferred codons are used, zero means random choice and less than zero implies greater use of non-preferred codons:

$$\mathrm{CBI} = \frac{o_{\mathrm{opt}} - e_{\mathrm{rand}}}{o_{\mathrm{tot}} - e_{\mathrm{rand}}}, \tag{13.8}$$

where o_{opt} is the number of preferred optimal codons, o_{tot} is the total number of codons, and e_{rand} is the expected number of optimal codons if random codon assignments were made for each amino acid.

e_{rand} is used to account for the random effect of codon usage and is computed as follows:

$$e_{\mathrm{rand}} = \sum_{a \in A} o_a \frac{n_a^{\mathrm{opt}}}{k_a}, \tag{13.9}$$

where o_a is the number of occurrences of amino acid a in the sequence, n_a^{opt} is the number of instances of optimal codons for amino acid a, and k_a the codon redundancy.

Amino acids with only one codon are excluded from the analysis, as are occasionally amino acids that show little preference towards a single codon (e.g. Asp in Yeast).

13.5.2.3 *Codon usage bias (B)*

The codon usage bias (B) assesses the codon bias of a test set of genes (or group of genes) relative to a second reference set of genes (Karlin and Mrázek, 1996; Karlin *et al.*, 1998). The reference set, composed of a gene class, an entire genome, or a single gene, is used as a standard to which other genes or groups of genes can be compared. This metric is defined as the amino acid frequency weighted sum of distances of the relative codon usage frequencies between the two sets, f and f^{ref}:

$$\mathrm{B} = \sum_{a \in A} F_a \, \mathrm{d}(\mathbf{f}_a, \mathbf{f}_a^{\mathrm{ref}}), \tag{13.10}$$

where F_a is the frequency of the amino acid a in the test set, vectors $\mathbf{f}_a$ and $\mathbf{f}_a^{\mathrm{ref}}$ are the codon frequency vectors for amino acid a in the test and reference set respectively, and d is the 1-norm distance between the codon vectors of amino acid a:

$$\mathrm{d}(\mathbf{f}_a, \mathbf{f}_a^{\mathrm{ref}}) = \sum_{c \in C_a} |f_{ac}, f_{ac}^{\mathrm{ref}}| \tag{13.11}$$

The possible values of B range from 0 to 2, rarely exceeding 0.5. The B measure is also referred to as the codon usage bias CUB.

The codon bias similarity statistic of Gladitz *et al*. (2005) resembles the B measure in several aspects. It differs in that it uses the square of the distance rather than the distance, emphasizing larger differences over many smaller differences. A weighting factor used also places higher weights on the 2-codon amino acids, since they are considered to have a more reliable signal.

The B measure can be used to infer the expression level by comparing the fraction of the distance of the query set with respect to all genes over the distance to a reference set, or a linear combination of reference sets (Karlin and Mrázek, 2000). Using the B measure in this way is then called the E measure (E for expression):

$$\mathrm{E} = \frac{\mathrm{B(all)}}{\mathrm{B(ref)}}. \tag{13.12}$$

13.5.2.4 Codon-enrichment correlation (CEC)

Codon usage in bona fide coding regions deviates from that in randomly generated sequences. There is a preference in amino acid composition as well as bias in the usage of synonymous codons. The codon enrichment correlation is the linear correlation coefficient of the codon enrichment vector $\mathbf{E} = \{\mathbf{E}_{c\in C}\}$ between an ORF and a reference set of genes (Ghaemmaghami *et al*., 2003). This reference set is based on all the ORFs that can confidently be assumed to be real coding sequences. The codon enrichment correlation is computed by:

$$\mathrm{CEC} = \mathrm{corr}(\mathbf{E}^{\mathrm{orf}}, \mathbf{E}^{\mathrm{ref}}). \tag{13.13}$$

The enrichment of each codon for the positive set is defined as the ratio of its frequency among the named ORFs by its expected frequency in random sequences:

$$E_c = \frac{f_c}{e_c}, \tag{13.14}$$

where f_c are the codon frequencies and the expected random codon usage $e_c = b_1 b_2 b_3$ is calculated as a product of the three nucleotide frequencies in codon c. The base frequencies can be taken from either the global nucleotide distribution in coding sequences or, alternatively, assigned using codon position specific nucleotide distribution.

The codon enrichment for the reference set ($\mathbf{E}^{\mathrm{ref}}$) is computed using the relative codon frequencies of all the ORFs in the reference set. The codon enrichment for an ORF ($\mathbf{E}^{\mathrm{orf}}$) is computed using the codon frequencies of the gene itself.

Together with expression data, CEC can be used to identify spurious open reading frames and can be used to detect incorrectly assigned ORFs that are not coding for a protein (Ghaemmaghami *et al*., 2003): if a sequence is not detected experimentally and the CEC is lower than the cutoff value, then the ORF is designated as spurious.

13.5.3 Measures based on the geometric mean

Most indices compute the contribution to the index value for each amino acid individually and then combine them in a second step. The differences between the indices result from differences in these two steps. Many indices use the geometric mean to combine the contribution of each amino acid and differ in the method used to compute contribution of each amino acid.

Most methods sum over the contribution from each amino acid rather than over the length of the sequence, which is essentially the the same, but conveniently then only the codon count vector is required.

13.5.3.1 Codon preference (P)

The codon preference P is a measure of the likelihood of a particular set of codons to a predetermined preferred usage (Gribskov *et al*., 1984). Originally P was computed for all three reading frames using a sliding window, and was used for locating genes and for detection of frame-shifts. The window size L was chosen to be small enough to discriminate genes from non-genes ($L = 25$ for genes smaller than 5000 bp and $L = 50$ otherwise). Here P is used as an index for known sequences and with a window of size L that covers the entire coding sequence. A window size smaller than the total gene length can be used to normalize P, so that it is less dependent on the length of the gene.

The likelihood ratio w_{ac}^{P} is the ratio of the frequency of observing a codon in a gene to the frequency of it being found randomly in the sequence based on the individual nucleotide frequencies in the sequence:

$$w_{ac}^{\mathrm{P}} = \frac{f_{ac}}{e_{ac}}. \qquad (13.15)$$

Note that we reuse the symbol w to denote the likelihood ratio. In Section 13.5.1 it is used for relative adaptiveness. The reason for this is that we want to emphasize the methodological similarities of the indices in this section.

The frequencies f_{ac} are the relative frequencies of codons, while the random codon usage is computed as $e_{ac} = b_1 b_2 b_3$, where b_i is the nucleotide frequency of the ith base of the codon of interest.

To compute P for a gene, we take the product of the likelihood ratios or better, the sum of log-likelihoods:

$$\mathrm{P} = (\prod_{i=1}^{L} w_c^{\mathrm{P}}(i))^{\frac{1}{L}} = \exp(\frac{1}{L} \sum_{i=1}^{L} \log w_c^{\mathrm{P}}(i)) \qquad (13.16)$$

13.5.3.2 Codon adaptation index (CAI)

The codon-adaption index is the most frequently used measure of codon usage bias (Sharp and Li, 1987). The CAI is similar to the codon preference statistic (Gribskov *et al.*, 1984) but instead of using the ratio of the likelihood of finding a codon in a highly expressed gene versus that of finding the codon in a random sequence, the CAI uses the relative adaptiveness (defined in Equation 13.5). The CAI defines translationally optimal codons as those that appear frequently in highly expressed genes (Sharp and Li, 1987).

The relative adaptiveness is computed from a defined subset of translationally optional codons, usually taken from genes that are highly expressed. Alternatively, the relative adaptiveness can be computed without knowledge of highly expressed genes using an iterative procedure that computes the relative adaptiveness from the dominating codon bias of the organism (Carbone *et al.*, 2003). It is also possible to use the codon frequencies of the ribosomal proteins that are known to have generally high expression.

CAI is computed as the geometric mean of the relative adaptiveness for each codon, r_{ac}:

$$\mathrm{CAI} = (\prod_{i=1}^{L} w_c(i))^{\frac{1}{L}} = \exp(\frac{1}{L} \sum_{i=1}^{L} \log w_c(i)) \quad (13.17)$$

This is equivalent to computing CAI from the ratio of the number of codons over the maximum number of codons (of the amino acid) that exists in query gene:

$$\mathrm{CAI} = \frac{\exp(\frac{1}{L}\sum_{i=1}^{L} \log o_{ac}^{\mathrm{ref}}(i))}{\exp(\frac{1}{L}\sum_{i=1}^{L} \log o_{a,\max}^{\mathrm{ref}}(i))}. \qquad (13.18)$$

Commonly CAI is computed by summing over the codons usage vector rather than over the length:

$$\mathrm{CAI} = \exp(\frac{1}{o_{\mathrm{tot}}} \sum_{c \in C} o_c \log w_c). \qquad (13.19)$$

Later improvements to CAI have targeted irregular cases that can cause errors (Xia, 2007) such as the problem encountered with amino acids that have a single codon, amino acids that are encoded by two separate codon families, or when the relative adaptiveness for a codon is zero.

13.5.3.3 Relative codon usage bias (RCB)

The relative codon usage bias (RCB; Roymondal *et al.*, 2009) is a measure that defines the contribution of the codons as:

$$w_c^{\mathrm{RCB}} = \frac{o_c - E[o_c]}{E[o_c]}, \qquad (13.20)$$

where o_c is the observed number of counts of codon c of the query sequence and $E[o_c]$ is the expected number of codon occurrences given the nucleotide distribution at the three codon positions ($b_1 b_2 b_3$).

RCB has the advantage that it does not rely on having a reference set. Unfortunately this index has the drawback that it depends on sequence length as pointed out by Fox and Erill (2010). The value of the index is over-estimated for short sequences. A possible remedy for this may be to introduce pseudo counts for short sequences based on the global codon and nucleotide distribution (Fox and Erill, 2010). Subtracting 1 from the result is used to shift the values such that zero indicates a lack of bias:

$$\mathrm{RCB} = \exp(\frac{1}{o_{\mathrm{tot}}} \sum_{c \in C} \log w_c^{\mathrm{RCB}}) - 1. \quad (13.21)$$

13.5.3.4 Relative codon adaptation (RCA)

The relative codon adaptation (RCA) index reuses the same idea as RCB to define the contribution of the codons but uses a subset of reference sequences (Fox and Erill, 2010). These can be taken either from expression data or determined by using methods to estimate the dominating codon bias (Carbone *et al.*, 2003). RCA, like RCB, has a possible advantage over CAI, in the sense that it considers the underlying nucleotide distribution at the three codon positions.

