Chapter 9
Bioinformatics and Translation Elongation

1 Introduction

We will first learn a few key definitions and notations on tRNA, its anticodon, and codon families. We will then outline the conceptual framework of codon adaptation, mediated by mutation and selection. This brings us to indices of codon usage bias, their calculation and interpretations, and factors that may confound their interpretations. There are codon-specific indices such as relative synonymous codon usage (RSCU, Sharp et al. 1986) or gene-specific indices such as index of translation elongation (I_{TE}, Xia 2015) and codon adaptation index (CAI, Sharp and Li 1987; Xia 2007c). All these indices are implemented in DAMBE (Xia 2013, 2017d).

I_{TE} takes background mutation bias into consideration, while CAI does not. I_{TE} is reduced to CAI if there is no background mutation bias. I will illustrate the applications of these indices in practical research. Keep in mind that a codon adaptation index is just one variable which will not be particularly interesting until you relate it to other variables and understand their relationships.

Two additional topics are dealt with close to the end of the chapter. The first involves how to discriminate between selection for translation efficiency and accuracy (Akashi 1994). The second is on the effect of amino acid usage on translation elongation efficiency. The general prediction concerning amino acid usage is that highly expressed proteins should maximize the use of amino acids that are abundant and energetically cheap (Akashi and Gojobori 2002) to make and have many tRNAs to carry them (Xia 1998a). The same argument has been used for transcription, i.e., an mRNA with many A nucleotides will be transcribed faster than one with many C nucleotides because A is in general far more abundant than C and it takes extra ATP to make CTP (Xia 1996; Xia et al. 2006).

X. Xia, *Bioinformatics and the Cell*, https://doi.org/10.1007/978-3-319-90684-3_9

1.1 Basic Notations, Definitions, and Abbreviations

Notations, definitions, and abbreviations are essential in science. We are lucky enough to have almost all of them unambiguous. If you were studying social sciences, you would have to come to define what is man and what is woman, and the debate on a proper definition will last forever, eventually with all debaters losing their mind and being called jerks.

1.1.1 tRNA Notation and Identification of tRNA Anticodon

The simplest notation of a tRNA is $tRNA^{AA}$, where AA is a specific amino acid. For example, $tRNA^{Gly}$ refers to all tRNAs that can be charged with amino acid glycine (Gly). A slightly more complicated notation is $tRNA^{AA/AC}$, where AC refers to tRNA anticodon. For example, $tRNA^{Gly/GCC}$ refers specifically to $tRNA^{Gly}$ with a GCC anticodon. The general notation of a tRNA is AA_2-$tRNA^{AA1/AC}$, where AA_1 is the amino acid the tRNA is supposed to carry, AA_2 is the amino acid that is actually carried by the tRNA, and AC is the anticodon. In most cases, AA_1 and AA_2 are the same. However, there are two cases where AA_1 and AA_2 can be different. The first is modification of AA_2 by a biochemist. The second occurs naturally in a number of species across all three domains of life (Sheppard et al. 2008; Yuan et al. 2008), where Gln-$tRNA^{Gln}$, Asn-$tRNA^{Asn}$, Cys-$tRNA^{Cys}$, and Sec-$tRNA^{Sec}$ are formed indirectly by two steps. Take Gln-$tRNA^{Gln}$ and Asn-$tRNA^{Asn}$, for example. Glu is first misacylated to $tRNA^{Gln}$, and Asp to $tRNA^{Asn}$, to form Glu-$tRNA^{Gln}$ and Asp-$tRNA^{Asn}$, respectively. The resulting misacylated tRNAs are then converted to Gln-$tRNA^{Gln}$ and Asn-$tRNA^{Asn}$ by a group of tRNA-dependent modifying enzyme.

Isoacceptor tRNA is a somewhat confusing term as it may carry two slightly different meanings. It could refer to a single tRNA decoding different synonymous codons, e.g., $tRNA^{Gly/GCC}$ decoding GGC and GGU codons. Alternatively, it could refer to a set of different tRNAs that carry the same amino acid but decode different synonymous codons. For example, $tRNA^{Gly/GCC}$, $tRNA^{Gly/CCC}$, and $tRNA^{Gly/UCC}$ are isoacceptor tRNAs. They all carry amino acid Gly but with different anticodons decoding different synonymous Gly codons. Different isoacceptor tRNAs could decode the same codon. For example, $tRNA^{Gly/CCC}$ decodes GGG, but $tRNA^{Gly/UCC}$ decodes both GGA and GGG, so GGG is decoded by both $tRNA^{Gly/CCC}$ and $tRNA^{Gly/UCC}$. Thus, isoacceptor tRNA refers to (1) one tRNA decoding different synonymous codons or (2) a set of tRNAs that carry the same amino acid but decode different sets of synonymous codons. The intersection of different sets of synonymous codons may not be empty. For example, the set of codons decoded by $tRNA^{Gly/CCC}$ is {GGG}, and the set of codons decoded by $tRNA^{Gly/UCC}$ is {GGA, GGG}. The intersection of the two sets is {GGG}.

Related to isoacceptor tRNA is another potentially confusing concept, near-cognate tRNA, which is defined in two ways. The first is based on empirical

evidence. If codon XYZ encoding amino acid AA1 can be misread by tRNA carrying amino acid AA2 (AA1 ≠ AA2), then that tRNA is a near-cognate tRNA for codon XYZ. The second definition is based on nucleotide similarity among codons. A codon XYZ has nine XYZ-like codons which differ from XYZ by a single nucleotide. Some of these XYZ-like codons are synonymous to XYZ and some not. The set of tRNAs that can decode any of those nonsynonymous XYZ-like codons are near-cognate tRNAs for codon XYZ because they can "potentially" misread codon XYZ. For example, $tRNA^{Asp}$ is a near-cognate for codons GAA and GAG because Asp is encoded by GAC and GAU which are GAA-like and GAG-like codons.

1.1.2 Genetic Codes and Associated Concepts and Definitions

It is through genetic code that the 64 codons are interpreted as encoding amino acids or translation stop. Nature is superfluous in her creation of genetic code. There are now 24 known genetic codes listed from 1 to 31 (Table 9.1). The standard genetic code is shown previously in Table 2.7.

Some codons do not change their meanings, e.g., Phe (UUY), Tyr (UAY), and Pro (CCN), whereas some others change their meaning frequently. Table 9.2 lists those codons with different meanings in different genetic codes. These codons tend to end with a purine, except for CUN. However, even within the CUR codon family, CUR codons are involved in recoding more often than CUY codons (Table 9.2).

We can build a distance tree from Table 9.2 by counting the pairwise number of reassignment events (i.e., when a codon for one amino acid is reassigned to a different amino acid or a stop codon). The only problem is how to treat reassignment between a sense codon and a stop. Such a change probably should occur less frequently than reassignments involving two sense codons. All pairwise comparisons among the 24 rows (24 genetic codes) generate 609 reassignments involving 2 sense codons and 445 reassignments between a sense codon and a stop codon. However, during the long evolutionary time, the more frequent reassignments will erase each other and the frequencies of their occurrences will be underestimated. So the actual difference between the two numbers must be much greater. If we count each reassignment between a sense codon and a stop codon as equivalent to four reassignments between two sense codons, we obtain a distance-based tree in Fig. 9.1. The topology remains the same if we treat each reassignment between a sense codon and a stop codon as equivalent to two, three, or five reassignments involving two sense codons.

Most bacteria use genetic code 11 which is the same as the standard code except for the difference in start codon usage. The wall-less bacteria including *Mycoplasma* and *Spiroplasma* use genetic code 4 which is identical to the mitochondrial genetic code used in a number of fungal lineages, red algae, and protozoa. The use of the same genetic code 4 by bacteria and mitochondria in eukaryotic lineages suggests two alternative hypotheses. First, it is convergence. Second, the ancestor of

Table 9.1 The 24 genetic tables named after representative species and corresponding translation tables (TT)

Name	TT
Standard	1
Vertebrate mitochondrial	2
Yeast mitochondrial	3
Mold, protozoan, and coelenterate mitochondrial code and the *Mycoplasma*/*Spiroplasma*	4
Invertebrate mitochondrial	5
Ciliate, Dasycladacean, and *Hexamita* nuclear	6
Echinoderm and flatworm mitochondrial	9
Euplotid nuclear	10
Bacterial, archaeal, and plant plastid	11
Alternative yeast nuclear	12
Ascidian mitochondrial	13
Alternative flatworm mitochondrial	14
Chlorophycean mitochondrial	16
Trematode mitochondrial	21
Scenedesmus obliquus mitochondrial	22
Thraustochytrium mitochondrial	23
Pterobranchia mitochondrial	24
Candidate division SR1 and *Gracilibacteria*	25
Pachysolen tannophilus nuclear	26
Karyorelict nuclear	27
Condylostoma nuclear	28
Mesodinium nuclear	29
Peritrich nuclear	30
Blastocrithidia nuclear	31

mitochondrial lineages in Cluster 3 (Fig. 9.1) is a *Mycoplasma*-like bacteria. This would imply multiple origin of mitochondrial lineages.

The main arguments for a single origin of mitochondria are (1) extensive phylogenetic reconstruction with rRNA sequences from diverse array of mitochondrial and bacterial lineages appears to recover mitochondrial lineages as a monophyletic taxon, with its closest phylogenetic relative being in *Alphaproteobacteria* lineages, especially *Rickettsiales* (Williams et al. 2007), and (2) all diverse mitochondrial genomes appear to represent reduced form of the mitochondrial genome from *Reclinomonas americana* (Lang et al. 1997). In particular, the closest phylogenetic relative for the mitochondrial genome from *R. Americana* among bacterial lineages is *Ehrlichia muris* strain AS145 within *Rickettsiales*. These lines of evidence, taken together, represent compelling evidence for the single-origin hypothesis of mitochondria.

Genetic codes also differ in start codons (Table 9.3). While AUG is used universally and dominantly as a start codon, other codons are used as well, although there has been no species in which a non-AUG codon is used as a start codon more

Table 9.2 Codons with different meanings in different translation tables (TT)

TT	UUA	UCA	UAA	UAG	UGA	CUU	CUC	CUA	CUG	AUA	AAA	AGA	AGG
1	L	S	*	*	*	L	L	L	L	I	K	R	R
2	L	S	*	*	W	L	L	L	L	M	K	*	*
3	L	S	*	*	W	T	T	T	T	M	K	R	R
4	L	S	*	*	W	L	L	L	L	I	K	R	R
5	L	S	*	*	W	L	L	L	L	M	K	S	S
6	L	S	Q	Q	*	L	L	L	L	I	K	R	R
9	L	S	*	*	W	L	L	L	L	I	N	S	S
10	L	S	*	*	C	L	L	L	L	I	K	R	R
11	L	S	*	*	*	L	L	L	L	I	K	R	R
12	L	S	*	*	*	L	L	L	S	I	K	R	R
13	L	S	*	*	W	L	L	L	L	M	K	G	G
14	L	S	Y	*	W	L	L	L	L	I	N	S	S
16	L	S	*	L	*	L	L	L	L	I	K	R	R
21	L	S	*	*	W	L	L	L	L	M	N	S	S
22	L	*	*	L	*	L	L	L	L	I	K	R	R
23	*	S	*	*	*	L	L	L	L	I	K	R	R
24	L	S	*	*	W	L	L	L	L	I	K	S	K
25	L	S	*	*	G	L	L	L	L	I	K	R	R
26	L	S	*	*	*	L	L	L	A	I	K	R	R
27	L	S	Q	Q	*w*	L	L	L	L	I	K	R	R
28	L	S	q	q	*w*	L	L	L	L	I	K	R	R
29	L	S	Y	Y	*	L	L	L	L	I	K	R	R
30	L	S	E	E	*	L	L	L	L	I	K	R	R
31	L	S	e	e	W	L	L	L	L	I	K	R	R

A small-case letter, such as *q* in translation table 28, means that the corresponding codon can mean either amino acid Q or a stop codon

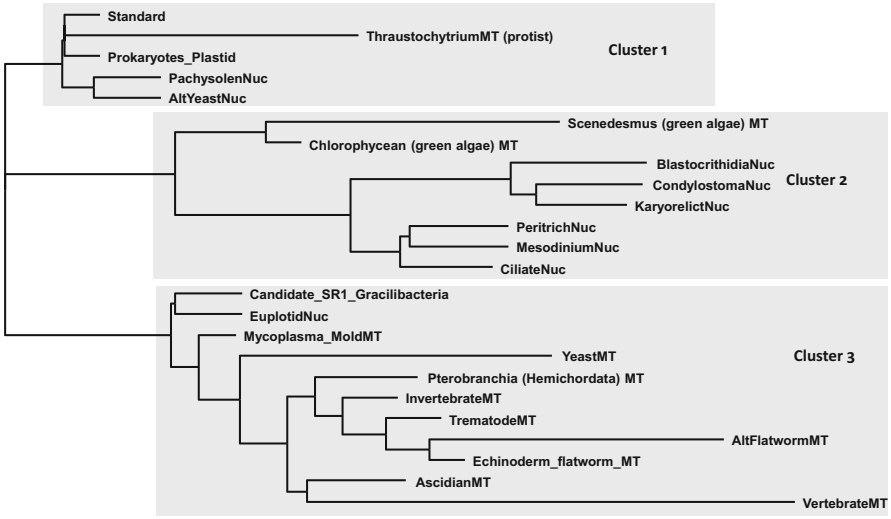

Fig. 9.1 "Phylogenetic tree" of 24 genetic codes with their differences shown in Table 9.2, based on pairwise number of codon reassignments. A reassignment between a sense codon and a stop codon is treated as equivalent to four codon reassignment events between two nonsynonymous sense codons. Leaves labeled with a "MT"-ending are mitochondrial genetic codes

frequently than AUG. For eukaryotic species where AUG is part of translation initiation signal such as in the Kozak consensus RxxAUGG, non-AUG codons are rarely used. In bacterial species where start codon is localized by pairing of Shine-Dalgarno (SD) sequences and anti-SD sequences, the requirement for AUG as a start codon is less stringent.

A synonymous codon family refers to all codons coding the same amino acids. For example, GGA, GGC, GGG, and GGU codons all code Gly and are collectively referred to as the Gly codon family or just Gly family. I may use "family" for "synonymous codon family" when there is no confusion. A codon family such as Gly family that differs only at the third codon position is a simple family. The Gly codon family is a simple family. In contrast, a codon family that differs not only at the third codon position but also at other codon positions is a compound codon family. For example, in standard genetic code, Leu is coded by UUR (where R stands for purine) and CUN (where N stands for any nucleotide) codons. Therefore, Leu codon family is a compound family. Other compound families in the standard code include Ser (coded by UCN and AGY, where Y stands for pyrimidine) and Arg (coded by CGN and AGR). Compound families are often divided into subfamilies. For example, the Ser family is broken into UCN subfamily and AGY subfamily.

The phenomenon that one amino acid may be encoded by multiple codons is called codon degeneracy. This gives rise to 4-fold, 3-fold, 2-fold, and 1-fold (0-fold is a misnomer) degenerate sites. An *n*-fold site is one that can be occupied by *n* different nucleotides without changing the meaning of the encoded amino acid. For example, the third site in the four Gly codons above is fourfold degenerate. In the

Table 9.3 The 24 translation tables (24) differ in start codon usage

TT	TTA	TTG	CTG	ATT	ATC	ATA	ATG	GTG
1	–	M	M	–	–	–	M	–
2	–	–	–	M	M	M	M	M
3	–	–	–	–	–	M	M	–
4	M	M	M	M	M	M	M	M
5	–	M	–	M	M	M	M	M
6	–	–	–	–	–	–	M	–
9	–	–	–	–	–	–	M	M
10	–	–	–	–	–	–	M	–
11	–	M	M	M	M	M	M	M
12	–	–	M	–	–	–	M	–
13	–	M	–	–	–	M	M	M
14	–	–	–	–	–	–	M	–
16	–	–	–	–	–	–	M	–
21	–	–	–	–	–	–	M	M
22	–	–	–	–	–	–	M	–
23	–	–	–	M	–	–	M	M
24	–	M	M	–	–	–	M	M
25	–	M	–	–	–	–	M	M
26	–	–	M	–	–	–	M	–
27	–	–	–	–	–	–	M	–
28	–	–	–	–	–	–	M	–
29	–	–	–	–	–	–	M	–
30	–	–	–	–	–	–	M	–
31	–	–	–	–	–	–	M	–

standard code, AUA, AUC, and AUU all encode amino acid Met, so that the third codon site is threefold degenerate. AAA and AAG both encode amino acid Lys, so that the third codon site is twofold degenerate. We may also have a twofold degenerate site at the first codon site. For example, both CUA and UUA encode amino acid Leu, so the first codon site is twofold degenerate. The second codon site of Gly codons is onefold degenerate because replacing it by any other nucleotide will change the meaning of the encoded amino acid.

A synonymous mutation refers to the change of a codon by another synonymous codon. A nonsynonymous mutation refers to codon replacement involving amino acid replacement. A substitution is a mutation that has spread to all individuals in the population. Synonymous substitutions occur often, but nonsynonymous substitutions occur rarely.

Throughout text, we will abbreviate highly and lowly expressed genes as HEGs and LEGs. Unless specified otherwise, HEGs and LEGs in this chapter pertain to protein expression, not mRNA expression. One may rank all proteins according to experimentally measured abundance and take the top and bottom 1/3 as HEGs and

LEGs, respectively. Non-HEGs are simply all genes from a genome that is not included in HEGs. Protein abundance values for most model species may be found in PaxDb (Wang et al. 2012).

1.2 Elongation Efficiency Depends on Amino Acid and Codon Usage

Many unicellular organisms, especially bacterial species, need to grow and replicate the cell rapidly in order not to be outcompeted by others. For example, an *E. coli* cell replicates once every 20 min with unlimited nutrients. To replicate a cell, not only the genome needs to be replicated, but a large amount of proteins have to be produced, with some proteins produced in nearly half a million copies in an *E. coli* cell. For such highly expressed proteins, it is very important for their coding genes to have efficient coding strategy to maximize the rate of translation. Translation involves three sub-processes, initiation, elongation, and termination. The previous chapter illustrates how natural selection can drive evolution toward more efficient translation initiation. This chapter addresses the question of how translation elongation can be improved through codon adaptation.

There are two obvious ways of increasing translation elongation efficiency for mass-produced proteins. The first is to optimize amino acid usage, i.e., to use energetically cheap and typically abundant amino acids as building blocks (Akashi and Gojobori 2002). The second is to maximize the usage of codons that match the anticodon of the most abundant cognate tRNA (Gouy and Gautier 1982; Ikemura 1992; Xia 1998a, 2005, 2009, 2015). For example, the amino acid glycine (Gly) can be coded by GGA, GGC, GGG, and GGU codons, but $tRNA^{Gly}$ species that decode GGY codons are more abundant than $tRNA^{Gly}$ species that decode GGR codons in *E. coli* cells. What codons should *E. coli* use to code glycine? Obviously natural selection should favor those that maximize the usage of GGY codons against GGR codons given the differential tRNA availability. However, selection and mutation may go in opposite directions, so any study of codon adaptation would be incomplete without considering both selection and mutation.

1.3 Empirical Illustration of Codon-Anticodon Adaptation

Ikemura's pioneering works established the relationship between differential tRNA abundance and its effect on codon usage in rapidly replicating bacterial species and unicellular eukaryotes (Ikemura 1981a, b, 1982, 1992). Many studies have since demonstrated a strong relationship not only between codon adaptation and gene expression (Coghlan and Wolfe 2000; Comeron and Aguade 1998; Duret and Mouchiroud 1999; Gouy and Gautier 1982; Xia 2007c) but also between

experimentally modified codon usage and protein production (Haas et al. 1996; Ngumbela et al. 2008; Robinson et al. 1984; Sorensen et al. 1989). These results have led to the explicit formulation of codon-anticodon coevolution and adaptation theory (e.g., Akashi 1994; Moriyama and Powell 1997; Ran and Higgs 2012; Xia 1998a, 2008) which states that (1) protein production is rate-limited by both translation initiation and elongation efficiency; (2) codon usage and tRNA anticodon coevolve to adapt to each other, resulting in increased production of correctly translated proteins; and (3) the increased elongation efficiency and accuracy represent the driving force for the HEGs to acquire a high degree of codon-anticodon adaptation.

1.3.1 Empirical Illustration of Codon-Anticodon Adaptation in Yeast

The baker's yeast, *Saccharomyces cerevisiae*, replicates rapidly and is expected to use codons with many decoding tRNAs and avoid codons with few decoding tRNAs. The earliest association between tRNA and codon usage was empirically demonstrated by Ikemura (1981a, b, 1992). Tables 9.4 and 9.5 show the association between tRNA gene copy number (T in Tables 9.4 and 9.5) in the genome and codon usage in highly expressed yeast genes (F in Tables 9.4 and 9.5). T is a good proxy for tRNA abundance (Percudani et al. 1997).

The association between T and F is obvious in Tables 9.4 and 9.5. Take the two Arg codons AGA and AGG in Table 9.4, for example. There are 11 $tRNA^{Arg/UCU}$ genes in the yeast genome that form perfect Watson-Crick base pair with AGA but

Table 9.4 Copy number of tRNA genes in the yeast *Saccharomyces cerevisiae* genome (T) and codon counts (F) in highly expressed yeast protein-coding genes, compiled in the Eyeastcai.cut file distributed with EMBOSS (Rice et al. 2000)

AA[a]	Codon[b]	T	F		AA[a]	Codon[b]	T	F
Arg	AGA	11	314		His	CAC	7	102
Arg	AGG	1	1		His	CAU	0	25
Asn	AAC	10	208		Leu	UUA	7	42
Asn	AAU	0	11		Leu	UUG	10	359
Asp	GAC	16	202		Lys	AAA	7	65
Asp	GAU	0	112		Lys	AAG	14	483
Cys	UGC	4	3		Phe	UUC	10	168
Cys	UGU	0	39		Phe	UUU	0	19
Gln	CAA	9	153		Ser	AGC	2	6
Gln	CAG	1	1		Ser	AGU	0	4
Glu	GAA	14	305		Tyr	UAC	8	141
Glu	GAG	2	5		Tyr	UAU	0	10

Only twofold codon families are included
[a]Amino acid carried by tRNA
[b]Codons forming Watson-Crick base pair with the anticodon of tRNA

Table 9.5 Copy number of tRNA genes in the yeast *Saccharomyces cerevisiae* genome (T) and codon counts (F) in highly expressed yeast protein-coding genes, compiled in the Eyeastcai.cut file distributed with EMBOSS (Rice et al. 2000)

AA	Codon	T	F	AA	Codon	T	F
Ala	GCA	5	6	Pro	CCA	10	211
Ala	GCG	0	0	Pro	CCG	0	0
Ala	GCC	0	130	Pro	CCC	0	2
Ala	GCU	11	411	Pro	CCU	2	10
Arg	CGA	0	0	Ser	UCA	3	7
Arg	CGG	1	0	Ser	UCG	1	1
Arg	CGC	0	0	Ser	UCC	0	133
Arg	CGU	6	43	Ser	UCU	11	192
Gly	GGA	3	1	Thr	ACA	4	2
Gly	GGG	2	2	Thr	ACG	1	1
Gly	GGC	16	9	Thr	ACC	0	164
Gly	GGU	0	459	Thr	ACU	11	151
Ile	AUA	2	0	Val	GUA	2	0
Ile	AUC	0	181	Val	GUG	2	5
Ile	AUU	13	149	Val	GUC	0	231
Leu	CUA	3	14	Val	GUU	14	278
Leu	CUG	0	1				
Leu	CUC	1	1				
Leu	CUU	0	2				

Only threefold and fourfold codon families are included. Symbols as in Table 9.4

only one $tRNA^{Arg/CCU}$ with AGG. So we expect yeast genes, especially highly expressed ones, to use AGA and avoid AGG, which is true (Table 9.4). The same applies to all other synonymous codon families or subfamilies, except for the Cys codon family. Why the rarely used Cys codon family should be exceptional remains unknown. It is possible that Cys codon UGC may happen to be followed by a GNN codon, leading to methylation of C at the third codon position which then changes to T via spontaneous deamination. Whether the yeast genome has cytosine methylation remains controversial, with both evidence for (Tang et al. 2012) and against (Capuano et al. 2014) the existence of methylation in *S. cerevisiae*. However, there is significant CpG deficiency and TpG and CpA surplus in genome, which is consistent with CpG-specific DNA methylation.

One can obtain tables similar to Tables 9.4 and 9.5 by downloading the yeast genome from GenBank and then using DAMBE to compile the data in three steps. First, read the GenBank files for yeast chromosome sequences into DAMBE (Xia 2013, 2017d) to extract the coding sequences (CDSs) and tRNA genes. Second, compute I_{TE} (Xia 2015) as a proxy of gene expression, and choose a subset of CDSs with highest I_{TE} as HEGs. Third, use DAMBE to obtain codon usage of these HEGs. In this way, a table similar to Table 9.4 can be generated in minutes.

1.3.2 Codon Usage Changes When tRNA Abundance Changes

An evolutionary change in tRNA composition or relative abundance is expected to alter codon-anticodon adaptation. This is not controversial theoretically, but empirically difficult to demonstrate. However, recent studies (Xia 2012c; Xia et al. 2007) have documented that changes in $tRNA^{Met}$ genes (where Met is the amino acid carried by the tRNA) in animal mitochondrial DNA (mtDNA) are associated with changes in Met codon usage.