13.5.3.5 Relative codon adaptation index (rCAI)

The relative CAI takes the background codon usage into account by using the two non-coding frames (Lee *et al.*, 2010). The relative adaptiveness of each codon, w_c^{rel}, is computed by normalizing with codon usage in the +1 and +2 reading frames:

$$w_c^{\mathrm{rel}} = \frac{w_c^0}{\sqrt{w_c^{+1}}\sqrt{w_c^{+2}}}. \quad (13.22)$$

rCAI was developed as a way to better discriminate between highly biased and unbiased genomic regions, i.e. to capture local codon bias patterns. Signals of codon bias are often found by smoothing over a region, since variation among individual codons is large. Investigation of codon bias in smaller regions can benefit from this noise reduction and signal improvement. It appears that a selective force is keeping codons in the +1 and −1 frames non-optimal, possibly preventing frameshifts during translation elongation.

13.5.3.6 An iterative approach to determining codon bias (GCB)

Rather than comparing codon usage to a predefined optimal set of genes, the GCB method iteratively recomputes the top scoring genes to define a reference set of biased genes (Merkl, 2003; similar to the approach by Carbone *et al.*, 2003). The iteration continues until convergence and the stop condition remains unchanged (within a tolerance). To avoid overweighting rare codons, the ratio of codon frequencies has a lower limit of −5. If a set of genes that are known to be highly expressed, e.g. protein participating in the translational process, then this set can be used as a starting point. From the fixed set of reference genes, species-specific scores are computed. The scores for codon c are defined as:

$$w_c^{\mathrm{GCB}} = \frac{f_c^{\mathrm{ref}}}{f_c^{\mathrm{all}}}, \quad (13.23)$$

where f_c^{ref} is the codon frequency in the set of reference genes from the genome and f_c^{all} is the mean frequency of codon c.

Once the CB scores have been fixed, the GCB score of an individual gene can be computed as below. In the original publication the authors have omitted the back-transformation from the log-space:

$$\mathrm{GCB} = \frac{1}{o_{\mathrm{tot}}} \sum_{c \in C} \log w_c^{\mathrm{GCB}}. \quad (13.24)$$

13.5.3.7 tRNA adaptation index (tAI)

The tRNA adaptation index is motivated by the assumption that tRNA availability is a driving force for translational selection. The tRNA adaptation index estimates the extent of adaptation of a gene to its genomic tRNA pool (dos Reis *et al.*, 2003, 2004). It is inspired by and reuses the same idea as the CAI by integrating the degree of adaptation of all codons. It mainly differs from the CAI and the P index in how the relative adaptiveness w is computed. First, the *absolute* adaptation W_c for codon c is computed:

$$W_c = \sum_t (1 - s_{ct}) T_{ct}. \quad (13.25)$$

The index t is summed over all the isoacceptor tRNAs that can recognize codon c and s_{ct} is the efficiency of the codon–anticodon coupling (dos Reis *et al.*, 2004). T_{ct} is the number of copies of the tRNA t that recognizes codon c. The *relative* adaptiveness is normalized to the absolute relative adaptiveness, by dividing with the maximum W_c value of the corresponding amino acid:

$$w_{ac}^{\mathrm{tAI}} = \frac{W_{ac}}{\max_{c \in C_a} W_{ac}}. \quad (13.26)$$

If W_c is zero, then the mean w_{mean} of the relative adaptiveness is used. Finally, the tAI of a gene is computed as the geometric mean of the relative adaptiveness values of its codons n:

$$\mathrm{tAI} = \exp\Big(\frac{1}{o_{\mathrm{tot}}}\sum_{c\in C} o_c \log w_c^{\mathrm{tAI}}\Big). \qquad (13.27)$$

One drawback of tAI is that it requires information that may not be known. Computation of the tAI requires: the codon recognition of tRNA, the properties of anticodon–codon interaction, the correct annotation of tRNA genes, and a subset of highly expressed genes (or alternatively a method to determine optimal codon frequencies). The values of the anticodon–codon affinity s_{ct} may also be difficult to assign correctly. Expression data is used to find the best correlation between expression levels and tAI values and from this the values for the possible anticodon–codon binding are derived. For humans, these values were found to be: G:U=0.41, I:C=0.28, I:A=0.99, U:G=0.68, L:A=0.89 (dos Reis *et al.*, 2004). Furthermore, it is unclear if the accuracy of this information is adequate to give reliable values of tAI. On the other hand, when this information was available, the index has performed favourably (Tuller *et al.*, 2007).

13.5.4 Measures based on deviation from an expected distribution

There is a large group of indices that measure deviation from the expected distribution of codons. These indices have the advantage of being easily understood from a statistical perspective. If the expected distribution can be estimated and a model formulated, the significance analysis and statistical tests can be performed, lending a big advantage.

13.5.4.1 Codon-preference bias measure (CPB)
The codon preference bias (CPB) measures how far observed codon usage deviates from the theoretical mean (McLachlan *et al.*, 1984). The CPB is not used often, perhaps due to the fact the method is quite theoretical and not very straightforward to implement. Like the P measure, it was used to detect *bona fide* coding sequences:

$$\mathrm{CPB} = \frac{U - \overline{U}}{\sigma_U}. \qquad (13.28)$$

The CPB measures the improbability of codon usage U, which is the negative log of the probability of observing a particular codon count vector:

$$U = -\log M(\mathbf{o}). \qquad (13.29)$$

The log transform of the codon count values accounts for the skew of the distribution. The distribution of a codon vector is computed from the multinomial distribution:

$$M(\mathbf{o}) = \frac{o_{\mathrm{tot}}!}{\prod_{c\in C}(o_c!)}\prod_{c\in C} f_c^{o_c}, \qquad (13.30)$$

where o_c is the observed counts of codons in the sequence, f_c is the expected frequency, and o_{tot} is the total sum of codon counts.

The expected frequency of codon c can be computed in several ways. The authors have chosen this to be equal for all synonymous codons, but arguably relative frequencies may be a better choice. The total number of codon counts o_{tot} can be quite large, making the probability distribution of U difficult to compute. Therefore, the authors apply an approximation based on the Poisson distribution.

13.5.4.2 Maximum-likelihood codon bias (MCB)
The Maximum-likelihood codon bias is useful to test a variety of null hypotheses (Urrutia and Hurst, 2001). The method is designed to account for background nucleotide composition and can also be adopted to correct for di-nucleotide biases. The MCB is not strictly a maximum-likelihood method, but the weight of each amino acid is estimated by the likelihood of occurrence of each amino acid given its frequency and codon degeneracy:

$$\mathrm{MCB} = \sum_{a\in A}\frac{B_a \log o_a}{o_{\mathrm{tot}}}, \qquad (13.31)$$

where o_a is the number of occurrences of amino acid a and o_{tot} is the total number of amino acid instances used to compute the index.

The more frequent the amino acid, the more prominent the bias. In such cases the compensation is logarithmic rather than linear, so as to not overemphasize for very frequent amino acids. B_a is the bias of an individual amino acid:

$$B_a = \sum_{c\in C_a}\frac{(o_c - e_c)^2}{e_c}. \qquad (13.32)$$

This is a χ^2-test using the observed o_c and expected e_c counts for each synonymous codon c.

To compute the expected values of codon counts, the nucleotide frequencies within a redundancy class (and the super-groups with larger size) are used. The classes are grouped according to the nucleotide at the third position (NNY, NNR, NNH, and NNN). To minimize uncertainty, all cases with less than 30 included sites are eliminated. The authors note that dinucleotide bias is not taken into account with this model.

13.5.4.3 The scaled χ^2 statistic

The 'scaled' χ^2 is a measure of the bias in silent codon usage (Shields *et al.*, 1988). It is computed as the deviation from the equal usage of synonymous codons divided by the total number of codons in the gene. That is, it is scaled by gene length:

$$\chi^2 = \frac{1}{o_{\text{tot}}} \sum_{a \in A} \sum_{c \in C_a} \frac{o_{ac} - k_a^{-1}}{k_a^{-1}}, \qquad (13.33)$$

where o_{ac} is the frequency of occurrence of codon c in amino acid a and k_a is the degeneracy of amino acid a.

Amino acids with single codons (Trp and Met) are excluded. The scaled χ^2 uses the deviance from equal usage of codons, rather than from the expected distribution of codons given the nucleotide distribution.

13.5.5 Measures based on information theory

Methods originating in statistical linguistics and information theory have also been used for the analysis of DNA sequences (Rao *et al.*, 1979; Konopka, 1984; Pavesi, 1999; Wang *et al.*, 2001; Frappat *et al.*, 2003). Zeeberg (2002) characterized codon usage bias based on the concepts of the Shannon information theory (Shannon, 1948). In the following subsections, two methods based on entropy are discussed.

13.5.5.1 Weighted sum of relative entropy (Ew)

Suzuki *et al.* (2004) suggested a logically sound usage of entropy in which the weighted sum of relative entropy is used to measure the degree of deviation away from equal codon usage. It is suggested that by using only the information in the gene under consideration, the measure is less dependent on biological assumptions, such as mutational biases and translational selection. They propose an index that takes into account the number of distinct amino acids, their relative frequencies, and their degree of codon degeneracy. This index, the weighted sum of relative entropy (Ew), is the sum of the relative entropy of each amino acid weighted by its relative frequency in the sequence and is computed as:

$$\text{Ew} = \sum_{a \in A} F_a E_a, \qquad (13.34)$$

where F_a is the relative frequency (the weight) of the amino acid in the sequence.

The relative entropy E_a is computed by normalizing the entropy H_a by the maximum entropy $\max(H_a) = \log_2 k_a$:

$$E_a = \frac{H_a}{\max(H_a)} = \frac{H_a}{\log_2 k_a}. \qquad (13.35)$$

H_a is the entropy that measures the uncertainty of codon usage in the sequence for amino acid a:

$$H_a = -\sum_{c \in C_a} f_{ac} \log_2 f_{ac}. \qquad (13.36)$$

As with many other indices, the sequence needs to be sufficiently long to avoid stochastic sampling effects. Potential drawbacks may be that Ew does not consider which codons are used and that two sequences may have identical Ew values but different codon usage bias.

13.5.5.2 Synonymous codon usage order (SCUO)

The synonymous codon usage order is an entropy-based measure of codon bias (Wan *et al.*, 2004). It is very similar to Ew and differs only in the way the entropy of each amino is computed. Instead of computing the relative entropy, the authors use the normalized difference between the maximum entropy and the observed entropy:

$$E_a = \frac{\max(H_a) - H_a}{\max(H_a)} = \frac{\log_2 k_a - H_a}{\log_2 k_a}. \qquad (13.37)$$

The SCUO can be computed just as Ew:

$$\text{SCUO} = \sum_{a \in A} F_a E_a. \qquad (13.38)$$

An online server, CodonO, is available for computing SCUO (Angellotti *et al.*, 2007).

13.5.6 Measures focusing on tRNA interaction

Many indices focus on the tRNA usage as the limiting factor. During the translational elongation step, an mRNA is at the ribosome with a codon in the ribosomal A-site. Ternary complexes composed of aminoacyl–tRNAs bound with elongation-factor Tu and GTP are thought to diffuse into the vicinity of the A-site and interact with the codon. If the codon in the A-site does not match the anticodon of the tRNA, it diffuses away and the process repeats until the correct tRNA is in position. At this time, elongation can occur. The indices are then based on the average number of codon–tRNA interactions during one elongation cycle. Several indices mentioned in other sections such as the tAI and Fop also have this property.

13.5.6.1 P1 index

The P1 index is a measure of the influence of tRNA availability (Gouy and Gautier, 1982) and measures the mean of the number of tRNA–codon interactions necessary for a correct recognition for a step in the elongation cycle. The influence is based on a simple model of protein synthesis dynamics, which relies on two strong assumptions. It assumes that all isoacceptor tRNAs have equal binding properties to all the codons they recognize and that the durations of the non-specific tRNA–codon interactions are all equal.

The probability p_c of a correct recognition of codon c is calculated from the relative concentrations of isoacceptor tRNA. These must have either been determined experimentally or predicted using the gene copy number as a proxy. The mean number of tRNA–mRNA interactions at the A-site of the ribosome is the inverse of p_c. For a gene, P1 is computed for each codon weighted by the corresponding codon frequency f_c:

$$\mathrm{P1} = \sum_{c \in C} \frac{f_c}{p_c}. \tag{13.39}$$

Genes that are optimized for a small number of tRNA discriminations are often highly expressed.

13.5.6.2 P2 index

The aim of the P2 index is to measure the bias for anticodon–codon interactions of intermediary strength (Gouy and Gautier, 1982). P2 is the fraction of pyrimidine-ending codons that have intermediate strength. Pyrimidine-ending codons always decode the same amino acid (if the two first positions of the codon are identical) and are almost always recognized by one tRNA isoacceptor. This is not true for purine-ending codons. If the first two positions of the codon are weakly binding nucleotides (W = A or T) then there is a bias for a strong nucleotide at the third position (S = G or C) and vice versa. The P2 index is:

$$\mathrm{P2} = \frac{o_{\mathrm{WWC}} + o_{\mathrm{SSU}}}{o_{\mathrm{WWY}} + o_{\mathrm{SSY}}} \tag{13.40}$$

and its values have been shown to be high for highly expressed genes and low for lowly expressed genes (Gouy and Gautier, 1982).

13.5.6.3 tRNA-pairing index (TPI)

The tRNA pairing index is a measure of synonymous codon ordering comparing of the number of changes of tRNA in a coding sequence to the total number of expected changes given a random distribution of the existing codons. It is worth emphasizing that the codons are not consecutive in the sequence but consecutive codons that encode the same amino acid. To understand the computation of the TPI, assume an example of an amino acid that occurs seven times and is translated by two tRNAs, A and B. The seven codons are extracted from the string and represented by their translating tRNA, e.g. ABBAABB. The most correlated sequences are AAABBBB and BBBBAAA. The most anticorrelated sequence is BABABAB. The number of changes of tRNA in the string quantifies the changes for this amino acid (e.g. six changes for BABABAB). This number is summed for all amino acids with at least two codons and two tRNAs. In yeast these are: Ala, Arg, Gly, Ile, Leu, Pro, Ser, Thr, Val.

To assess the significance, the observed number of changes is compared to the distribution of expected changes given the coding sequence and using the genome-wide codon frequencies. For each amino acid, the number of different tRNAs and the number of times they occur in the coding sequence are used to compute the expected frequency of occurrence for each possible number of changes. Efficient recursions of these distributions have been shown Friberg *et al*. (2006). This results in nine distributions for the nine relevant amino acids. These distributions are then convolved to give a distribution of total expected number of changes. The value of the TPI is:

$$\mathrm{TPI} = 1 - 2p, \tag{13.41}$$

where p is the value of the cumulative density function at the point of the observed number of changes.

This normalization results in a TPI value of 1 for a completely ordered sequence and a TPI of −1 for a completely unordered sequence. The average TPI of the *Saccharomyces cerevisiae* genome was found to be 0.124, biased toward ordering of the codons by their decoding tRNA (Cannarozzi *et al*., 2010).

13.5.7 Measures based on intrinsic properties of codon usage

Some measures that do not fall into the common categorizations of methods are described here.

13.5.7.1 Base composition at silent sites

There is selection for optimal codons in highly abundant proteins. These codons tend to have pyrimidines at the third position, in particular C. Therefore, GC content at silent sites is often correlated with gene expression (Shields *et al*., 1988). Multivariate analyses of codon usage often gives the result that nucleotide content at the third-codon position corresponds to the first principal component, and thus, explains the largest fraction of the variance. Much evidence of selection acting on silent-site base composition exists (Stenico *et al*., 1994; Eyre-Walker, 1999). The G and C nucleotides are strong, that is, they bind more strongly to each other than A and T and non-Watson–Crick base pairings. Hence, they are likely to be more influential on codon usage.

The base composition at silent sites measures the GC content at the third position of synonymous codons (GC3s) and can be used as an index of codon bias. Amino acids with six codons need special handling. They have to be divided into two groups: one of size four and one of size two, where the nucleotides at the two first positions are identical. The following formula describe the GC content at the third codon position, excluding non-degenerate codons:

$$\mathrm{GC3s} = \frac{o_{\mathrm{NNS}}}{o_{\mathrm{tot}}}, \tag{13.42}$$

where o_{NNS} is the number of G- or C-ending codons (S = strong).

It is certainly possible to measure any nucleotide fraction content, but GC3s is the most common.

13.5.7.2 Effective number of codons (Nc)

The effective number of codons (Nc) is the total number of different codons used in a sequence (Wright, 1990). The values of Nc range from 20, where only one codon is used per amino acid, to 61 (for standard genetic code), where all possible synonyms codons are used with equal frequency. Nc measures bias toward the use of a smaller subset of codons, away from equal use of synonymous codons. For example, as mentioned above, highly expressed genes use fewer codons due to selection.

The underlying idea of Nc is similar to the concept of zygosity from population genetics, which refers to the similarity for a gene from two organisms. In the context of codon usage, multiple synonymous codons are treated analogously to multiple alleles. Homozygosity for an amino acid Z_a measures the degree of similarity and is computed based on the relative codon frequencies f_{ac}:

$$Z_a = \frac{o_a \sum_{c \in C_a} f_{ac}^2 - 1}{o_a - 1}. \tag{13.43}$$

The number of effective codons for an amino acid is the inverse of homozygosity:

$$N_a = Z_a^{-1}. \tag{13.44}$$

The value of N_a ranges from 1 to the number of synonymous codons k_a (the codon degeneracy). With equal codon usage, homozygosity is minimal and the value of N_a is the number of synonymous codons. The overall number of effective codons for a gene (Nc) is a sum of average homozygosities Z_a for different redundancy classes k (in set K of all redundancy classes):

$$\mathrm{Nc} = \sum_{k \in K} n_k \overline{N}_{a=k}, \tag{13.45}$$

where for each redundancy class:

$$\overline{N}_{a=k} = \frac{1}{n_k} \sum_{a \in K_k} N_a. \tag{13.46}$$

When the codon usage pattern is more uniform than expected, it is possible to obtain Nc > 61, in which case it is readjusted to 61. If an amino acid is not observed, or is very rare, then the value is replaced by the average homozygosity of the amino acids in the same redundancy class. If Ile is missing (the only member in the redundancy class with three synonymous codons), then the corresponding Z is estimated from the average homozygosity of the other redundancy classes (Fuglsang, 2004). For example, in the case of isoleucine:

$$\overline{Z}_{k=3} = \frac{1}{3}\Big((\frac{2}{\overline{Z}_{k=2}} - 1)^{-1} + (\frac{2}{3\overline{Z}_{k=4}} - \frac{1}{3})^{-1} + (\frac{2}{5\overline{Z}_{k=6}} - \frac{3}{5})^{-1}\Big) \tag{13.47}$$

When there is a large discrepancy among the amino acids for a gene, the sum of Nc for all individual amino acids can be used instead of taking the sum of the averages of each redundancy class (Fuglsang, 2004):

$$\mathrm{Nc} = \sum_{a \in A} \mathrm{N}_a. \tag{13.48}$$

Novembre (2002) proposed a modification of Nc to account for biased background nucleotide distribution. It may be particularly important for phylogenetic studies where the nucleotide distribution may differ among organisms. Novembre uses Person's χ^2 statistic to describe departure of codon usage from the expected regarding the nucleotide distribution.

Nc is a popular index, perhaps due to the fact that the resulting values are easy to interpret, and no knowledge of optimal codons is required.

13.5.7.3 Measure independent of length and composition (MILC)

The MILC is a measure that aims to be independent of gene length and nucleotide composition, as indicated by its name (Supek and Vlahovicek, 2005):

$$\mathrm{MILC} = \frac{1}{L} \sum_{a \in A} M_a - K, \tag{13.49}$$

where L is the number of codons in the sequence, M_a is the goodness of fit test of the observed codon usage to the expected, and K is a correction factor described below.

M_a is based on a log-likelihood ratio similar to the statistical G-test of goodness of fit:

$$M_a = 2 \sum_{c \in C_a} o_{ac} \log \frac{o_{ac}}{e_{ac}}. \tag{13.50}$$

The expected number of codons e_{ac} can be computed in several ways, the simplest being the assumption of equal codon usage. The correction factor K is used to compensate for sampling errors in short sequences where the number of observations is small:

$$K = \frac{1}{L} \sum_{a \in A} (k_a - 1) - \frac{1}{2}. \tag{13.51}$$

The last term $\frac{1}{2}$ is to compensate for extremely unbiased genes as to avoid negative values of MILC.

MILC is used for the prediction of expression level by taking the ratio of the MILC of a gene to the MILC of a reference set of highly expressed proteins, e.g. ribosomal proteins:

$$\mathrm{MELP} = \frac{\mathrm{MILC}^{(\mathrm{gene})}}{\mathrm{MILC}^{(\mathrm{ref})}}. \tag{13.52}$$

13.5.7.4 Intrinsic codon bias index (ICDI)

The ICDI is an index that does not require knowledge of the optimal codons (Freire-Picos *et al.*, 1994). In this sense, it is related to Nc. The value of ICDI ranges from 0 for equal usage to 1 for extremely high-biased genes. The authors estimate that, in general, a bias over 0.5 is high and a bias below 0.3 means little bias (in fungi). The ICDI, a relatively simple index that is highly correlated with Nc and CBI, is computed based on S_a values for each of the 18 amino acids with k-fold degeneracy:

$$S_a = \frac{1}{k_a(k_a - 1)} \sum_{c \in C_a} (r_{ac} - 1)^2, \tag{13.53}$$

where r_{ac} is the relative synonymous codon usage and k_a is the degeneracy of amino acid a in the sequence.