In mtDNA of most animal species, Met is coded by AUA and AUG codons. In some animal species, e.g., vertebrates, these two codons are translated by a single $tRNA^{Met/CAU}$ species (where CAU is the anticodon in the 5′ to 3′ orientation) with a modified C (i.e., f^5C) at the first anticodon position (Grosjean et al. 2010) to allow C/A pairing. In other animal species, e.g., tunicates, an additional $tRNA^{Met/UAU}$ gene is present in the mtDNA. One would expect that, when $tRNA^{Met/UAU}$ is absent, Met should be preferably coded by AUG with a reduced AUA usage. The gain of $tRNA^{Met/UAU}$ would favor more Met to be coded by AUA.

In addition to tunicates, MtDNA in bivalve species also have two $tRNA^{Met}$ genes. In some bivalve species (e.g., *Acanthocardia tuberculata*, *Crassostrea gigas*, *C. virginica*, *Hiatella arctica*, *Placopecten magellanicus*, and *Venerupis philippinarum*), both $tRNA^{Met}$ genes have a CAU anticodon forming Watson-Crick base pair with codon AUG. In some other bivalve species (e.g., *Mytilus edulis*, *Mytilus galloprovincialis*, and *Mytilus trossulus*), one $tRNA^{Met}$ has a CAU anticodon, and the other has a UAU anticodon forming Watson-Crick base pair with the AUA codon. One would predict that the latter should be more likely to code Met by AUA than the former, i.e., the proportion of AUA codon within the AUR codon family, designated P_{AUA}, should be greater in the latter with both a $tRNA^{Met/CAU}$ and a $tRNA^{Met/UAU}$ gene than in the former with $tRNA^{Met/CAU}$ gene only (Xia et al. 2007).

One complication in testing the prediction is that AUA usage will increase with genomic AT%. To control for this effect, one may use another A-ending codon, such as UUA as a reference. Thus, given the same P_{UUA} (the proportion of UUA codon in the UUR codon family), P_{AUA} in the three Mytilus mtDNA with both a $tRNA^{Met/CAU}$ and a $tRNA^{Met/UAU}$ gene should be higher than that in the six bivalve species without a $tRNA^{Met/UAU}$ gene. This is supported by empirical evidence (ANCOVA test, $p = 0.0111$, Fig. 9.2a). Thus, the presence of $tRNA^{Met/UAU}$ increases AUA usage significantly.

A similar comparison can be performed between the urochordates (tunicates, with both $tRNA^{Met/CAU}$ and $tRNA^{Met/UAU}$ genes in their mtDNA) and cephalochordates (lancelets, with only a $tRNA^{Met/CAU}$ gene in their mtDNA). Figure 9.2b shows that P_{AUA} is much smaller in lancelets than in tunicates at the same P_{UUA} level. Thus, AUA usage is consistently increased by the gain of a $tRNA^{Met/UAU}$ gene (or consistently decreased by the loss of a $tRNA^{Met/UAU}$ gene) in animal mtDNA.

A gain of a $tRNA^{Met/UAU}$ gene is also associated with a surplus of AUG→AUA substitutions in animal mitochondrial coding sequences (results not shown). Similar associations can also be observed with other gain/loss of tRNA genes in animal

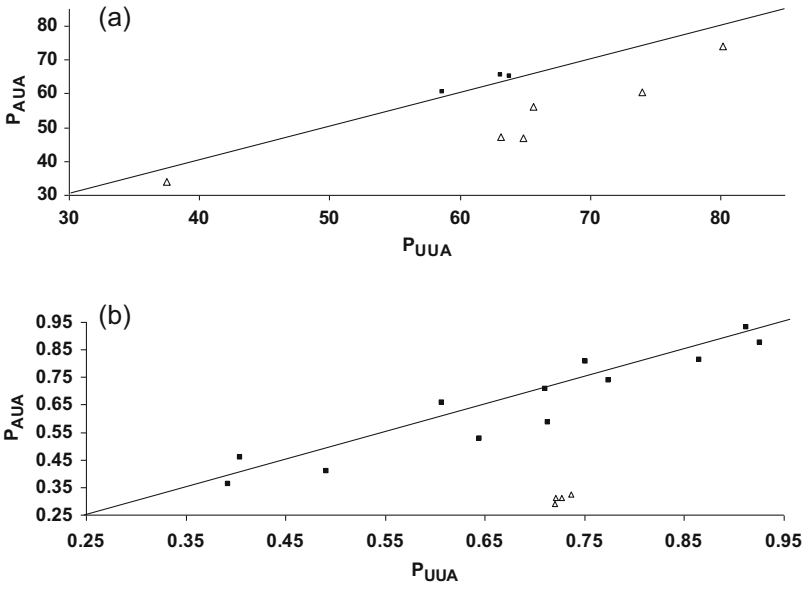

Fig. 9.2 Relationship between PAUA and PUUA, highlighting the observation that PAUA is greater when both a $tRNA^{Met/CAU}$ and a $tRNA^{Met/UAU}$ are present than when only $tRNA^{Met/CAU}$ is present in the mtDNA, for bivalve species (**a**) and chordate species (**b**). The filled squares are for mtDNA containing both $tRNA^{Met/CAU}$ and $tRNA^{Met/UAU}$ genes, and the open triangles are for mtDNA without a $tRNA^{Met/UAU}$ gene

mitochondrial. In contrast, a gain/loss of tRNA genes in plant mtDNA appears to have little effect on nucleotide substitutions or codon usage, presumably because such gain/loss events do not significantly alter the tRNA pool in plant cells where nuclear tRNAs are mass-imported into plant mitochondria.

1.4 Effect of Biased Mutation on Codon Usage and Some Misconceptions

Biased mutation has long been known to affect codon usage (Muto and Osawa 1987; Sueoka 1964; Xia and Yuen 2005; Xia et al. 2002). The third codon position is the most amenable to mutation bias (Fig. 9.4) because most nucleotide substitutions at the third codon position are synonymous. Nucleotide substitutions are synonymous at some first codon positions but nonsynonymous at all second codon position. Furthermore, all nucleotide substitutions at the second codon positions typically involve rather different amino acids and therefore should be subject to strong purifying selection (Xia 1998b; Xia and Li 1998). One therefore would predict that the third codon position should increase more rapidly with the genomic GC%

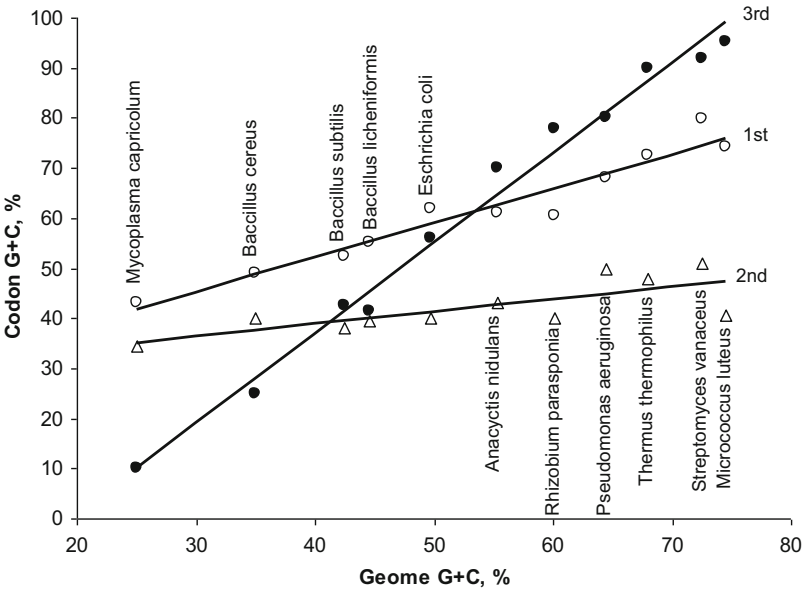

Fig. 9.3 Correlation of GC% between genomic DNA and first, second, and third codon positions (Muto and Osawa 1987). While the actual position of the points may be substantially revised with new genomic data (e.g., the GC% for the first, second, and third codon positions for *Mycoplasma capricolum* is 35.8%, 27.4%, and 8.8% based on all annotated CDSs in the genomic sequence), the general trend remains the same

than the first codon position which in turn should have its GC% increase more rapidly with the genomic GC% than the second codon position. The empirical results (Fig. 9.3) strongly support the prediction (Muto and Osawa 1987).

However, the pattern in Fig. 9.3, while consistent with the mutation hypothesis, has resulted in two misconceptions. First, the pattern shown by the third codon position is often interpreted to reflect mutation bias. This interpretation is incorrect because the third codon position is subject to selection by differential availability of tRNA species (Carullo and Xia 2008; Xia 1998a, 2005, 2008; Xia et al. 2007). We may contrast a GC-rich *Streptomyces coelicolor* and a GC-poor *Mycoplasma capricolum* as an illustrative example. *M. capricolum* has no tRNA with a C or G at the wobble site for fourfold codon families (Ala, Gly, Pro, Thr, and Val), i.e., the translation machinery would be inefficient in translating C-ending or G-ending codons. This implies selection in favor of A-ending or U-ending codons and will consequently reduce GC% at the third codon position. This most likely has contributed to the low GC% at the third codon position in *M. capricolum.* In contrast, most of the tRNA genes translating the five fourfold codon families in the GC-rich *S. coelicolor* have G or C at the wobble site, and should favor the use of C-ending or G-ending codons. This most likely has contributed to the high GC% at the third codon position in *S. coelicolor.* In these two cases, mutation bias and tRNA-mediated

selection are in the same direction to drive up or down GC% at the third codon position. The same pattern is observed for twofold codon families. The most conspicuous one is the Gln codon family (CAA and CAG). There is only one $tRNA^{Gln}$ gene in *M. capricolum* with a UUG anticodon favoring the CAA codon. In contrast, there are two $tRNA^{Gln}$ in *S. coelicolor*, both with a CUG anticodon favoring the CAG codon. Thus, the high slope for the third codon position in Fig. 9.3 is at least partially attributable to the tRNA-mediated selection. Relative contribution of mutation and tRNA-mediated selection to codon usage has been evaluated in several recent studies (Carullo and Xia 2008; Xia 2005, 2008; Xia et al. 2007).

The second misconception arising from Fig. 9.3 is that the frequency of G-ending and C-ending codons will increase and A-ending and U-ending codons decrease, with genomic GC% or GC-biased mutation (Kliman and Bernal 2005). This is not generally true (Palidwor et al. 2010). Take the arginine codons, for example. Given the transition probability matrix for the six synonymous codons shown in Table 9.6, the equilibrium frequencies (π) for the six codons are

$$\begin{aligned} \pi_{AGA} &= \frac{1}{2k^2+3k+1} \\ \pi_{AGG} = \pi_{CGA} = \pi_{CGT} &= \frac{k}{2k^2+3k+1} \\ \pi_{CGC} = \pi_{CGG} &= \frac{k^2}{2k^2+3k+1} \end{aligned} \tag{9.1}$$

The three solutions correspond to the number of GC in the codon, with AGA having one, AGG, CGA and CGT having two, and CGC and CGG having three G or C. One may note that the G-ending codon AGG has the same equilibrium frequency as that of the A-ending CGA and the T-ending CGT. Thus, we should not expect A-ending or T-ending codons to always decrease or G-ending and C-ending codons always increase, with increasing genomic GC% or GC-biased mutation. In fact, according to the solutions in Eq. (9.1), π_{AGG}, π_{CGA}, and π_{CGT} will first increase with k until k reaches $\sqrt{2}/2$ and will then decrease with k when $k > \sqrt{2}/2$ (Palidwor et al. 2010).

1.5 Two Hypotheses on Translation Elongation Efficiency

It is controversial as to what degree is protein production limited by translation elongation. Early theoretical considerations (Andersson and Kurland 1983; Bulmer 1990, 1991; Liljenstrom and von Heijne 1987) tend to favor the argument that translation elongation is not rate-limiting in protein production, but translation initiation is. This hypothesis does not deny the existence of codon adaptation, but it asserts that codon-anticodon adaptation and increased elongation efficiency are not related to protein production. Instead, the benefit of codon adaptation and increased elongation efficiency is to increase ribosomal availability for global translation. This

Table 9.6 Transition probability matrix for the six synonymous arginine codons, with α for transitions (C↔T and A↔G), β for transversions, and k modeling AT-biased mutation ($0 \leq k \leq 1$) or GC-biased mutation ($k > 1$)

	CGT	CGC	CGA	CGG	AGA	AGG
CGT		$k\alpha$	β	$k\beta$	0	0
CGC	α		β	β	0	0
CGA	β	$k\beta$		$k\alpha$	β	0
CGG	β	β	α		0	β
AGA	0	0	$k\beta$	0		$k\alpha$
AGG	0	0	0	$k\beta$	α	

We ignore nonsynonymous substitutions because nonsynonymous substitution rate is often negligibly low compared to synonymous rate. The diagonal is constrained by the row sum equal to 1

hypothesis was explicitly formulated only recently and empirically tested (Kudla et al. 2009).

We thus have two alternative hypotheses attributing different benefits to codon-anticodon adaptation. The first assumes that protein production is rate-limited by both initiation and elongation and codon-anticodon adaptation would result in higher elongation efficiency and more efficient and accurate protein production, especially for HEGs. The second claims that protein production is rate-limited only by initiation efficiency but improved codon adaptation and consequently increased elongation efficiency have the benefit of increasing ribosomal availability for global translation.