The value of the index is then computed as:

$$\mathrm{ICDI} = \sum_{a \in A} F_a S_a. \tag{13.54}$$

The ICDI gives equal weight to all amino acids included, that is, all values of F_a are $\frac{1}{18}$.

13.5.7.5 HK measure

The HK measure, named by the initials of the authors, relies on a multivariate statistical method (Hey and Kliman, 2002). First, the variation caused by nucleotide content and gene length is removed by regression, using the synonymous codon frequencies and the GC content from non-coding DNA, as well as the length of the protein. The residual variation after the regression is then used for factor analysis. The HK measure is the primary factor from the factor analysis.

13.5.7.6 Strength of mRNA secondary structure

There are many indications that the effects of secondary mRNA structures have to be taken into account (Iserentant and Fiers, 1980). The strength of folding from positions −4 to +38 relative to the initiation codon in the mRNA influences protein expression levels (Kudla *et al.*, 2009). If the ribosome cannot access its binding site because of mRNA secondary structure formation, initiation is prolonged and expression of protein is hampered. Methods for codon optimization utilize folding programs to predict the occurrence of mRNA structures that have strongly-bound folding patterns (Freyhult *et al.*, 2005).

13.5.7.7 Evolutionary rate (ER)

The use of the evolutionary rate for codon usage (denoted ER), was motivated by the observation that codon usage is similar in closely related species and changes much more dramatically over large evolutionary distances and thus is correlated with evolutionary distance (Grantham *et al.*, 1981). As highly expressed genes evolve more slowly, the evolutionary rate can be used to predict the level of expression.

Wall *et al.* (2005) estimated evolutionary rates in four yeasts and examined the correlation between evolutionary rates and both expression level and protein dispensability, which was estimated by the growth rate of mutants deficient for the protein. They found that dispensability and expression both have independent and significant effects on the rate of protein evolution, although they could not yet accurately estimate the relative strengths of these effects.

Drummond *et al.* (2006) used principal component analysis of seven predictors (gene expression level, dispensability, protein abundance, codon adaptation index, gene length, number of protein–protein interactions, and the gene's centrality in the interaction network) to find which had the largest effect on protein evolutionary rates. They found that the dominant component is almost entirely determined by the gene expression level, protein abundance, and codon bias as measured by the CAI.

13.5.7.8 Codon volatility

The codon volatility measures the proportion of the point-mutation neighbours of a codon that encodes different amino acids (Plotkin and Dushoff, 2003). It is based on the observation that codons differ with respect to the likelihood that a point mutation will cause a nonsynonymous mutation.

The volatility $v(c)$ of a codon c is defined as the sum over all one-point neighbouring codons of the distances between corresponding amino acids:

$$v(c) = \sum_{i=1}^{9} \mathrm{d}(A(c_i), A(c)), \qquad (13.55)$$

where $A(c)$ is the amino acid of the corresponding codons, and d quantifies the distance between two amino acids.

The simplest distance is the hamming distance: zero if the amino acids are the same, one if they are different. Alternatively, the Miyata metric can be used, which measures the impact of the hydrophobicity and volume of an amino acid (Miyata *et al.*, 1979). The distance from any amino acid to a stop codon is dependent on the application of the index. In the original publication, zero was used but this may not be biologically valid. The significance of the observed volatility can be computed by comparing it to a bootstrap distribution of alternate synonymous sequences, based on the genomic codon frequencies.

13.5.7.9 Partial least squares regression (PLS)

Welch *et al.* (2009) completed a systematic analysis of gene design parameters in *E. coli* and identified codon usage within a gene as a critical determinant of protein expression levels. For two different genes, they constructed a set of 40 genes, each coding for the same amino acid sequence but differing in their synonymous codon usage. The difference in expression for these synonymous sequences ranged from undetectable to 30% of cellular protein. Using partial least squares regression (PLS; Eriksson *et al*., 2004), the correlation of protein production levels was tested against parameters reported to affect expression. PLS does not provide the optimal codon usage, rather it suggests which codons should differ from their averages as well as the direction. Only a subset of ten amino acids was shown to have an impact on expression levels in *E. coli*: Ala, Gly, Phe, Ser, Lys, Pro, Asp, Leu, Gln, Thr (Welch *et al*., 2009). The codon frequencies that are preferred and disfavoured coincide with the isoacceptor tRNAs that are sensitive to starvation of amino acid (Elf and Ehrenberg, 2005).

13.5.7.10 Synonymous codon usage bias maximum-likelihood estimation (SCUMBLE)

The synonymous codon usage bias maximum-likelihood estimation (SCUMBLE) algorithm is based on an probabilistic model of codon usage for a set of genes (Kloster and Tang, 2008) and is similar to Bailly-Bechet *et al*. (2006). It was proposed to estimate the degree of contribution by different sources ('trends') and their effects on a gene ('offsets' or β_i). Each gene is assigned a given number of offsets $i(g)$ that describe the extent to which a gene g is affected by the estimated bias ('trend') number i. Each trend can be described by a 'preference function' $E_i(c)$, which indicates how much trend i favours or disfavours codon c.

The dimensions that best explain the observed codon usage of the gene set are determined by maximum-likelihood estimation. Although similar to principal component analysis, the authors suggest their model can capture nonlinearities between expression levels and codon usage, while use of the maximum likelihood framework ensures good statistical performance and reduces the risk of artefacts.

When translational selection was found to be the major source of bias (as in *S. cerevisiae*), the first offset (β_1) was highly correlated with gene expression. In contrast, in *Helicobacter pylori*, β_3 was found to be the highest correlating offset. A subsequent study pointed out some weaknesses of SCUMBLE but considered it complementary to rCAI or CAI (Lee *et al*., 2010).

13.5.7.11 Stochastic evolutionary model of protein production rate (SEMPPR)

The SEMPPR, a stochastic evolutionary model of protein production rate, assumes that selection to reduce the cost of nonsense errors drives the evolution of codon bias, which is counteracted by mutation and drift (Gilchrist, 2007). The SEMPPR starts by linking the coding sequence to its protein production cost. This is then linked to fitness and a population genetic model is used to compute the probability of an allele being fixed. In a Bayesian framework, the SEMPPR then generates a posterior probability distribution for the protein production rate of a given gene based on the codon sequence.

This can be conceptualized as a fitness landscape built from protein production costs. The sequences with the minimal and maximal protein production costs are represented as the highest peak and lowest point. The location of an observed sequence is a consequence of selection, mutation, and drift. The height of the peaks and valleys of the fitness landscape scale with the production rate of the gene. Genes with low production rates will have a smaller difference in the energetic usage between the highest peak and the lowest valley than will those with high production rates. Inferences about production rate are not only a function of the absolute difference between the observed and the minimum production rate but also depend on where the observed rate lies with respect to the entire set of possible protein production costs. The results indicate predictions made using this method are as reliable as index-based ones.

13.5.8 Measures for total codon usage in genomes

At times it can be useful to compare the level of codon bias at a genomic level. Some organisms (e.g. yeast and *E. coli*) have a much higher level of codon bias than other organisms (e.g. human and *D. melanogaster*).

13.5.8.1 Mean dissimilarity index (Dmean)

The intention of the mean dissimilarity index (Dmean) is to quantify the level of diversity in synonymous codon usage among all genes (or a subset of genes) within a genome (Suzuki *et al.*, 2009). The synonymous codon usage of a coding sequence can be represented by a vector of length 59 (excluding stop codons and amino acids with only one codon) with values w_{ac} defined as previously:

$$w_{ac} = \frac{o_{ac}}{\max o_{ac}}, \tag{13.56}$$

where o_{ac} is the number of occurrences of this codon and $\max o_{ac}$ is the number of occurrences of the most frequently used synonymous codon for this amino acid, rendering the vector less dependent of gene length, amino acid composition, and codon degeneracy.

The distance between two genes is the Pearson correlation distance (one minus Pearson's product moment correlation coefficient between relative adaptiveness vectors **w**.) Dmean is the normalized mean distance between all pairs of genes (Watve and Gangal, 1996):

$$\text{Dmean} = \frac{2}{G(G-1)} \sum_{i,j \in \text{all pairs}} \{1 - \text{cor}(\mathbf{w}_{(i)}, \mathbf{w}_{(j)})\} \tag{13.57}$$

where G is the total number of genes.

13.6 Dependencies of measures

Indices of codon bias may target different aspects of codon usage, but in general it is desirable that an index is not influenced by properties other than those intended to be measured. Therefore, it is important to be aware of dependencies of underlying properties. Sequence simulation is a useful tool to investigate such dependencies of various indices. To this end, we simulate the effects of nucleotide composition, gene length, codon degeneracy, codon usage discrepancy, and amino acid discrepancy on the performance of various indices. Herein, we generally use the following approach, adapted for the property we want to measure: (1) Draw the amino acids from a distribution based on the codon frequencies, or any other defined amino acid usage; (2) draw the relative frequencies of synonymous codons for that amino acid from the predefined codon distribution; (3) simulate the start and stop codons separately and concatenate them with the rest of the sequence. We make the assumption that there are no interactions among the codons and that the probabilities of the codons are independent of each other. The length of the sequences is fixed to 500 amino acids, unless length dependence is being investigated.

13.6.1 Dependence on nucleotide composition

The nucleotide composition is a result of mutational biases that can cause dependencies for codon bias indices. Often, the underlying nucleotide bias is not the focus of the analysis, but rather the codon usage bias given the background distribution of nucleotides. Dependence is not necessarily a nuisance; for example, the Nc-plot (Wright, 1990) a plot of Nc versus GC3, is used to investigate codon usage patterns across genes. Nevertheless, the dependence on the nucleotide frequencies is unwanted for some indices.

To examine the effects of nucleotide composition, sequences are simulated using a gradient of GC content and a fixed protein length (500 amino acids). First, the individual nucleotide frequencies are set such that the desired GC content is achieved, and the frequency of A and T is equal, and the frequency of G and C are equal. Assuming that the codon frequency is the product of the three nucleotide base frequencies, the codon distribution, and thus the amino acid frequency distribution, can be derived and are used to generate random sequences. The value of the index is computed from these sequences with each point on the plot representing the average of five such sequences.