How should we go about testing these two hypotheses? Note that the two hypotheses make different predictions about the relationship among three variables: (1) translation initiation efficiency, (2) translation elongation efficiency, and (3) protein production. Before we can test these two hypotheses, we need to understand how these variables can be measured. The previous chapter outlines a few factors contributing to translation initiation efficiency. Here we first learn a few indices of codon usage bias as a proxy for translation elongation efficiency and then include them in the test of the two hypotheses in the section illustrating the application of index of translation elongation (Xia 2015).

1.6 Wobble Hypothesis and Its Extensions

The wobble hypothesis is proposed to explain how a set of tRNA molecules can decode all sense codons which are much larger in number. The wobble-pairing rules are specified in Fig. 9.4, together with the numbering system used here for individual codon and anticodon sites that is more precise than, but different from, the conventional one. The original wobble hypothesis (Crick 1966), with its extended codon-anticodon base pairs (Fig. 9.4), played a crucial role in understanding the working of the translation machinery. It explains why $tRNA^{Ile/IAU}$, where I in IAU is inosine derived from A, is able to translate all three Ile codons (AUC, AUU, and, albeit

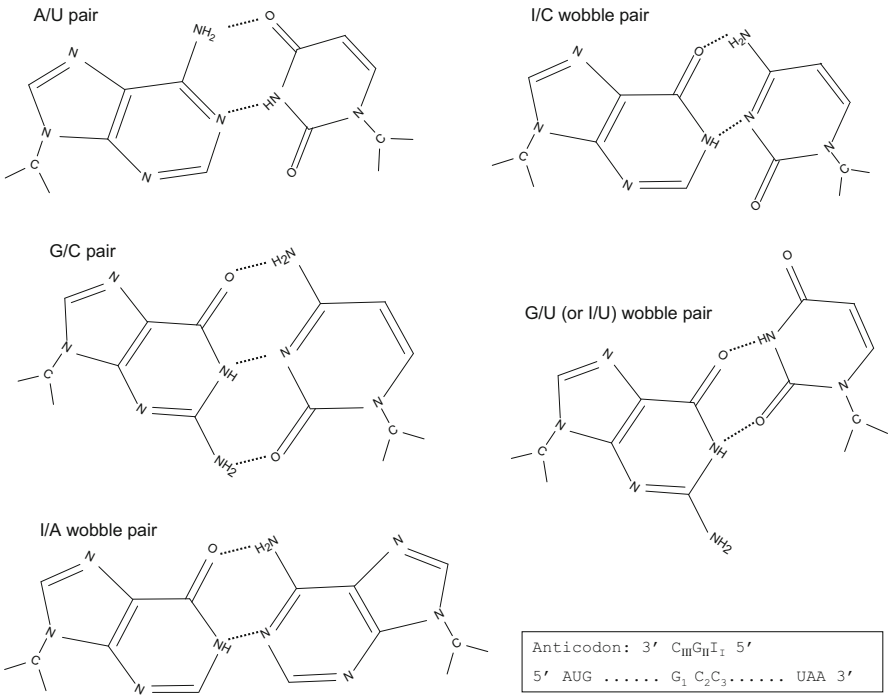

Fig. 9.4 Base pairs between nucleotides at the first anticodon site (which can have I, G, C, U but rarely A) and the third codon site. The inset shows the site numbering system of codon and anticodon, with codon sites subscripted with 1, 2 and 3 and anticodon sites subscripted with I, II, and III, which is illustrated by the paring of II/C3, GII/C2, CIII/G1.

inefficiently, AUA), why a tRNA with a G_I can translate Y-ending codons (where Y stands for C or U), and why a tRNA with a U_I can translate R-ending codons (where R stands for A or G). The hypothesis also explains the lack of A_I in tRNA genes for decoding twofold Y-ending codon family because such a tRNA, when its A_I is modified to I_I, would misread the near-cognate R-ending codons.

Wobble pairing reduces the number of tRNAs needed for translation and simplifies the translation machinery. As an example of parsimonious tRNA usage, the Y-ending codons, be they in twofold or fourfold codon families, are decoded by tRNAs with either a I_I or a G_I, but never both. This rule is obeyed in all three kingdoms of life. Almost all fourfold codon families in *Mycoplasma pulmonis* (including the Ser UCN codon family and Leu CUN codon family) are decoded by a single tRNA species with a U_I, except for the Thr ACN and Arg CGN codon families which are each decoded by two tRNA species, one with a U_I and other with a G_I. The most dramatic simplification of tRNome is observed in vertebrate mitochondria, e.g., vertebrate mitochondrial genomes which contain only 22 tRNA genes, with each tRNA species decoding a codon family. Instead of separate initiation $tRNA^{iMet/CAU}$ and elongation $tRNA^{eMet/CAU}$ present in all nuclear genomes, a single

$tRNA^{Met/CAU}$, with a modified C_I, decodes both the initiation AUG codon and internal Met AUR codons. Each Y-ending codon family is decoded by a single tRNA species with a wobble G_I and each R-ending codon family by a single tRNA with a wobble U_I which is modified to prevent its pairing with U or C. All fourfold codon families are decoded by a tRNA with a wobble U_I which is not modified.

Wobble pairing is not without cost as it often reduces translation efficiency and accuracy and is generally avoided (Xia 2008). For example, an I_I/A_3 pair is bulky because it involves two purines (Fig. 9.4) in contrast to other base pairs which typically involve a large purine and a small pyrimidine. For this reason, Ile is rarely coded by AUA except for certain viruses with a strong A-biased mutation (van Weringh et al. 2011). Among a set of highly expressed genes in the yeast (*Saccharomyces cerevisiae*), AUA is not used at all (Table 9.5). Similarly, a tRNA with a U_I can translate A-ending codons better than G-ending codons (Grosjean et al. 2010; Xia 2008). Most of the yeast $tRNA^{Arg}$ have a U_I, and only one AGG codon is found in contrast to 314 AGA codons in highly expressed yeast genes (Table 9.4). Yeast genomic data also suggest that a tRNA with a G_I can translate C-ending codons better than U-ending codons. For example, the yeast $tRNA^{Asn}$ genes translating the Asn AAY codon family all have a G_I. Among 219 Asn codons in highly expressed yeast genes, only 11 are AAU codons, suggesting strong selection against AAU codons in favor of AAC codons (Table 9.4). Note that the yeast genome is strongly AT-biased. If there is no selection against AAU codons, we would expect more AAU codons than AAC codons, which is contrary to the observed frequencies. However, the selection against G_I/U_3 pair is in general much weaker than that against U_I/G_3 pair. In fungal mitochondrial genomes, there is no avoidance of G_I/U_3 pair in favor of G_I/C_3 pair, although U_I/G_3 pair is strongly avoided in favor of U_I/A_3 pair (Xia 2008). The weak, or lack of, selection against G_I/U_3 can explain several puzzling counterexamples against the codon-anticodon adaptation theory (Bulmer 1991; Ikemura 1981b; Xia 1998a) which states that the most frequently used codon in each synonymous codon family should form Watson-Crick base paring with the anticodon of the most abundant tRNA species to reduce translation error and increase translation efficiency. For example, Cys codons (UGY) are translated by $tRNA^{Cys/GCA}$ in both cytoplasm and mitochondria in the yeast, yet most Cys codons have U_3. If there is little selection against G_I/U_3 pair (i.e., G_I/U_3 pair is as efficient and accurate as G_I/C_3 pair), then the frequencies of UGC and UGU will be mostly determined by AT-bias. Because the yeast nuclear and mitochondrial genomes are both AT-rich, we have more UGU codons than UGC codons, in spite of G_I in $tRNA^{Cys}$. The weak selection against G_I/U_3 but strong selection against U_I/G_3 also explains why Y-ending codons are typically translated by a tRNA with a G_I, whereas R-ending codons are typically translated by two different tRNAs, one with a U_I and the other with a C_I (Xia 2008).

The wobble hypothesis points to the necessity of nucleotide modification in tRNA to either increase or decrease the wobble versatility to improve accuracy and efficiency of translation. The observation that an unmodified U_I can pair with all N_3 in many mitochondrial genomes suggests that U_I in tRNA for twofold R-ending codon families needs to be modified to restrict its wobble versatility to

avoid misreading the near-cognate Y-ending codons. Chemical modification of U_I to restrict its pair versatility to R_3 in twofold R-ending codon family is universal in all three kingdoms of life and in organelles (Grosjean et al. 2010; Lim 1994). On the other hand, the $tRNA^{Met/CAU}$ in vertebrate mitochondria need to read both the initiation AUG codon and the internal AUG and AUA codons, and its C_I is modified to f^5C_I to increase its wobble versatility so as to form a f^5C_I/A_3 pairing between the anticodon and the AUA codon. Nucleotide modification in tRNA has been extensively reviewed (Grosjean et al. 2010) and chemically detailed in MODOMICS (Czerwoniec et al. 2009).

Wobble pairing implies the theoretical possibility of adding new base pairs of novel nucleotides to protein-coding genes to increase the coding capacity (Hirao and Kimoto 2010). A single novel base pair, involving two novel nucleotides, would increase the number of codons from 64 to 216 ($=6^3$), and one can then use these extra codons, together with engineered tRNAs to recognize these codons and to carry new amino acid analogs, to produce novel proteins.

The wobble hypothesis can be extended to explain the lack of UCG anticodon in Arg CGN codon family in a large number of evolutionary lineages. A tRNA species with a wobble U_I is almost always present among tRNA species decoding fourfold codon families and twofold R-ending codon families, with most exceptions observed in the Arg CGN codon family. In the mitochondrial genomes of *Caenorhabditis elegans* (metazoan), *Marchantia polymorpha* (plant), *Pichia canadensis* (fungus), and *Saccharomyces cerevisiae* (fungus), there is no $tRNA^{Arg/UCG}$, and Arg CGN codon family is decoded by $tRNA^{Arg/ACG}$ (Xia 2005). The lack of $tRNA^{Arg/UCG}$ in the mitochondrial genome of these diverse taxa suggests that the lack is an ancestral state and that the presence of $tRNA^{Arg/UCG}$ in vertebrate mitochondria is a derived state. This is substantiated by the fact that almost all eubacterial species, from which the mitochondrion was originally derived, lack $tRNA^{Arg/UCG}$ (Grosjean et al. 2010).

The expanded wobble hypothesis for the lack of $tRNA^{Arg/UCG}$ requires an extension of the wobble hypothesis by invoking wobble paring between the third anticodon site (N_{III}) and the first codon site (N_1), conditional on a C_{II}/G_2 or G_{II}/C_2 with three hydrogen bonds. Thus, the anticodon UCG would wobble-pair with stop codon UGA through a wobble G_{III}/U_1 pair and should therefore be strongly selected against (Carullo and Xia 2008). This explains not only the absence of $tRNA^{Arg/UCG}$ in diverse evolutionary lineages but in particular why $tRNA^{Arg/UCG}$ is absent in most eubacterial species and ancestral mitochondrial lineages where UGA is used as a stop codon and why it is present in derived mitochondrial lineages such as vertebrate mitochondrial genomes where UGA is no longer used as a stop codon.

2 Commonly Used Codon Usage Indices

There are two key factors contributing to codon usage bias: the mutation bias (Osawa et al. 1987) and the tRNA-mediated selection (Ikemura 1981a, 1982, 1992; Xia 1998a, 2015). There are also two types of codon usage indices, but they do not

correspond to the two factors shaping codon usage. The first type of codon usage indices is codon-specific best represented by relative synonymous codon usage (RSCU, Sharp et al. 1986), which measures deviation of codon usage from equal usage. The second type of codon usage indices is gene-specific with several well-known representatives including codon adaptation index effective number of codons (ENC, Sun et al. 2013; Wright 1990), codon adaptation index (CAI, Sharp and Li 1987; Xia 2007c), codon bias index (CBI, Bennetzen and Hall 1982), frequency of optimal codons (Fop, Ikemura 1985), tRNA adaptation index (tAI, dos Reis et al. 2004), and index of translation elongation (I_{TE}, Xia 2015).

ENC aims to measure deviation of codon usage from equal usage and may be considered as the gene-specific equivalent of the codon-specific RSCU. They are both descriptive and do not distinguish between mutation bias or tRNA-mediated selection in their contribution to codon usage bias. All other gene-specific indices aim to measure the intensity of the tRNA-mediated selection on codon usage bias. A gene encoding a mass-produced (highly expressed) protein is expected to be under stronger selection to optimize its codon usage corresponding to differential tRNA availability than a gene encoding lowly expressed protein, and we expect CAI, CBI, tAI, and I_{TE} to be greater for the highly expressed gene than the lowly expressed gene. However, CAI, CBI, and tAI ignore background mutation bias. I_{TE} is a generalization of CAI, by incorporating background mutation, and is reduced to CAI when there is no background mutation bias (Xia 2015).

Codon indices that aim to measure tRNA-mediated selection (i.e., CAI, CBI, F_{op}, tAI, and I_{TE}) all define a translationally optimal codon (TOC) within each codon family, and the codon usage index value will be the highest if all codons in a gene are TOCs. However, TOC is defined differently among these indices. CBI, F_{op}, and tRNA define a TOC mainly as one that corresponds to the most abundant isoacceptor tRNA, with CBI incorporating gene expression information as well. CAI defines a TOC as one in its codon family that is used most frequently in HEGs. I_{TE} defines a TOC as one in its codon family that is used most frequently in HEGs after adjustment of mutation bias reflected in LEGs. Comparative studies (Coghlan and Wolfe 2000; Comeron and Aguade 1998) suggest that CAI is better than ENC, CBI, and F_{op} in predicting gene expression levels, tAI is better than CAI (dos Reis et al. 2004; Tuller et al. 2010), and I_{TE} is better than CAI and tAI (Xia 2015). However, such comparison depends not only on the methods but also on the quality of the software that implements the methods. A good method could be conceptually sound but implemented erroneously and generate poor results. Moreover, the same index could be implemented differently. For example, one implementation could treat all synonymous codons into one family so that some codons could have six or even eight synonymous codons (trematode mitochondrial code has eight Ser codons: UCN and AGN), whereas another implementation would break all compound codon families, such as Leu, Ser, and Arg codon families, into separate fourfold and twofold codon families.

2.1 RSCU (Relative Synonymous Codon Usage)

RSCU measures codon usage bias for each codon within each codon family. It is essentially a normalized codon frequency so that the expectation is 1 when there is no codon usage bias. A codon is overused if its RSCU value is greater than 1 and underused if its RSCU value is less than 1. It is computed directly from input sequences.

2.1.1 Calculation of RSCU

The general equation for computing RSCU is

$$\mathrm{RSCU}_{ij} = \frac{\mathrm{CodFreq}_j}{\frac{\left(\sum_{j=1}^{\mathrm{NumCodon}_i} \mathrm{CodFreq}_i\right)}{\mathrm{NumCodon}_i}} \tag{9.2}$$

where i refers to a codon family and j to a specific codon within the family. For example, i may refer to the alanine codon family with four codons (GCU, GCC, GCA, and GCG) and j to a specific codon such as GCU. In this case, the numerator is the frequency of GCU, and the denominator is the summation of the four codon frequencies divided by the number of codons in the codon family, i.e., 4.

For biology students, it is always easier to learn by numerical examples. Suppose we counted the codon frequencies of one particular protein-coding sequence and have obtained the codon frequencies (Table 9.7). The RSCU for the GCU codon is computed, according to Eq. (9.2), as

$$\mathrm{RSCU}_{\mathrm{GCU}} = \frac{52}{\frac{(52+91+103+2)}{4}} = 0.84 \tag{9.3}$$

which is displayed in Table 9.7. Biology students are recommended to cover up the last column in Table 9.7 and finish the computation of the rest of the RSCU values.

Table 9.7 Data for illustrating the calculation of RSCU

Codon	AA	*N*	RSCU
GCU	Ala	52	0.84
GCC	Ala	91	1.47
GCA	Ala	103	1.66
GCG	Ala	2	0.03
GAA	Glu	78	1.64
GAG	Glu	17	0.36
...	...	...	...

AA amino acid, *T* codon frequency

2.1.2 Illustration of RSCU Applications

As I mentioned earlier, a variable such as RSCU is often not interesting by itself, but it becomes more interesting when you relate the variable to some other variables. Figure 9.5 shows the correlation of RSCU for *Escherichia coli* genes and that for the *E. coli* double-stranded DNA (dsDNA) phage TLS. This strong and positive correlation suggests adaptation of host tRNA pool. This adaptation the phage genes and the host genes to the same tRNA pool in *E. coli* cells and the evolution of the very similar codon usage patterns is an example of convergent evolution, i.e., phylogenetically remote organisms evolving similar features not due to coancestry, but in response to the same selection regime induced by the same environment.

What explanation would you offer if we find little correlation in RSCU between a phage and its host? There are in fact a large number of cases in which a virus and its host share little similarity in codon usage. Will such cases invalidate our convergent evolution explanation for the strong and positive correlation between phage TLS and its *E. coli* host? Science thrives in questions, and such questions immediately drive us to search for answers, and the answers enrich our explanatory conceptual framework. Ronald Fisher once said that "No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or ideally, one question at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will respond to a logical and carefully thought-out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed" (Fisher 1926).

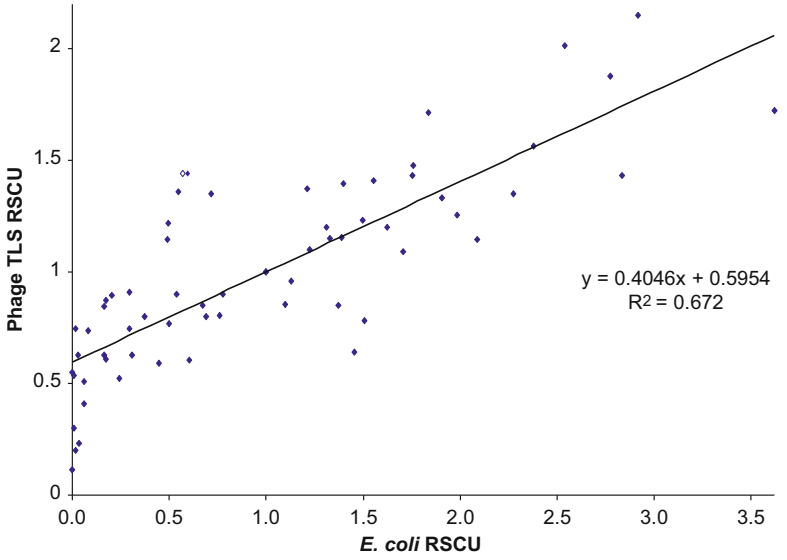

Fig. 9.5 Correlation in RSCU between *Escherichia coli* and its double-stranded DNA phage TLS

There are at least six factors that will weaken the correlation in RSCU between a virus and its host. First, some dsDNA phages carry many tRNA genes of their own genome, and the transcription of these tRNA genes would modify the host tRNA pool. For example, another dsDNA *E. coli* phage, enterobacteria phage WV8, carries 20 tRNA genes on its genome. In such cases, the phage genes would adapt to the modified tRNA pool which may be different from the tRNA pool where *E. coli* mRNAs are translated normally (i.e., without phage infection). Partly for this reason, the correlation in RSCU between enterobacteria phage WV8 and its *E. coli* host is much weaker than that shown in Fig. 9.5 (Chithambaram et al. 2014a). Phage TLS (Fig. 9.5) happens to have a genome that does not encode any tRNA genes of its own. So it depends entirely on the host tRNA pool to decode the codons of its genes.

Second, codon usage adaptation takes time. If a phage having adapted to one host has switched to a new host, and if the original host and the new host differ in their tRNA pools, then the phage codon usage will be more similar to that of the original host than the new host. This may be applicable to phage PRD1 which belongs to the peculiar *Tectiviridae* family with members parasitizing both gram-negative and gram-positive bacteria. Phage PRD1 is the only species in the family known to parasitize gram-negative bacteria, with other members of the family, i.e., phages PR3, PR4, PR5, L17, and PR772, parasitizing gram-positive bacteria (Bamford et al. 1995; Grahn et al. 2006). It is reasonably safe to assume that the phage PRD1 lineage has switched host from gram-positive to gram-negative bacteria. Furthermore, there is only one amino acid difference in the coat protein between phages PRDl and PR4 (Bamford et al. 1995). This suggests that PRD1 is phylogenetically close to its relative parasitizing gram-positive, i.e., the host-switching may have occurred quite recently. In fact, codon usage in phage PRD1 is more similar to that in gram-positive bacteria than in gram-negative bacteria (Chithambaram et al. 2014b). Among 87 bacterial genomes covering major groups of bacterial species, the host species with codon usage most similar to that of phage PRD1 are strains in the gram-positive *Geobacillus* (NC_014206, NC _012793, NC_014650, NC_014915, NC_013411).

Third, a phage with a wide range of host species may imply diverse tRNA pools that would represent fluctuating selection with different optima. Phage PRD1 mentioned above does have a variety of gram-negative bacteria as hosts, including *Salmonella*, *Pseudomonas*, *Escherichia*, *Proteus*, *Vibrio*, *Acinetobacter*, and *Serratia* species (Bamford et al. 1995; Grahn et al. 2006). However, this diverse array of hosts actually have rather similar codon usage, so host variability is not a good explanation for the lack of similarity in codon usage between PRD1 and *E. coli* (Chithambaram et al. 2014b).

Fourth, the tRNA-mediated selection differs in its effectiveness between temperate phages (i.e., those with lysogeny) and virulent phages (i.e., those without lysogeny). The lysogenic phase effectively hides protein-coding genes of the phage from tRNA-mediated selection, and the phage codon usage will be at the mercy of mutation bias in the host genome. In contrast, virulent phages have their codon usage under tRNA-mediated selection every time they enter the host cell. For this reason, one would expect better codon usage adaptation in virulent phages than in temperate phages, which is true (Prabhakaran et al. 2015).

Fifth, mass translation of phage mRNA often occurs in the late infection phase when the host cellular environment has already been dramatically altered, presumably with a quite different tRNA pool in the late phase from that in the early phase. In vaccinia virus, the degradation of host mRNA appears nearly complete 6 h after the viral infection as no host poly(A) mRNA is detectable at/after this time (Katsafanas and Moss 2007). Shutdown or drastic alteration of host protein and RNA expression implies that many tRNA species are no longer sequestered for host translation, which would dramatically alter availability of different tRNA species. Many other viruses, including hepatitis C (Chan and Egan 2009), SARS (Minakshi et al. 2009), Japanese encephalitis virus (Su et al. 2002), and coxsackie B2 virus (Zhang et al. 2010), can induce stress responses such as the UPR (unfolded protein response) in late phase. URP often results in the shutdown of transcription of ribosomal RNAs as well as repression of translation via phosphorylation of eukaryotic translation initiation factor eIF-2α (DuRose et al. 2009). All these suggest that the tRNA pool in the late phase differs from that in the normal cell. If codon usage of phage genes adapts to the altered tRNA pool in the late phase, whereas that of host genes adapts to the tRNA pool and normal cells, then we should not expect the parasite and the host share high similarity in codon usage. Interestingly, HIV-1 early genes have RSCU positively correlated with RSCU of human genes, but HIV-1 late genes have RSCU values negatively correlated with RSCU of human genes (van Weringh et al. 2011).

Sixth, if mutation bias is in different direction from tRNA-mediated selection, e.g., if tRNA-mediated selection favors Y-ending codons whereas mutation bias favors R-ending codons (where Y and R stand for pyrimidine and purine, respectively), then strong mutation bias will disrupt selection. This may well be the case for the poor codon adaptation in HIV-1. According to a recent compilation of tRNAs in human genome (Chan and Lowe 2009), the AUC codon can be translated by 17 $tRNA^{Ile}$ species (14 $tRNA^{Ile/IAU}$ and 3 $tRNA^{Ile/GAU}$) and AUU can be translated by 14 $tRNA^{Ile/IAU}$ species, whereas AUA can be translated by only 5 $tRNA^{Ile/UAU}$ species. In agreement with the tRNA-mediated selection, human genes code Ile mostly by AUC and least by AUA. In contrast, HIV-1 genes code Ile mostly by AUA and least by AUC (Haas et al. 1996; Nakamura et al. 2000). The poor codon adaptation of HIV-1 (Fig. 9.6a) reduces the translation efficiency of HIV-1 genes. Modifying HIV-1 codon usage according to host codon usage has been shown to increase the production of viral proteins (Haas et al. 1996; Ngumbela et al. 2008). The high frequency of maladaptive AUA codons in HIV-1 genes is due to high A-biased mutation at the third codon position of HIV-1 genes (Jenkins and Holmes 2003). The A-bias is mediated by the error-prone reverse transcriptase (Martinez et al. 1994; Vartanian et al. 2002) and the human APOBEC3 protein (Yu et al. 2004). The frequency of A can reach up to 40% in some HIV-1 genomes (Vartanian et al. 2002), resulting in a preponderance of A-ending codons which are typically rarely used in the human HEGs (Kypr and Mrazek 1987; Sharp 1986).