Figure 13.1a summarizes the dependence of several indices on GC content using a normalized mean. The normalized mean is the mean minus the total mean divided by the total sample standard deviation ($(x - \bar{x})/s_x$). Several indices show dependencies; for example, Nc shows the characteristic parabola used for the Nc-plot, mentioned previously. Also, Fop and CBI have dependencies, some partly due to the GC profile of the defined optimal codons. Of the indices considered, the CAI is the least affected by the GC content. Figure 13.1b

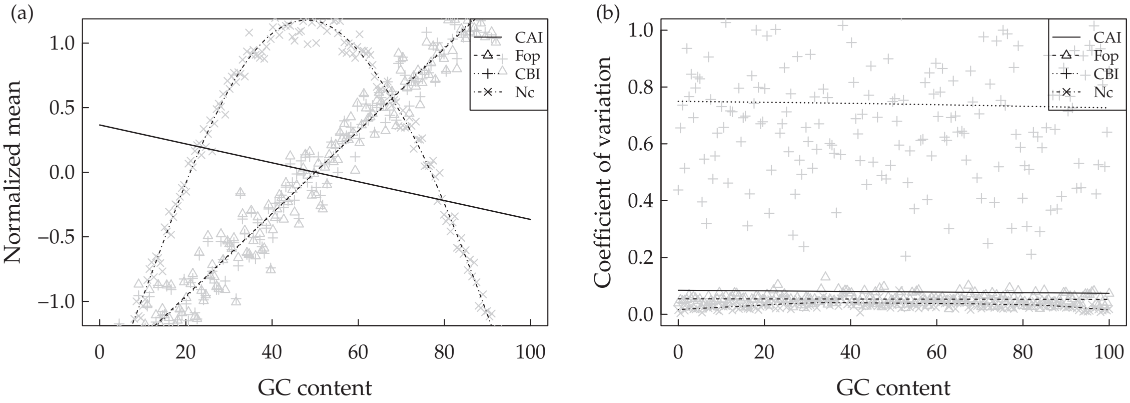

Figure 13.1 The dependence of indices on the GC content is shown for the indices CAI (circle), Fop (triangle), CBI (plus), and Nc (cross). (a) The normalized mean values for each index, where the mean value of the index of the samples at X% GC are subtracted from the total mean divided by the total sample standard deviation ($(x - \bar{x})/s_x$). (b) The values for each index of the coefficient of variation (CV), which is the sample variance divided by the sample mean ($s_x/\bar{x}$).

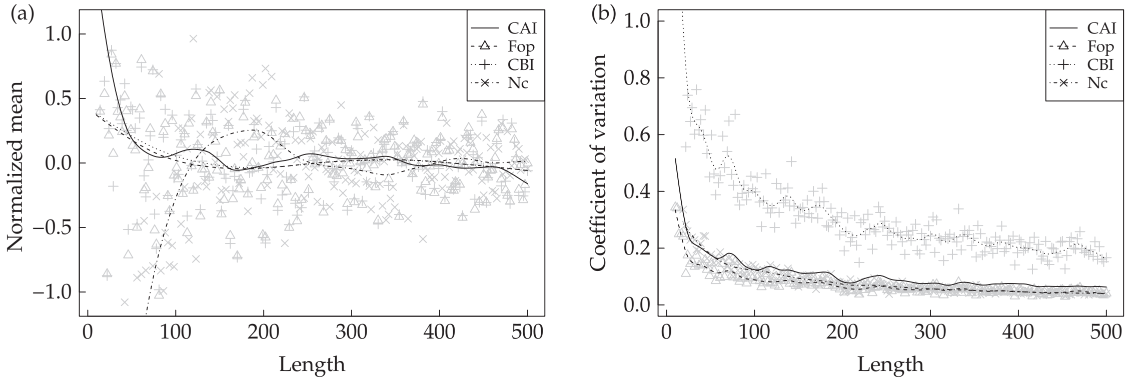

Figure 13.2 Length-dependence of codon indices: (a) the normalized mean of the indices at increasing gene length (number of codons); (b) the coefficient of variation. From the plots it can be seen that at short gene lengths the variance is higher and the estimates tend to deviate from the expected value (based on random sequences).

shows for each index the values of the coefficient of variation (CV), the sample variance divided by the sample mean ($s_x/\bar{x}$). The coefficient of variation (CV) provides a way to compare the variation, irrespective of the value of the mean. The variation is not affected by GC content, albeit CBI has a much larger variance than the other indices.

13.6.2 Dependence on gene length

To examine the dependency caused by differing gene lengths we simulate sequences of different lengths with a fixed codon distribution (using that of *E. coli*). Figure 13.2 shows the dependence of CAI, Fop, CBI, and Nc (a) and their variances (b) on sequence length. Clearly the variation is higher for shorter sequences. This undesirable behaviour is due to stochastic sampling effects and many authors advise against using sequences shorter that 100 amino acids.

13.6.3 Dependence on the degree of codon degeneracy

The degree of degeneracy has been shown to correlate with codon bias indices (Urrutia and Hurst, 2001). Here we define 'degree of degeneracy' as the percentage of four- and six-fold degenerate

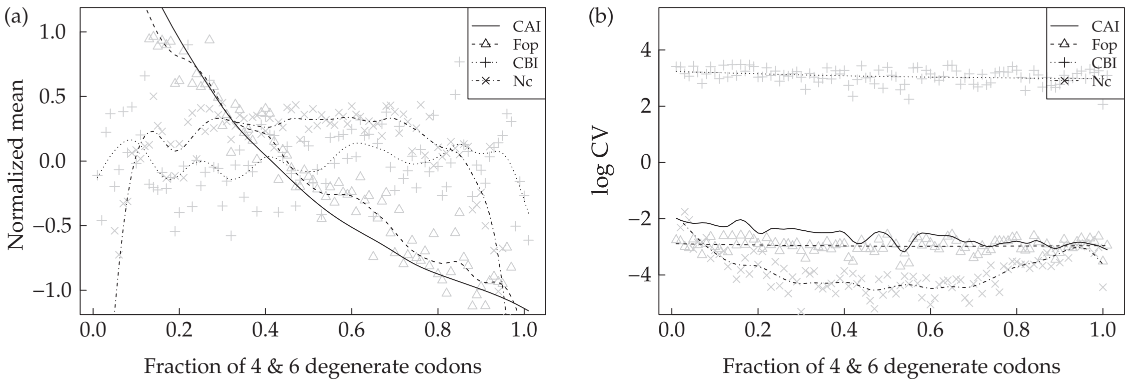

Figure 13.3 The dependence of indices on the fraction of four- and six-fold degenerate codons: See Figure 13.1 caption for the definitions. (a) the normalized means for the indices at increasing degrees of degeneracy; (b) the coefficient of variation for the indices in log space as they are comparatively small.

amino acids of a sequence. In the simulations (Figure 13.3a), it can be seen that CAI and Fop are very dependent on the frequency of degenerate codons, the reason being that these indices rely on a reference set that may have a different set of optimal codons. The indices that do not use a set of preferred codons are less sensitive to this. The behaviour of lower values of Nc at the extremes is expected when the sequences consist of all or none of these four- and six-fold degenerate codons. The variance (Figure 13.3b) shows little change at different degrees of codon degeneracy, although the variance of the CBI is considerably higher than the others.

13.6.4 Dependence on the skewness of synonymous codon usage

Codon usage skewness is the non-uniformity of the synonymous codons. This is due to the underlying nucleotide distribution and is, in fact, very similar to codon usage bias. The reason why we make this distinction is that sometimes it is desirable to measure the codon bias 'on top' of the expected codon frequencies. For example, some organisms with extreme GC content have codon frequencies that are very non-uniform and we would like to detect the sequences that have higher degree of codon bias. We define the maximum discrepancy as to occurring when only one codon is used and the minimum discrepancy as occurring when all codons are used equally. That is, at discrepancy 1, there is a complete uniform distribution of synonymous codons, at 0.5 there is a decay from the first codon of the amino acid to the last. At a discrepancy close to zero, only one randomly selected synonymous codon is used. The sequences are simulated by a discrepancy parameter d that reduces the frequency of the ith synonymous codon by d^{i-1}. For example, when $d = \frac{1}{2}$ for a four-fold amino acid, the codon frequencies will be proportional to $\left\{\frac{1}{2}^0, \frac{1}{2}^1, \frac{1}{2}^2, \frac{1}{2}^3\right\}$, which results in $\left\{\frac{8}{15}, \frac{4}{15}, \frac{2}{15}, \frac{1}{15}\right\}$ after normalization. Figure 13.4 shows the dependencies of CAI, Fop, CBI, and Nc on codon discrepancy. In terms of the normalized mean (Figure 13.4a) CBI and Fop have values close to zero, while CAI shows a slight bias. The number of effective codons Nc measures from the deviation from uniform codon bias (i.e. ranges from 20 to 61) and is obviously dependent on the codon skewness, since this is the underlying property that Nc aims to measure.

13.6.5 Dependence on amino acid discrepancy

Amino acid composition and codon bias are often correlated. Biophysical properties of the protein (content of aromatic amino acids, hydrophobicity, isoelectric point, etc.) can cause dependencies for codon bias indices (Lobry and Gautier, 1994). For example, the content of hydrophobic amino acids in the membrane-bound regions of proteins is high. Here, we look at how the skewness of amino acid usage affects the codon bias indices.

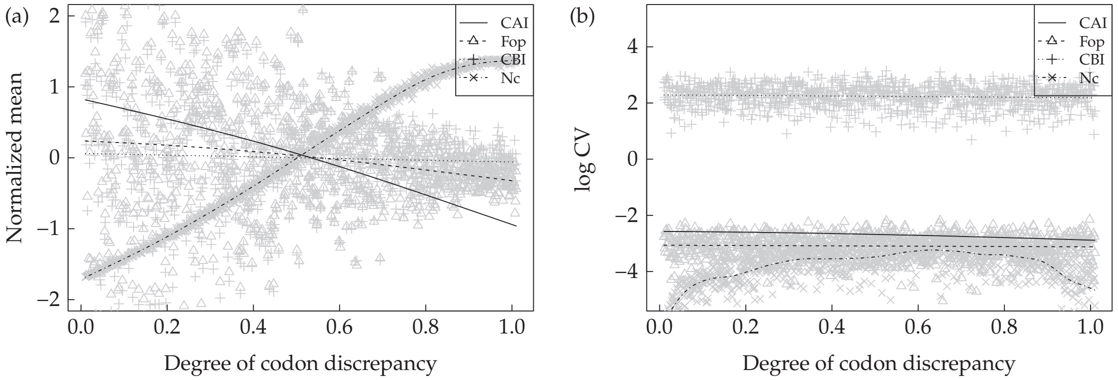

Figure 13.4 Dependence of indices on the codon skewness: (a) The normalized mean and (b) log(CV) are shown for CBI, Fop, CAI, and Nc.