One would predict a better correlation in RSCU between HIV-1 genes and highly expressed human genes. One viral species that may shed light on this prediction is HTLV-1 which infects the same type of host cell as HIV-1. Both HIV-1 and HTLV-1 are retroviruses with RNA genomes, but HTLV-1 is exceptional in that it does not

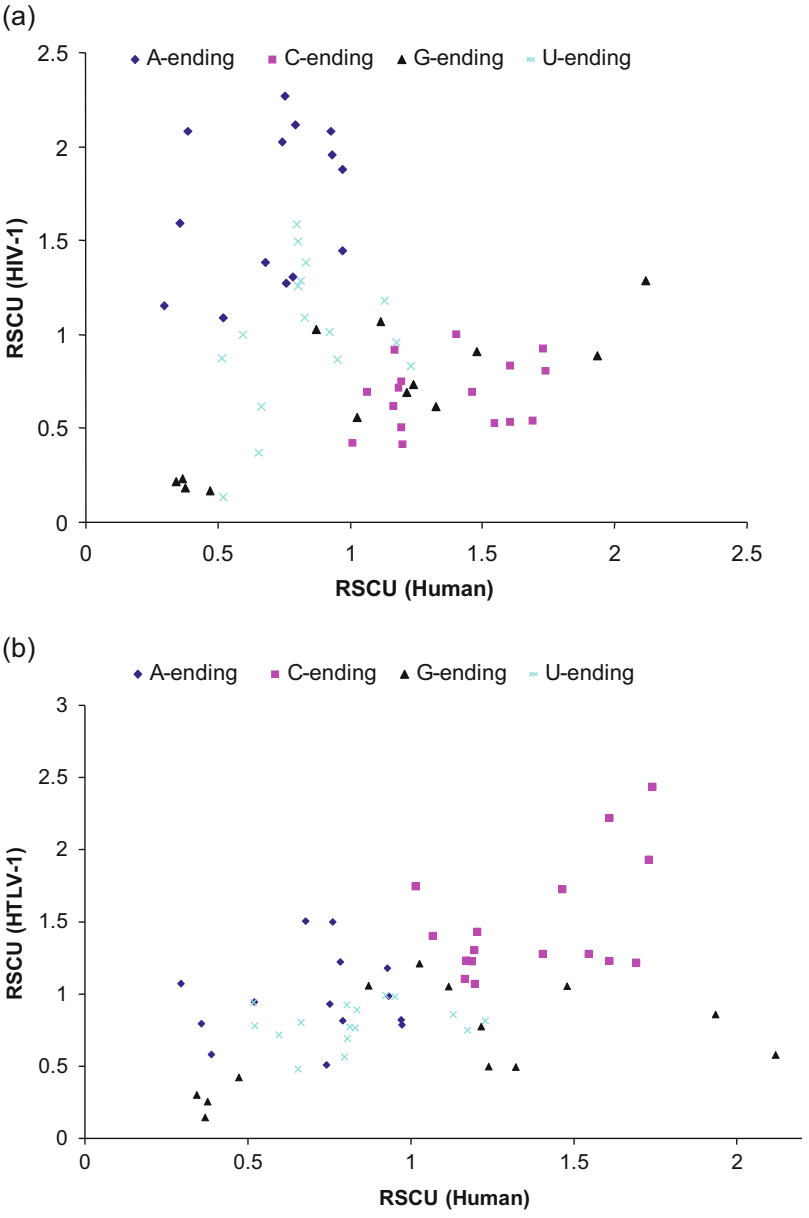

Fig. 9.6 Relative synonymous codon usage (RSCU) of HIV-1 (**a**) and HTLV-1 (**b**) plotted against RSCU of highly expressed human genes. Modified from van Weringh et al. (2011)

have a strong A-biased mutation (Van Dooren et al. 2004; van Hemert and Berkhout 1995). HTLV-1 relies for the most part on the host polymerase to replicate through clonal expansion of infected cells rather than undergoing iterative replication cycles

like HIV-1 (Strebel 2005). The substitution rate of HTLV-1 is consequently lower, about 5.2×10^{-6} substitutions/site/year (Hanada et al. 2004; Van Dooren et al. 2004), whereas that of HIV-1 is around 2.5×10^{-3} substitutions/site/year (Hanada et al. 2004). Thus, although HTLV-1 infects the same cells as HIV-1, i.e., human CD4+ T cells (Rimsky et al. 1988), and both viruses are therefore subject to the same selective pressures on codon usage by the host tRNA pool, mutations are less likely to disrupt codon-anticodon adaptation in HTLV-1 than in HIV-1 as they occur at a lower rate in the former. The positive correlation in RSCU between HTLV-1 and highly expressed human genes (Fig. 9.6b) is highly significant (Pearson $r = 0.4982$, $p < 0.0001$, Spearman $r = 0.4688$, $p = 0.0002$).

2.2 CAI (Codon Adaptation Index)

CAI has been used extensively in biological research. Other than its primary use for measuring the efficiency of translation elongation, it has contributed to the finding that functionally related genes are conserved in their expression across different microbial species (Lithwick and Margalit 2005), to the prediction of protein production (Futcher et al. 1999; Gygi et al. 1999), and to the optimization of DNA vaccines (Ruiz et al. 2006).

2.2.1 Calculation of CAI

While RSCU characterizes codon usage bias in each codon family, CAI quantifies the codon usage bias in one gene. It is based on (1) the codon frequencies of the gene and (2) the codon frequencies of a set of known HEGs (often referred to as the reference set). The reference set of genes is used to generate a column of w values computed as

$$w_{ij} = \frac{\text{RefCodFreq}_{ij}}{\text{RefCodFreq}_{i.\max}} \tag{9.4}$$

where RefCodFreq_{ij} is the frequency of codon j in synonymous codon family i and $\text{RefCodFreq}_{i.\max}$ is the maximum codon frequency in synonymous codon family i. For example, if the four alanine codons GCA, GCC, GCG, and GCU have frequencies 20, 4, 4, and 2, respectively, then their associated w value are 1, 0.2, 0.2, and 0.1, respectively. The codon whose frequency is $\text{RefCodFreq}_{i.\max}$ is often referred to as the major codon (whose w is 1), and the other codons in the synonymous codon family are referred to as minor codons. The major codon is assumed to be the translationally optimal codon.

It is easy to see the relationship between w_{ij} and RSCU. The former is obtained by dividing each RSCU by the largest RSCU value within each codon family. With the

w values for a particular species, we can now compute the CAI value of any protein-coding sequence from the species by using the following equation:

$$\mathrm{CAI} = e^{\left(\frac{\sum_{i=1}^{n} \left[\mathrm{CodFreq}_i \ln\left(w_i\right) \right]}{\sum_{i=1}^{n} \mathrm{CodFreq}_i} \right)} \tag{9.5}$$

where n is the number of sense codons (excluding codon families with a single codon, e.g., AUG for methionine and UGG for tryptophan in the standard genetic code). Note that the exponent is simply a weighted average of $\ln(w)$. Because the maximum of w is 1, $\ln(w)$ will never be greater than 0. Consequently, the exponent will never be greater than 0. Thus, the maximum CAI value is 1. The minimum CAI depends on the w values for minor codons in each codon family. If the minor codons all have w values close to zero, then the minimum CAI will also be very close to zero.

The calculation of CAI is numerically illustrated in Table 9.8 for a gene whose observed codon frequency is in column ObsFreq (Table 9.8). The codon frequency of the highly expressed reference set is in column "RefCodFreq." The column "w" is obtained by dividing RefCodFreq values by the largest value in the codon family. For example, the first w value in the table, 0.606, is obtained by dividing RefCodFreq value 195 by the largest RefCodFreq value in the alanine codon family, i.e., 322. We take a weight average of $\ln(w)$ as shown in Eq. (9.5) and then exponentiate it to obtain CAI.

The way w is calculated implies that, if a protein contains only methionine and tryptophan, both encoded by a single codon (AUG and UGG, respectively, in

Table 9.8 Illustration of CAI calculation for a gene whose observed codon frequencies are in column "ObsFreq"

Codon	AA	ObsFreq	RefCodFreq	w
GCA	A	1	195	0.606
GCU	A	15	322	1.000
GCG	A	0	81	0.252
GCC	A	8	242	0.752
UGC	C	3	123	1.000
UGU	C	3	112	0.911
GAU	D	9	69	1.000
GAC	D	11	40	0.580
GAG	E	11	289	0.863
GAA	E	14	335	1.000
UUU	F	3	118	0.554
UUC	F	9	213	1.000
...	...	...		...

The codon frequency of the highly expressed reference set is in column "RefCodFreq." The column "w" is obtained by dividing RefCodFreq values by the largest value in the codon family

standard code), then the gene will have the highest CAI value of 1 because *w* values are 1 for such codons. Similarly, a gene with many AUG and UGG codons would have high CAI values even if it is not under any tRNA-mediated selection. For this reason, a good implementation of CAI should exclude single-member codon families from CAI calculation.

I have previously mentioned that codon usage indices such as CAI can be implemented differently with different classification of codon families, so gene A could have a higher CAI value than gene B from one software, but the opposite from another software. I wish to illustrate this so that the reader can better interpret their results.

In highly expressed yeast genes (e.g., compiled in the Eyeastcai.cut in EMBOSS distribution), CGU is by far the most frequent codon in the CGN (coding for arginine) codon family. The overuse of CGT and the avoidance of CGG, CGA, and CGC codons in highly expressed yeast genes make sense because the yeast genome contains six $tRNA^{Arg}$ genes with anticodon ACG forming Watson-Crick base pairing with the CGT codon, but no other $tRNA^{Arg}$ gene forming Watson-Crick base pairing with the other three CGN codons (the nucleotide A in anticodon ACG is modified to inosine but still pairs with U better than with other nucleotides). While this illustrates well the codon-anticodon adaptation, it causes practical problems with computing CAI.

Suppose we now use a sequence consisting entirely of CGU codons and expect the resulting CAI to be 1 by using the Eyeastcai.cut reference set. The resulting CAI value from the EMBOSS.cai program is 0.140 instead of 1. It turns out that amino acid arginine is coded by two codon subfamilies, the CGN codon family we have mentioned and the AGR codon family. The largest codon frequency among these six codons is 314 (for AGA codon) in Eyeastcai.cut. So the *w* value for CGT is not 1 (43/43) as we have thought but is only 0.1369 (= 43/314). For this reason, some CAI-calculating programs, e.g., DAMBE (Xia 2013, 2017d), may separate compound codon families such as the arginine family into two separate families, one twofold and one fourfold.

2.2.2 Illustration of CAI Applications

The most obvious application of CAI or related codon usage indices is to optimize codon usage to optimize protein expression. Many experiments have demonstrated increased protein production by optimizing codon usage and decreased protein production if codons are replaced by rarely used ones (Haas et al. 1996; Kaishima et al. 2016; Ngumbela et al. 2008; Robinson et al. 1984; Sorensen et al. 1989). There are claims that codon optimization does increase protein production (e.g., Kudla et al. 2009), but these claims were found to be due to wrong data analysis (Tuller et al. 2010; Xia 2015) and will be dealt with on a later section on I_{TE} (Xia 2015). Below I list two less obvious applications of CAI.

2.2.2.1 Does High Mutation Rate Prevent HIV-1 Genes from Evolving Codon Adaptation?

I have mentioned in the section on RSCU that the lack of concordance in codon usage between HIV-1 and human genes was conventionally explained by high mutation rate in HIV-1, based on the observation that (1) HIV-1 genome is known to experience strongly A-biased mutations, (2) usage of A-ending codons in HIV-1 genes is particularly different from that of the host genes, and (3) HTLV-1 that parasitizes the same human CD4+ T cells but has reduced mutation rate does have codon usage similar to human genes (Fig. 9.6b). Thus, the lack of concordance in codon usage between HIV-1 and human genes is interpreted as poor codon adaptation caused by high mutation rate disrupting codon adaptation.

However, van Weringh et al. (2011) objected to this interpretation. They argued that the lack of concordance in codon usage between HIV-1 and human genes is not due to poor codon adaptation in the part of HIV-1 genes, but because HIV-1 genes, especially the late genes, have adapted to a tRNA pool that is fundamentally different from that in a normal human CD4+ T cell. What originally prompted them to formulate this hypothesis is the observation that CAI for HIV-1 early genes are significantly greater than CAI for HIV-1 late genes when highly expressed human genes are used as reference genes. These late genes encode mass-translated HIV-1 structural proteins and are typically expected to have higher CAI than the relatively lowly expressed early genes. So it is thus a surprise to see late genes having smaller CAI than early genes, unless the mass-translated late genes adapt to a tRNA pool different from the early genes.

van Weringh et al. (2011) investigated experimentally measured tRNA abundance in the human cell when the late HIV-1 genes are translated and HIV-1 virions are produced. The tRNA pool for the late genes is indeed different in the expected direction, supporting their hypothesis that the lack of concordance in codon usage between HIV-1 and human genes is not due to poor codon adaptation in HIV-1 genes but because HIV-1 genes, especially the late genes, have adapted to a tRNA pool different from the one with which highly expressed human genes are translated (van Weringh et al. 2011).

2.2.2.2 Detecting Horizontally Transferred Genes

CAI has also been used jointly with a reformulated effective number of codons (N_c, Sun et al. 2013) to detect horizontally transferred genes. *E. coli* genes with a strong codon usage bias typically have high CAI values. However, three genes (*yagF*, *yagG*, and *yagH*) from the defective CP 4–6 prophages of *E. coli* (Wang et al. 2010) have strongly biased codon usage (small N_c values) but relatively small CAI values. This codon usage pattern sets the three genes apart from the rest of *E. coli* genes (Fig. 9.7) which highlight the value of using the "N_c versus CAI" plot to detect

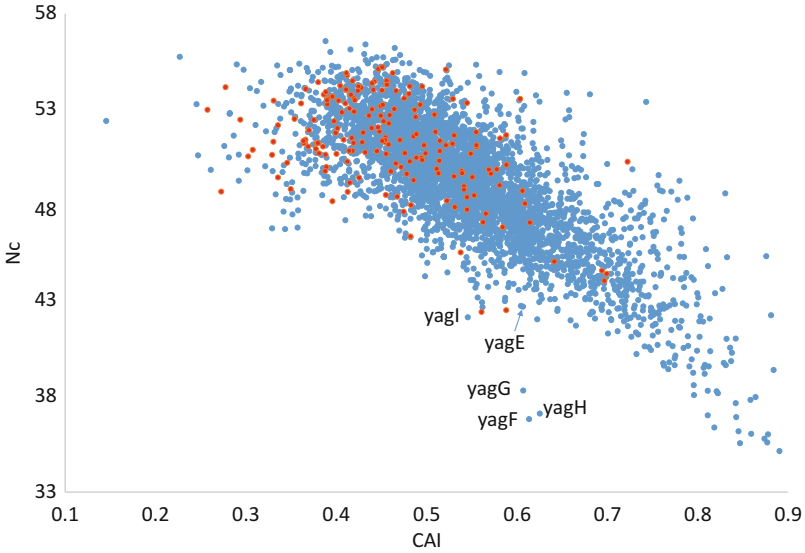

Fig. 9.7 Plot of CAI against a reformulated effective number of codons (N_c, Sun et al. 2013) for *E. coli* genes facilitates the detection of newly "immigrant" genes that exhibit codon usage bias different from the "native" genes. Three *E. coli* genes (*yagF*, *yagG*, and *yagH*) from the defective CP 4–6 prophages of *E. coli* (Wang et al. 2010) have strongly biased codon usage (relatively small N_c) but relatively poor codon adaptation (mediocre CAI values). The red points represent 179 annotated *E. coli* pseudogenes (NC_000913) that have not accumulated frameshifting mutations

recently horizontally transferred genes. These genes have been "naturalized" in *E. coli* genome and contribute to *E. coli* survival and growth (Wang et al. 2010).

The largest mucin gene (*mucin 14A*) in *Drosophila melanogaster* also exhibits strong codon usage bias ($N_c = 38.6$), but in the direction opposite to those highly expressed *D. melanogaster* genes. Its CAI value is equal to 0.1277, which is the second smallest among all *D. melanogaster* genes. It is unknown how and why the gene has evolved to have such a peculiar feature.

The distribution of CAI values for the 179 annotated pseudogenes are indicated in red. These pseudogenes have not accumulated frameshifting mutations and presumably were pseudogenized only recently. They tend to be clustered on the lower end of CAI distribution, suggesting that genes with high CAI values require tRNA-mediated selection to maintain the high CAI values.

The gene with the smallest CAI is *mgtL*, which has only 17 sense codons and is a bacterial mRNA leader that controls the expression of the downstream *mgtA* (Park et al. 2010). The low CAI is not due to stochastic fluctuation due to small number of codons but because almost all used codons are minor codons. This may represent a real case of a gene preferring minor codons to facilitate its regulatory function.

2.2.3 Problems with CAI and Other Gene-Specific Codon Usage Indices

There are major problems with CAI and other commonly used codon usage indices. While some minor problems have been addressed before (Xia 2007c), the key issue of properly inferring translationally optimal codons (TOCs) remains unresolved. These gene-specific codon usage indices all need to infer TOCs, by using two types of information. The first, represented by tAI (dos Reis et al. 2004), uses the most abundant tRNA and its anticodon to infer TOC within each codon family, i.e., the codon that base-pairs best with the most abundant tRNA is the TOC. The second, represented by CAI, considers the most frequent codon in HEGs as the TOC within each codon family. I will outline the problems to pave the way for the presentation of a new index of translation elongation in the next section (I_{TE}, Xia 2015).

2.2.3.1 Problem with Codon Usage Indices Using tRNA Abundance to Infer TOCs

For indices such as tAI that use tRNA abundance information to define TOCs, the main problem is that TOCs cannot be inferred reliably from tRNA gene copy numbers or experimentally measured tRNA abundance. For example, inosine is expected to pair best with C and U, less with A (partly because of the bulky I/A pairing involving two purines), and not with G. However, $tRNA^{Val/IAC}$ from rabbit liver pairs better with GUG codon than with other synonymous codons (Jank et al. 1977; Mitra et al. 1977). No one would have identified GUG as the best codon for $tRNA^{Val/IAC}$ without actually seeing the experimental result.

Similarly, the *Bacillus subtilis* genome codes $tRNA^{Ala/GGC}$ for decoding GCY codons. One would have thought that GCC codon, which forms Watson-Crick base pairing with the anticodon, would be translationally more optimal than GCU. However, GCU is used much more frequently than GCC in HEGs than LEGs in *B. subtilis*. We have encountered a similar example in Table 9.4 involving Cys codon usage in HEGs. There are four $tRNA^{Cys}$ genes with the same anticodon GCA forming Watson-Crick base pair with UGC codon, but no $tRNA^{Cys}$ gene with anticodon forming Watson-Crick base pair with the alternative UGU codon. We would have taken UGC as the TOC. However, UGU is used far more frequently than UGC codon in highly expressed yeast genes relative to LEGs. In short, in all these cases we would be wrong to use the most abundant tRNA species and its matching codon to infer TOC.

There is one more reason for tRNA abundance not able to reliably predict TOCs. What matters in translation elongation is not the abundance of transcribed tRNAs but the availability of charged tRNAs. It is tedious to determine the level of charged tRNAs, and researchers typically would use transcriptionally determined tRNAs or even the number of tRNA genes in the genome as a proxy of charged tRNAs. Unfortunately, the abundance of tRNAs often do not reflect the abundance of charged tRNA (Elf et al. 2003).

Furthermore, codon-anticodon base pairing is known to be context-dependent (Lustig et al. 1989). For example, a wobble cmo^5U in the anticodon of $tRNA^{Pro}$, $tRNA^{Ala}$, and $tRNA^{Val}$ can read all four synonymous codons in the respective codon family, but the same cmo^5U in $tRNA^{Thr}$ cannot read C-ending codons (Nasvall et al. 2007). For this reason, the optimal codon usage is likely better approximated by the codon usage of HEGs than what we can infer based on codon-anticodon pairing. Consistent with this proposition, CAI, which is based on the codon usage of HEGs (HEGs), performs better in predicting protein production or abundance than other indices based on tRNAs (Coghlan and Wolfe 2000; Comeron and Aguade 1998; Duret and Mouchiroud 1999).

2.2.3.2 Problem with Using Codon Usage of HEGs to Infer TOCs

Codon usage indices such as CAI that use codon usage of HEGs to infer TOCs also have problems. Other than those previously outlined (Xia 2007c), it often leads to wrong interpretation of tRNA-mediated selection. I illustrate this problem here with the Ala codon subfamily GCR (where R stands for either A or G). The frequencies of GCA and GCG in *E. coli* HEGs, as compiled and distributed with EMBOSS (Rice et al. 2000), are 1973 and 2654, respectively, which may lead one to think that *E. coli* translation machinery prefers GCG over GCA. However, the codon frequencies of GCA and GCG for *E. coli* non-HEGs are 25,511 and 43,261, respectively. Thus, GCA is relatively more frequent in *E. coli* HEGs than in *E. coli* non-HEGs. This suggests that mutation bias favors GCG, but tRNA-mediated selection favors GCA. The battle between the mutation bias and tRNA-mediated selection leads to increased usage of GCA in *E. coli* HEGs relative to LEGs, although GCA is still not as frequent as GCG in HEGs. This interpretation is corroborated by the *E. coli* genome encoding three $tRNA^{Arg}$ genes for GCR codons, all with a UGC anticodon forming perfect Watson-Crick base pair with codon GCA.

The example above illustrates the point that mutation bias is reflected to codon usage of lowly expressed genes. This is what has driven the formulation, development, and implementation of a new codon usage index, I_{TE} (Xia 2015).

2.3 $\mathbf{I}_{TE}$ *(Index of Translation Elongation)*

2.3.1 Illustration of I_{TE} Calculation

I_{TE} is implemented in DAMBE (Xia 2013, 2017d). There are in fact four different implementations of I_{TE} in DAMBE, depending on how one would classify codons into codon families. The first implementation is the most extreme (unconventional) and classifies all sense codons into NNR or NNY codon families or subfamilies. For example, the fourfold alanine codon is broken into GCR and GCY subfamilies. For such an NNR or NNY codon family or subfamily i, we first define $P_{i.\text{HEG}}$ and $P_{i.\text{non-}}$

HEG as the proportion of codon i within its R-ending or Y-ending family for *E. coli* HEGs and non-HEGs. Take data for codons GCA and GCG in Table 9.9, for example:

$$\begin{aligned} P_{\text{GCA.HEG}} &= \frac{N_{\text{GCA.HEG}}}{N_{\text{GCR.HEG}}} = \frac{1973}{1973+2654} = 0.42641 \\ P_{\text{GCA.non-HEG}} &= \frac{N_{\text{GCA.non-HEG}}}{N_{\text{GCR.non-HEG}}} = \frac{25511}{25511+43261} = 0.37095 \end{aligned} \tag{9.6}$$

$$\begin{aligned} S_{\text{GCA}} &= \frac{P_{\text{GCA.HEG}}}{P_{\text{GCA.non-HEG}}} = 1.1495 \\ S_{\text{GCG}} &= \frac{P_{\text{GCG.HEG}}}{P_{\text{GCG.non-HEG}}} = 0.9118 \end{aligned} \tag{9.7}$$

where S_{GCA} and S_{GCG} may be viewed as relative codon frequencies of HEGs corrected for the "background" non-HEGs. Codon i is considered selected for if $S_i > 1$ and against if $S_i < 1$. Thus, codon GCA is considered selected for because, according to Eq. (9.7), $S_{\text{GCA}} > 0$. This insight would be obscured if we use codon frequency data from *E. coli* HEGs only which would have suggested that codon GCA is selected against. The S_i values for the four sense codons in *E. coli* are listed in Table 9.9.

We now compute w_i as follows:

$$\begin{aligned} w_i &= \frac{S_i}{\text{Max}(S_i)}, \text{e.g.,} \\ w_{\text{GCA}} &= \frac{1.1495}{1.1495} = 1; w_{\text{GCG}} = \frac{0.9118}{1.1495} = 0.7932 \end{aligned} \tag{9.8}$$

The index of translation elongation (I_{TE}) is then calculated in the same way as CAI except that, in this particular codon family classification, the computation is applied to NNR and NNY codon subfamilies:

$$I_{\text{TE}} = e^{\frac{\sum_{i=1}^{N_s} F_i \ln w_i}{\sum_{i=1}^{N_s} F_i}} \tag{9.9}$$

Table 9.9 Codon frequency (CF) for *E. coli* highly expressed genes (HEGs) and non-HEGs, as well as the computed S_i values according to Eq. (9.7)

AA	Codon	CF_{HEG}	$\text{CF}_{\text{non-HEG}}$	S_i
A	GCA	1973	25,511	1.1495
A	GCG	2654	43,261	0.9118
A	GCC	1306	33,463	0.5646
A	GCU	2288	18,526	1.7865
...	...	...	...	...

where F_i is the frequency of codon i and N_s is the number of sense codons (excluding those in single-codon families). For example, AUG for methionine, AUA for isoleucine, and UGG for tryptophan in the standard genetic code are excluded from computing I_{TE}. Just like CAI, tAI, and N_c, I_{TE} is a gene-specific index of codon usage bias.

One may note that CAI is a special case of I_{TE} when there is absolutely no codon usage bias in non-HEGs in all codon subfamilies. That is, when $N_{GCA.non\text{-}HEG} = N_{GCG.non\text{-}HEG}$, $N_{GCC.non\text{-}HEG} = N_{GCU.non\text{-}HEG}$, and so on. The range of I_{TE} is the same as CAI, i.e., between 0 and 1.

Readers may demand a justification for the extreme classification of all sense codons into NNR and NNY codon families. The main reason is that, for genes encoded by the nuclear genome, the R-ending codons are typically decoded by two types of tRNA species (one with a wobble C and the other with a wobble U), whereas the Y-ending codons are decoded typically by a single type of tRNA species with either a wobble G or a wobble A modified to inosine, but never by both (Grosjean et al. 2007; Marck and Grosjean 2002). For this reason, the R-ending and Y-ending codons, even within a single fourfold codon family, are subject to different tRNA-mediated selection and therefore should be treated separately. Such implementation is also relevant for certain experimental settings that induce mutation almost exclusively in NNY codons, which is the case in Kudla et al. (2009). However, for comparative purposes, I have included two alternative I_{TE} implementations in DAMBE (Xia 2013, 2017d): (1) with compound sixfold and eightfold codon families broken into twofold and fourfold codon families and (2) lumping all synonymous codons into one codon family. One may access the function by clicking "Seq.Analysis|Codon usage|Index of translation elongation" and then choosing the desired implementation.