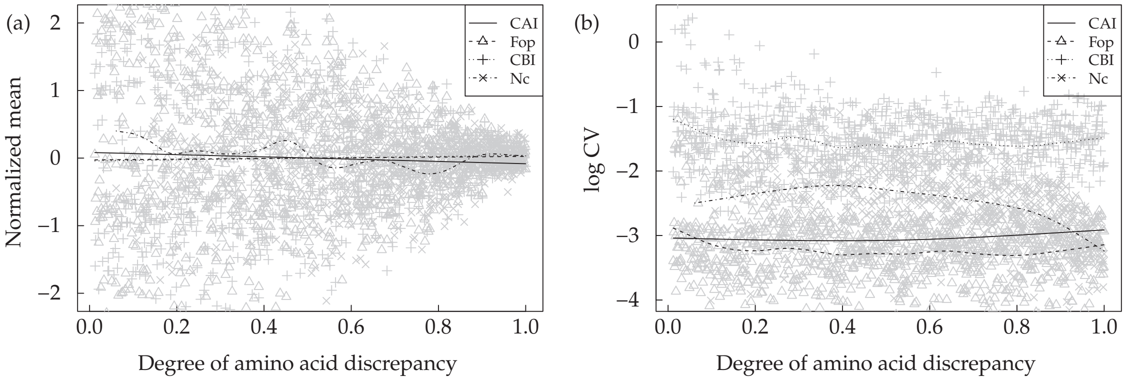

Figure 13.5 Dependence of the indices on the amino acid distribution: (a) the dependency of CAI, Fop, CBI, and Nc on the skew in the amino acid distribution; (b) the coefficient of variation in log space.

Skewness ranges from equal amino acid usage to the hypothetical case of a protein consisting of a single amino acid. Figure 13.5a shows that estimates of CAI, Fop, CBI, and Nc tend to converge towards 0 with increasing discrepancy, while the variance of the estimates is generally low, although CBI has a larger CV than the others.

Note that indices measuring amino acid usage are commonly computed together with codon indices. Two common indices of this type are: the GRAVY and the AROMA. The grand averages of hydropathy (GRAVY) score measures the hydropathicity of a protein (Kyte and Doolittle, 1982) and is the average hydropathy value Y of all the amino acids:

$$\text{GRAVY} = \sum_{a \in A} F_a Y_a, \qquad (13.58)$$

where F_a is the relative frequency and Y_a is the hydropathy index of the amino acids.

The hydropathy values of the amino acids are: A = 1.8, R = –4.5, N = –3.5, D = –3.5, C = 2.5, Q = –3.5, E = –3.5, G = –0.4, H = –3.2, I = 4.5, L = 3.8, K = –3.9, M = 1.9, F = 2.8, P = 1.6, S = –0.8, T = –0.7, W = –0.9, Y = –1.3, V = 4.2. The rationale for the GRAVY index is that the hydropathy of the encoded proteins is a factor influencing the codon usage in some bacteria (de Miranda *et al.*, 2000). The aromaticity score (AROMA) is the aromaticity of a protein, defined as the frequency of aromatic amino acids in a protein (Lobry and Gautier, 1994):

$$\text{AROMA} = \sum_{a \in A_\phi} F_a, \qquad (13.59)$$

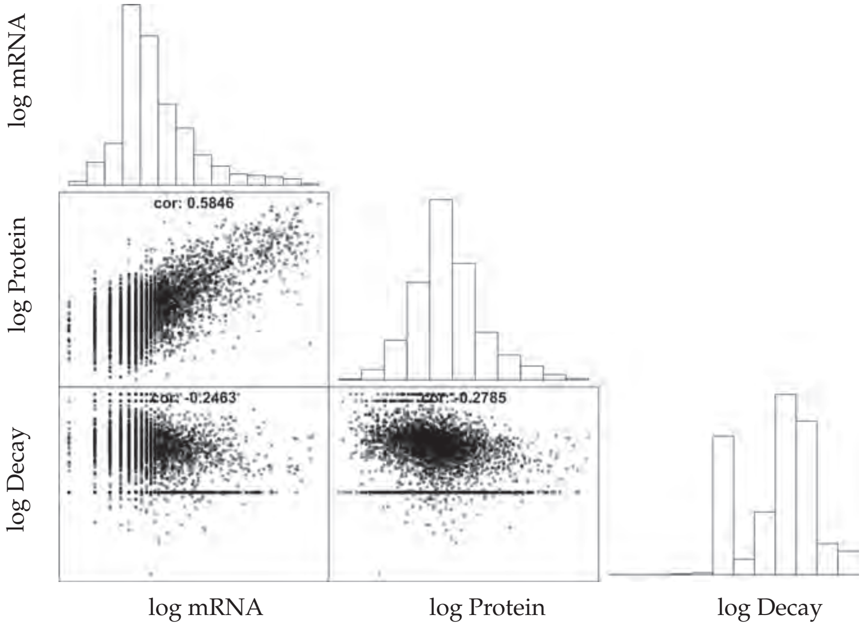

Figure 13.6 Correlation among experimental data. The plots show the correlation between mRNA level, protein level, and rate of protein decay in the yeast *S. cerevisiae* using a integrated dataset von der Haar, 2008. The diagonals show the histograms of the logs of the values of the measurements for the whole genome.

where A_ϕ is the subset of the amino acids that are aromatic (i.e. Phe, Tyr, and Trp), and F_a is the relative frequency of that amino acid the protein.

13.7 Comparisons using biological data

A common usage of codon bias indices is to predict the level of protein abundance. For certain organisms (e.g. *S. cerevisiae* and *E. coli*), there is a clear correlation between protein abundance and codon usage bias. Here we show correlations of codon bias indices with experimental whole-genome measurements of mRNA level, protein abundance and the rate of protein turnover in yeast (von der Haar, 2008) as summarized in Figure 13.6. It is often assumed that the protein level should be dependent on the mRNA level of a gene. However, only a part of the variance of protein levels can be explained by mRNA levels (Spearman correlation coefficient: 0.58). A likely reason for this is that proteins decay at very different rates and this decay influences the protein level. The average rate of protein turnover rate in yeast is 2.2% per hour, but some proteins have rates of almost 10%, while others have rates close to zero (Pratt *et al.*, 2002). Protein decay has a weak inverse correlation with protein and mRNA levels (Figure 13.6), which suggests that abundant proteins tend to have slower decay.

The processes of transcription, translation, and post-translation (e.g. turnover rate and modifications) imposes limits on what it is possible to measure with codon bias indices. For example, fast-growing proteins that are only expressed at a certain time point of development may have values that indicate high abundance, but the overall protein concentration is low. Furthermore, the experimental data that we use for validation have errors. For one thing, there are often systematic biases in expression data originating from the detection limit of the method. For example, smaller proteins are less likely to be detected correctly, since shorter peptides diffuse more readily on 2D gels, which

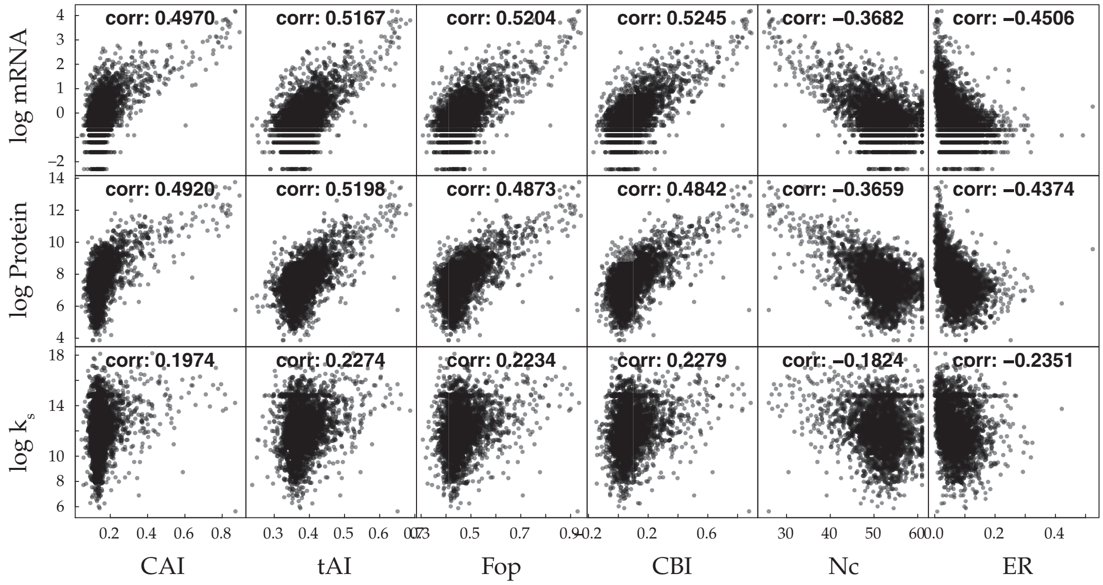

Figure 13.7 Correlation with experimental data. The correlations of CAI, tAI, Fop, CBI, Nc, and ER with the logs of mRNA concentration, protein concentration, and the the rate of protein synthesis k_S in *S. cerevisiae* are shown.

decrease the intensity of the spots. The correlations between these data measurements and the codon bias indices are discussed in the following sections.

13.7.1 Correlation with transcript and protein levels

Codon usage correlates with the mRNA transcript levels and protein abundance in yeast because of selection for optimal elongation. If elongation is inefficient, larger quantities of ribosomes are occupied on the mRNA and are not available to engage in initiation of translation. Therefore, codon patterns that promote efficient translation are preferred. For example, ribosomal proteins are among the most abundant proteins and typically have a high codon bias.

Indices that have a high correlation with expression levels are desirable for the prediction of expression. The top row of Figure 13.7 shows the correlation between various indices and the logarithm of the mRNA levels for *S. cerevisiae*. The highest correlation coefficients are found for the four indices based on distance to the optimal codon usage, CBI and Fop with CAI and tAI close behind. The correlation of the protein level with the indices is similar to that of mRNA levels, although CAI and tAI show a slightly higher correlation than CBI and Fop (middle row in Figure 13.7).

13.7.2 Correlation with rate of protein synthesis

It appears that the rate of protein turnover (e.g. protein degradation) is not the same for all proteins and that normalizing mRNA concentration without accounting for protein degradation may be an oversimplification. Here we look at the correlation of codon bias indices with the rate of protein synthesis k_s.

Figure 13.8 shows a simplified scheme of protein synthesis, in which the concentration of protein

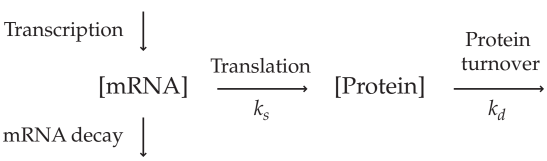

Figure 13.8 Model of protein synthesis. The protein concentration depends on the concentration of mRNA, the rate of synthesis k_S, and the rate of protein degradation k_d.

depends on the mRNA level and protein decay. The concentration of mRNA depends on the rate of transcription and mRNA decay. We will not pay particular interest to the dynamics of mRNA and use the mRNA concentrations directly. We assume that the speed of translation can be inferred from the rate of translation, which can be inferred if the concentrations of mRNA and protein are known, along with the protein degradation rate. The rate of synthesis k_s can be defined as the following. Assume that the change in the concentration of a particular protein is:

$$\frac{d[\text{Protein}]}{dt} = k_s[\text{mRNA}] - k_d[\text{Protein}]. \qquad (13.60)$$

Assuming steady-state for protein and mRNA concentrations, we can find an approximation of the rate of protein synthesis from the ratio of the concentrations of protein and mRNA, and the rate of protein decay:

$$k_s = k_d \frac{[\text{Protein}]}{[\text{mRNA}]}. \qquad (13.61)$$

The protein decay rate can be determined from the protein half-life time $k_d = \ln 2/t_{\frac{1}{2}}$. Although the whole-genome measurements of protein degradation is much less studied than protein and mRNA abundances, the half-lives of proteins have been determined for yeast (Belle *et al*., 2006). The correlation between the indices and the synthesis rate is shown in the bottom row of Figure 13.7. The indices show less correlation with the protein synthesis rate than with the protein and mRNA abundance. One potential reason for this is that the experimental data is often associated with large errors and in our model for the protein synthesis rate k_s, the errors from three different separate experiments are cumulated. Also, k_d can have big error since the whole genome measures are performed under different conditions.

13.8 Limitations of codon usage indices

All codon indices map some aspect of codon usage to one single number. The loss of information by this reduction in dimensionality means that indices cannot capture the entire extent of the underlying biological phenomena. Limitations and shortcomings of all codon bias indices are also present. A common theme is that indices fail to exclude the confounding effects of other biases. As mentioned in the introduction, there are several such effects. The amino acid composition of a protein can strongly influence the codon usage, as well as the nucleotide distribution. The length of a gene can be a strong feature for determining the codon bias, in particular for very short sequences, where all amino acids and codons may not be present. A potential remedy for missing or rare codons or amino acids is to use pseudo-counts for the codon distribution. An intra-genic variation of codon usage also exists, in which the amount and direction of codon bias can vary along the gene (Qin *et al*., 2004). For example, slow codons at the start of the coding regions serve to slowly load ribosomes onto the mRNA to avoid congestion (Tuller *et al*., 2010). Such position-specific codon biases further complicate the estimates and care must be taken to account for variable codon usage along the gene. Although not discussed in this chapter, significant dicodon-biases exist: for example, two consecutive rare codons are generally avoided, since this increases the probability of ribosome drop-off (Cruz-Vera *et al*., 2004).

Sometimes overlooked is the fact that some organisms use alternative genetic codes. The reason for this is that the two most common genetic codes, the standard (1) and the bacterial (11) are identical apart from that the bacterial has several different start codons. Several indices ignore the start codon, since it is being read by a designated tRNA that is not part of the elongation. Nevertheless there are many organisms that use other alternative genetic codes and most indices have to be adopted to account for this.

13.9 Conclusions

This chapter summarizes many codon bias indices and unifies their notation to facilitate visualization of their similarities. We have classified the indices into categories based on historical and methodological similarities. In addition to reviewing the indices, we have outlined methodologies to evaluate them, evaluated a few indices to illustrate their behaviour, and suggested methodologies for further studies. To evaluate all indices is beyond the scope of this review, since for many indices no implementation is available.

We have investigated the dependence of the indices on properties of the sequences using simulations. We have also estimated the extent to which the indices capture different aspects of expression-based experimental data. To this end, we measured the correlation of the indices with mRNA and protein abundance data, as well as an estimated rate of synthesis. A statistical framework in which all methodologies could be evaluated in a systematic manner would be desirable to answer questions of performance.

The choice of index depends on the task, as different indices measure different aspects of codon usage. To predict protein yield for over-expression of heterologous proteins, the PLS measure performs well when the goal is to optimize yield in protein production (Welch *et al*., 2009). In such cases, tRNA depletion becomes a limiting factor and thus codons less sensitive to starvation become preferable. The codon adaptation index (CAI) is a long used method for measuring codon usage bias and has the advantage of being widely known and understood. In particular, the version by Carbone *et al*. (2003) is convenient and remains a good choice for measuring codon usage bias (Friberg *et al*., 2004), as it does not require external knowledge, such as optimal codons or anticodon–codon mapping. Other studies have also provided recommendations for which indices to use (Supek and Vlahovicek, 2005).

Several different complementary indices can be used to understand the diversity of codon usage among genes and organisms as they sometimes capture different aspects of evolution. It may be that an amalgam of indices may provide improved performance. For example, a combination of indices that capture different aspects of translation can be used as a better classifier for predicting translation efficiency (Tuller *et al*., 2004).

In our opinion there is room for improvement, in particular, for predictions towards functionality, regulation, and lowly expressed genes. In addition to the obvious requirements of being theoretically sound and adequately described, a few points should be observed when devising a new index of codon usage. Any new index should have an accessible implementation. If possible, the source code of the implementation should be accessible in order to facilitate verification and understanding. A web-interface (preferably including a web API) and downloadable binaries are essential. If the index is to reach the intended audience, the importance of a proper implementation can not be underestimated.

References

Adzhubei, A.A., Adzhubei, I.A., Krasheninnikov, I.A., and Neidle, S. (1996). Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS Lett*, 399(1-2): 78–82.

Akashi, H. (1994). Synonymous codon usage in drosophila melanogaster: natural selection and translational accuracy. *Genetics*, 136(3): 927–35.

Angellotti, M.C., Bhuiyan, S.B., Chen, G., and Wan, X.-F. (2007). CodonO: codon usage bias analysis within and across genomes. *Nucleic Acids Res*, 35(Web Server issue):W132–6.

Bailly-Bechet, M., Danchin, A., Iqbal, M., Marsili, M., and Vergassola, M. (2006). Codon usage domains over bacterial chromosomes. *PLoS Computational Biology*, 2(4):e37.

Begley, U., Dyavaiah, M., Patil, A., Rooney, J.P., DiRenzo, D., Young, C.M. *et al*. (2007). Trm9-catalyzed tRNA modifications link translation to the DNA damage response. *Mol Cell*, 28(5): 860–70.

Belle, A., Tanay, A., Bitincka, L., Shamir, R., and O'Shea, E.K. (2006). Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci USA*, 103(35): 13004–9.

Bennetzen, J.L. and Hall, B.D. (1982). Codon selection in yeast. *J Biol Chem*, 257(6): 3026–3031.

Beutler, E., Gelbart, T., Han, J.H., Koziol, J.A., and Beutler, B. (1989). Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proc Natl Acad Sci USA*, 86(1): 192–6.

Bodilis, J. and Barray, S. (2006). Molecular evolution of the major outer-membrane protein gene (oprF) of Pseudomonas. *Microbiology*, 152(Pt 4): 1075–88.

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature*, 325(6106): 728–30.

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C. *et al*. (2010). A role for codon order in translation dynamics. *Cell*, 141(2): 355–67.

Carbone, A., Zinovyev, A., and Kepes, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, 19(16): 2005–15.

Chamary, J.V., Parmley, J.L., and Hurst, L.D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*, 7(2): 98–108.

Charif, D., Thioulouse, J., Lobry, J.R., and Perriere, G. (2005). Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics*, 21(4): 545–7.

Clarke, B. (1970). Darwinian evolution of proteins. *Science*, 168(934): 1009–11.

Coghlan, A. and Wolfe, K.H. (2000). Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast*, 16(12): 1131–45.

Comeron, J.M. and Aguadé, M. (1998). An evaluation of measures of synonymous codon usage bias. *J Mol Evol*, 47(3): 268–74.

Cortez, D.Q., Lazcano, A., and Becerra, A. (2005). Comparative analysis of methodologies for the detection of horizontally transferred genes: a reassessment of first-order Markov models. *In Silico Biol*, 5(5-6): 581–92.

Cruz-Vera, L.R., Magos-Castro, M.A., Zamora-Romo, E., and Guarneros, G. (2004). Ribosome stalling and peptidyl-trna drop-off during translational delay at aga codons. *Nucleic acids research*, 32(15): 4462–8.

de Miranda, A.B., Alvarez-Valin, F., Jabbari, K., Degrave, W.M., and Bernardi, G. (2000). Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *J Mol Evol*, 50(1): 45–55.

D'Onofrio, G., Jabbari, K., Musto, H., and Bernardi, G. (1999). The correlation of protein hydropathy with the base composition of coding sequences. *Gene*, 238(1): 3–14.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res*, 32(17): 5036–44.

dos Reis, M., Wernisch, L., and Savva, R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res*, 31(23): 6976–85.

Drummond, D.A., Raval, A., and Wilke, C.O. (2006). A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*, 23(2): 327–37.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current opinion in genetics & development*, 12(6): 640–9.

Duret, L. and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci USA*, 96(8): 4482–7.

Elf, J. and Ehrenberg, M. (2005). What makes ribosome-mediated transcriptional attenuation sensitive to amino acid limitation? *PLoS Comput Biol*, 1(1):e2.

Elf, J., Nilsson, D., Tenson, T., and Ehrenberg, M. (2003). Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science*, 300(5626): 1718–22.

Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F. *et al*. (2004). Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal Bioanal Chem*, 380(3): 419–29.

Eyre-Walker, A. (1999). Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics*, 152(2): 675–83.

Faux, N.G., Huttley, G.A., Mahmood, K., Webb, G.I., de la Banda, M.G., and Whisstock, J.C. (2007). RCPdb: An evolutionary classification and codon usage database for repeat-containing proteins. *Genome Res*, 17(7): 1118–27.

Fijalkowska, I.J., Jonczyk, P., Tkaczyk, M.M., Bialoskorska, M., and Schaaper, R.M. (1998). Unequal fidelity of leading strand and lagging strand DNA replication on the Escherichia coli chromosome. *Proc Natl Acad Sci USA*, 95(17): 10020–5.

Fox, J.M. and Erill, I. (2010). Relative codon adaptation: a generic codon bias index for prediction of gene expression. *DNA Res*, 17(3): 185–96.

Frappat, L., Minichini, C., Sciarrino, A., and Sorba, P. (2003). Universality and Shannon entropy of codon usage. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(6 Pt 1): 061910.

Freire-Picos, M.A., González-Siso, M.I., Rodríguez-Belmonte, E., Rodríguez-Torres, A.M., Ramil, E., and Cerdán, M.E. (1994). Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes. *Gene*, 139(1): 43–9.

Freyhult, E., Gardner, P.P., and Moulton, V. (2005). A comparison of RNA folding measures. *BMC Bioinformatics*, 6: 241.

Friberg, M., von Rohr, P., and Gonnet, G. (2004). Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in *Saccharomyces cerevisiae*. *Yeast*, 21(13): 1083–93.

Friberg, M.T., Gonnet, P., Barral, Y., Schraudolph, N.N., and Gonnet, G.H. (2006). Measures of codon bias in yeast, the tRNA pairing index and possible DNA repair mechanisms. *Algorithms in Bioinformatics, Proceedings*, 4175: 1–11.

Fuglsang, A. (2004). Bioinformatic analysis of the link between gene composition and expressivity in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *Antonie Van Leeuwenhoek*, 86(2): 135–47.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N. *et al*. (2003). Global analysis of protein expression in yeast. *Nature*, 425(6959): 737–41.

Gilchrist, M.A. (2007). Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol Biol Evol*, 24(11): 2362–72.

Gladitz, J., Shen, K., Antalis, P., Hu, F.Z., Post, J.C., and Ehrlich, G.D. (2005). Codon usage comparison of novel genes in clinical isolates of Haemophilus influenzae. *Nucleic Acids Res*, 33(11): 3644–58.

Goetz, R.M. and Fuglsang, A. (2005). Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. *Biochem Biophys Res Commun*, 327(1): 4–7.

Gouy, M. and Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res*, 10(22): 7055–74.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res*, 9(1):r43–74.

Gribskov, M., Devereux, J., and Burgess, R.R. (1984). The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res*, 12(1 Pt 2): 539–49.

Grosjean, H.J., de Henau, S., and Crothers, D.M. (1978). On the physical basis for ambiguity in genetic coding interactions. *Proc Natl Acad Sci USA*, 75(2): 610–4.

Harrison, R.J. and Charlesworth, B. (2011). Biased gene conversion affects patterns of codon usage and amino acid usage in the Saccharomyces sensu stricto group of yeasts. *Mol Biol Evol*, 28(1): 117–29.

Hershberg, R. and Petrov, D.A. (2008). Selection on codon bias. *Annu Rev Genet*, 42: 287–99.

Hey, J. and Kliman, R.M. (2002). Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics*, 160(2): 595–608.

Ikemura, T. (1981a). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol*, 146(1): 1–21.

Ikemura, T. (1981b). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol*, 151(3): 389–409.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*, 2(1): 13–34.

Iserentant, D. and Fiers, W. (1980). Secondary structure of mRNA and efficiency of translation initiation. *Gene*, 9(1-2): 1–12.

Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H., and Ikemura, T. (2001). Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* o157 genome. *Gene*, 276(1-2): 89–99.

Karlin, S. and Mrázek, J. (1996). What drives codon choices in human genes? *J Mol Biol*, 262(4): 459–72.

Karlin, S. and Mrázek, J. (2000). Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol*, 182(18): 5238–50.

Karlin, S., Mrazek, J., and Campbell, A.M. (1998). Codon usages in different gene classes of the *Escherichia coli* genome. *Mol Microbiol*, 29(6): 1341–55.

Kaufmann, W.K. and Paules, R.S. (1996). DNA damage and cell cycle checkpoints. *FASEB J*, 10(2): 238–47.

Kimchi-Sarfaty, C., Oh, J.M., Kim, I.-W., Sauna, Z.E., Calcagno, A.M., Ambudkar *et al*. (2007). A 'silent' polymorphism in the MDR1 gene changes substrate specificity. *Science*, 315(5811): 525–8.

Kloster, M. and Tang, C. (2008). SCUMBLE: a method for systematic and accurate detection of codon usage bias by maximum likelihood estimation. *Nucleic Acids Res*, 36(11): 3819–27.

Knight, R.D., Freeland, S.J., and 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 Biol*, 2(4):RESEARCH0010.

Konopka, A. (1984). Is the information content of DNA evolutionarily significant? *J Theor Biol*, 107(4): 697–704.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, 324(5924): 255–8.

Kunkel, T.A., Pavlov, Y.I., and Bebenek, K. (2003). Functions of human DNA polymerases eta, kappa and iota suggested by their properties, including fidelity with undamaged DNA templates. *DNA Repair (Amst)*, 2(2): 135–49.

Kyte, J. and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1): 105–32.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J. *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860–921.

Lao, P.J. and Forsdyke, D.R. (2000). Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res*, 10(2): 228–36.

Lee, S., Weon, S., Lee, S., and Kang, C. (2010). Relative codon adaptation index, a sensitive measure of codon usage bias. *Evol Bioinform Online*, 6: 47–55.

Lithwick, G. and Margalit, H. (2005). Relative predicted protein levels of functionally associated proteins are conserved across organisms. *Nucleic Acids Res*, 33(3): 1051–57.

Lobry, J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*, 13(5): 660–5.

Lobry, J.R. and Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res*, 22(15): 3174–80.

Lynn, D.J., Singer, G.A.C., and Hickey, D.A. (2002). Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res*, 30(19): 4272–7.

Macaya, G., Thiery, J.P., and Bernardi, G. (1976). An approach to the organization of eukaryotic genomes at a macromolecular level. *Journal Mol Biol*, 108(1): 237–54.

McInerney, J. (1998). Gcua: general codon usage analysis. *Bioinformatics*, 14(4): 372–33.

McLachlan, A.D., Staden, R., and Boswell, D.R. (1984). A method for measuring the non-random bias of a codon usage table. *Nucleic Acids Res*, 12(24): 9567–75.

McLean, M.J., Wolfe, K.H., and Devine, K.M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol*, 47(6): 691–6.

Merkl, R. (2003). A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. *J Mol Evol*, 57(4): 453–66.

Miyata, T., Hayashida, H., Yasunaga, T., and Hasegawa, M. (1979). The preferential codon usages in variable and constant regions of immunoglobulin genes are quite distinct from each other. *Nucleic Acids Res*, 7(8): 2431–8.

Novembre, J.A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol*, 19(8): 1390–4.

Pagani, F. and Baralle, F.E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet*, 5(5): 389–96.

Parmley, J.L. and Huynen, M.A. (2009). Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation. *PLoS Genet*, 5(7):e1000548.

Pavesi, A. (1999). Relationships between transcriptional and translational control of gene expression in *Saccharomyces cerevisiae*: a multiple regression analysis. *J Mol Evol*, 48(2): 133–41.

Pavlov, Y.I., Mian, I.M., and Kunkel, T.A. (2003). Evidence for preferential mismatch repair of lagging strand dna replication errors in yeast. *Curr Biol*, 13(9): 744–8.

Peden, J.F. (2000). *CodonW*, p. 1; http://codonw.source.forge.net/(last accessed September 2011).

Plotkin, J.B. and Dushoff, J. (2003). Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proc Natl Acad Sci USA*, 100(12): 7152–7.

Pratt, J.M., Petty, J., Riba-Garcia, I., Robertson, D.H.L., Gaskell, S.J., Oliver, S.G. *et al*. (2002). Dynamics of protein turnover, a missing dimension in proteomics. *Mol Cell Proteomics*, 1(8): 579–91.

Qin, H., Wu, W.B., Comeron, J.M., Kreitman, M., and Li, W.-H. (2004). Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, 168(4): 2245–60.

Ran, W. and Higgs, P.G. (2010). The influence of anticodon–codon interactions and modified bases on codon usage bias in bacteria. *Mol Biol Evol*, 27(9): 2129–40.

Rao, G.S., Hamid, Z., and Rao, J.S. (1979). The information content of DNA and evolution. *J Theor Biol*, 81(4): 803–7.

Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite. *Trends Genet*, 16(6): 276–7.

Roymondal, U., Das, S., and Sahoo, S. (2009). Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome. *DNA Res*, 16(1): 13–30.

Ruiz, L.M., Armengol, G., Habeych, E., and Orduz, S. (2006). A theoretical analysis of codon adaptation index of the *Boophilus microplus* bm86 gene directed to the optimization of a DNA vaccine. *J Theor Biol*, 239(4): 445–9.

Saunders, R. and Deane, C.M. (2010). Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res*, 38(19): 6719–28.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27: 379–423, 623–56.

Sharp, P.M. and Li, W.H. (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, 15(3): 1281–95.

Sharp, P.M., Tuohy, T.M., and Mosurski, K.R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*, 14(13): 5125–43.

Shields, D.C., Sharp, P.M., Higgins, D.G., and Wright, F. (1988). 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol*, 5(6): 704–16.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C. *et al*. (2002). The BioPerl toolkit: Perl modules for the life sciences. Genome Res, 12(10): 1611–8.

Stenico, M., Lloyd, A.T., and Sharp, P.M. (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res*, 22(13): 2437–46.

Sugaya, N., Sato, M., Murakami, H., Imaizumi, A., Aburatani, S., and Horimoto, K. (2004). Causes for the large genome size in a cyanobacterium Anabaena sp. PCC7120. *Genome Inform*, 15(1): 229–38.

Supek, F. and Vlahovicek, K. (2004). Inca: synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics*, 20(14): 2329–30.

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

Suzuki, H., Saito, R., and Tomita, M. (2004). The 'weighted sum of relative entropy': a new index for synonymous codon usage bias. *Gene*, 335: 19–23.

Suzuki, H., Brown, C.J., Forney, L.J., and Top, E.M. (2008). Comparison of correspondence analysis methods for synonymous codon usage in bacteria. *DNA Res*, 15(6): 357–65.

Suzuki, H., Saito, R., and Tomita, M. (2009). Measure of synonymous codon usage diversity among genes in bacteria. *BMC Bioinformatics*, 10: 167.

Tsirigos, A. and Rigoutsos, I. (2005). A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res*, 33(3): 922–33.

Tuller, T., Kupiec, M., and Ruppin, E. (2007). Determinants of protein abundance and translation efficiency in S. cerevisiae. *PLoS Comput Biol*, 3(12):e248.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J. *et al*. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2): 344–54.

Uemura, S., Aitken, C.E., Korlach, J., Flusberg, B.A., Turner, S.W., and Puglisi, J.D. (2010). Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature*, 464(7291): 1012–7.

Urrutia, A.O. and Hurst, L.D. (2001). Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, 159(3): 1191–9.

von der Haar, T. (2008). A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol*, 2: 87.

Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaever, G., Eisen, M.B. *et al*. (2005). Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci USA*, 102(15): 5483–8.

Wan, X.-F., Xu, D., Kleinhofs, A., and Zhou, J. (2004). Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol Biol*, 4: 19.

Wang, H.C., Badger, J., Kearney, P., and Li, M. (2001). Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Mol Biol Evol*, 18(5): 792–800.

Watve, M.G. and Gangal, R.M. (1996). Problems in measuring bacterial diversity and a possible solution. *Appl Environ Microbiol*, 62(11): 4299–4301.

Weiss, M., Schrimpf, S., Hengartner, M.O., Lercher, M.J., and von Mering, C. (2010). Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics*, 10(6): 1297–306.

Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J. *et al*. (2009). Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One*, 4(9):e7002.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene*, 87(1): 23–9.

Xia, X. (1996). Maximizing transcription efficiency causes codon usage bias. *Genetics*, 144(3): 1309–20.

Xia, X. (2007). An improved implementation of codon adaptation index. *Evolutionary Bioinformatics*, 3: 53–8.

Zeeberg, B. (2002). Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. *Genome Res*, 12(6): 944–55.

Zhou, T., Lu, Z.H., and Sun, X. (2005). The correlation between recombination rate and codon bias in yeast mainly results from mutational bias associated with recombination rather than hill-robertson interference. *Conf Proc IEEE Eng Med Biol Soc*, 5: 4787–90.