2.3.2 A Major Controversy Resolved by the Application of I_{TE}

Highly expressed genes in bacteria and unicellular eukaryotes overuse codons that match the anticodon of the most abundant tRNA (Ikemura 1981a, b, 1982, 1992). When such codons are replaced by rarely used codons, protein production is reduced (Robinson et al. 1984; Sorensen et al. 1989). Similarly, when codon usage is optimized, protein production is increased (Haas et al. 1996; Kaishima et al. 2016; Ngumbela et al. 2008). However, to what degree is translation elongation rate-limiting has been controversial. Early theoretical considerations (Andersson and Kurland 1983; Bulmer 1990, 1991; Liljenstrom and von Heijne 1987) tend to favor the argument that translation elongation is not rate-limiting in protein production, but translation initiation is. This hypothesis states that codon-anticodon adaptation and increased elongation efficiency are not related to protein production. Instead, the benefit of codon adaptation and increased elongation efficiency is to increase ribosomal availability for global translation and timely response to environmental perturbations.

To test these two alternative hypotheses, Kudla et al. (2009) engineered a synthetic library of 154 genes, all encoding the same green fluorescent protein in *Escherichia coli*, but differing in synonymous sites (and consequently the degree of codon adaptation, as measured by codon adaptation index or CAI). All sequences share an identical 5′ UTR of 144 nt long, so there is no variation in the Shine-Dalgarno sequence. Because the engineered genes all encode the same protein, it is justifiable to use protein abundance as a proxy for protein production (assuming that protein molecules sharing the same amino acid sequence have the same degradation rate).

Kudla et al. (2009) used minimum folding energy (MFE), computed from sites −4 to +37 (where ribosomes position themselves at the initiation codon), as a proxy for initiation efficiency. The rationale for using MFE as a measure of translation initiation is that an initiation codon would be inaccessible if it is embedded in a strong secondary structure and that accessibility of the initiation codon is a key determinant of translation initiation efficiency (Nakamoto 2006). Stable secondary structure in sequences positioned at or before the start codon has been experimentally shown to inhibit translation initiation (Osterman et al. 2013), presumably because it embeds SD and start codon in a structural stem and consequently hiding the SD and start codon signals from ribosomes. The previous chapter on translation initiation has already highlighted the point that mRNAs in bacteria and unicellular eukaryotes tend to have much weaker secondary structure near the start codon than elsewhere, especially those from highly expressed.

Kudla et al. interpreted CAI as a proxy of translation elongation. If both translation initiation and elongation contribute to translation efficiency, then protein production is expected to depend on both MFE and CAI. If only translation initiation is important, then protein production will depend on MFE only. They found that MFE accounts for 44% of the variation in protein production but CAI is essentially unrelated to protein production. They concluded consequently that "translation initiation, not elongation, is rate-limiting for gene expression."

The conclusion by Kudla et al. (2009), however, is based on two critical assumptions. First, MFE and CAI are good proxies of translation initiation and elongation efficiencies, respectively. Second, the effect of translation elongation is independent on translation initiation. The problem with the second assumption has been pointed out recently (Supek and Smuc 2010; Tuller et al. 2010) who reanalyzed the data in addition to providing an overwhelming amount of additional empirical evidence to demonstrate the joint effect of both translation initiation and elongation on protein production. In short, protein production rate is expected to increase with elongation efficiency only when translation initiation is efficient. If translation initiation is slow, then increasing elongation rate is not expected to increase protein production. Kudla et al. (2009) ignored the dependence of elongation effect on translation initiation.

Xia (2015) reanalyzed the experimental data in Kudla et al. (2009) with two improvements, by replacing CAI by I_{TE} and by incorporating translation initiation and elongation into one model. Three points are worth highlighting in Fig. 9.8a. First, in contrast to a nonsignificant relationship between protein abundance and CAI, the protein abundance and I_{TE} are highly significantly correlated ($p = 0.0001$,

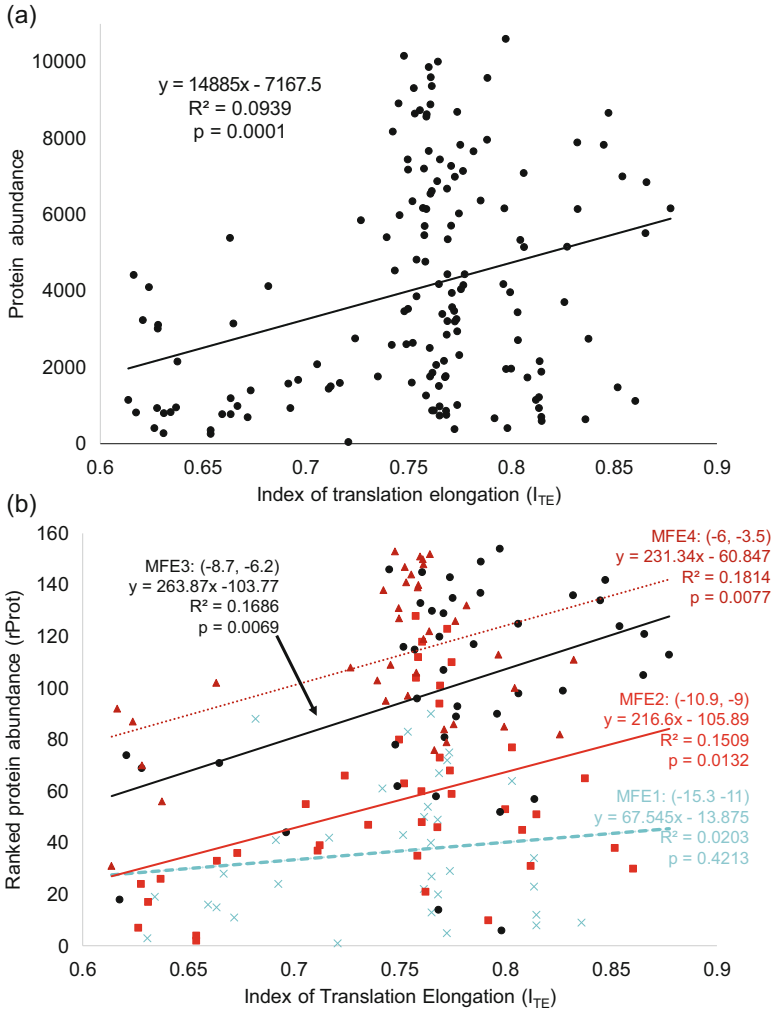

Fig. 9.8 Relationship between protein abundance (measured by GFP normalized fluorescence; data kindly provided by Dr. Plotkin) translation elongation efficiency (I_{TE}). (**a**) Without considering translation initiation. (**b**) The relationship between protein abundance and I_{TE} is characterized separately for four groups of data, with MFE1, MFE2, MFE3, and MFE4 corresponding to groups of genes with increasing translation initiation efficiency. (Modified from Xia 2015)

Fig. 9.8a). Second, when I_{TE} is small (e.g., $I_{TE} < 0$), protein abundance is generally low, suggesting that translation elongation is limiting. Third, a large I_{TE} (efficient translation elongation) does not imply high protein production, e.g., when translation initiation is very slow. One expects a large I_{TE} to be associated with increase protein production only when translation initiation is efficient.

Xia (2015) binned MFE into four MFE categories, from strong secondary structure to weak secondary structure (−15.3, −11), (−10.9, −9), (−8.7, −6.2), and (−6, −3.5), representing translation initiation from the lowest to the highest, and designated as MFE1-MFE4 (Fig. 9.8b). The intervals are chosen in such a way that all MFE values fall into four roughly equal-sized groups with within-group MFE being as small as possible. The benefit of binning is that one can exclude the MFE variable so that the effect of I_{TE} can be modeled more explicitly. It is for the same reason that Tuller et al. (2010) also used binned analysis for this data set.

In the MFE1 group, translation initiation is the lowest, and we should expect little increase of protein production with translation elongation efficiency (I_{TE}). This is consistent with the empirical result (Fig. 9.8b) where the relationship between I_{TE} and protein abundance is not statistically significant in the MFE1 group ($b = 67.545$, $p = 0.4213$, Fig. 9.8b), with I_{TE} accounting for only 2% of total variation in ranked protein abundance (rProt). In contrast, when translation initiation is more efficient in groups MFE2-MFE4, rProt increases significantly with I_{TE}, with the simple linear model consistently accounts for about 17% of the total variation in rProt (Fig. 9.8b, with b varying from 216.60 to 263.87). Thus, the contribution of translation elongation (I_{TE}) to protein production is much greater than previously documented for this data set, i.e., absent (Kudla et al. 2009) or less than 3% of the total variation in protein production (Tuller et al. 2010). Readers may consult Xia (2015) for more explicit modeling of the protein abundance on translation initiation and elongation.

One might wonder why previous studies, although not taking translation initiation into consideration, almost always consistently show positive relationship between translation efficiency and codon adaptation. There are two explanations. First, previous experimental studies were carried out typically on highly expressed genes with efficient translation initiation efficiency. Such studies are equivalent to excluding the MFE1 group in Fig. 9.8b. Second, for correlational studies, nature generally does not generate bacterial genes with high translation initiation efficiency but poor codon adaptation or low translation initiation with high codon adaptation. However, the experiment by Kudla et al. (2009) generated both of these unnatural associations, leading to a lack of positive association between protein production and codon adaptation. This example highlights the point that a well-intended and well-done experiment can mislead us. It represents another illustration of Simpson's Paradox in which wrong conclusion is reached when one omits a contributing variable.

3 Translation Elongation Efficiency and Accuracy

Given a fixed translation initiation efficiency, our conceptual model for the relationship between codon adaptation (CA) and tRNA-mediated selection, in its simplest form, is

$$\mathrm{CA} = \alpha + \beta \mathrm{S_E} \tag{9.10}$$

where CA is tRNA-mediated codon adaptation often measured by CAI or I_{TE} (Xia 2015) and S_E is selection for translation efficiency (in unit of protein produced per mRNA molecule). The slope b is typically positive, i.e., stronger selection for translation efficiency leads to better codon adaptation. Many studies have demonstrated a strong relationship between codon adaptation and gene expression (Coghlan and Wolfe 2000; Duret and Mouchiroud 1999; Gouy and Gautier 1982).

One key deficiency in Eq. (9.10) is that it does not distinguish between selection due to translation efficiency or that due to translation accuracy (Akashi 1994). Take Asn codons AAC and AAU in *E. coli*, for example. AAC is a major codon (heavily used by highly expressed genes and decoded by the most abundant isoacceptor tRNA), whereas AAU is a rarely used minor codon. A major codon is typically translated faster than a minor codon, and highly expressed *E. coli* genes use AAC almost exclusively to code for Asn, so one could argue that the overuse of AAC is driven by S_E. However, AAC and AAU also differ in misreading rate, in particular by $tRNA^{Lys}$ which ideally should decode only AAA and AAG codons but does misread AAC and AAU, leading to Asn replaced by Lys. This misreading error rate is six times greater for AAU than for AAC, with the error ratio maintained in both Asn-starved and Asn-non-starved conditions (Johnston et al. 1984) or with streptomycin used to inhibit translation (Johnston and Parker 1985). Thus, the overuse of AAC could be driven either by selection for increased translation efficiency or increased translation accuracy or both. Designating S_A as selection for translation accuracy, we have three alternative hypotheses expressed, in the simplest form, as

$$\mathrm{CA} = \alpha + \beta_1 S_E \tag{9.11}$$

$$\mathrm{CA} = \alpha + \beta_1 S_A \tag{9.12}$$

$$\mathrm{CA} = \alpha + \beta_1 S_E + \beta_2 S_A + \beta_3 S_E S_A \tag{9.13}$$

Akashi (1994) classified amino acid sites into conserved sites (assumed to be functionally important with high S_A) and variable sites (assumed to experience low S_A). He reasoned that, if codon adaptation is due to selection for translation efficiency, then all codons in the gene should be subject to similar selection regardless of whether the codon is in a functionally important or unimportant site. In contrast, if codon adaptation is driven by selection for translation accuracy, then the selection is stronger in functionally important sites than in functionally unimportant sites. So we should observe greater codon usage bias in functionally important codon sites than functionally unimportant codon sites. He found greater codon adaptation in conserved amino acid sites than in variable amino acid sites and concluded that this difference between the conserved and variable sites to have resulted from selection for accuracy.

There is a problem with the conclusion. Take lysine codons (AAA and AAG) and glutamate codons (GAA and GAG), for example. Suppose that AAA codon is favored by selection in lysine codon family and GAG favored in glutamate codon family. Also suppose that an ancestral gene has good codon adaptation with lysine coded by AAA and glutamate coded by GAG. Now some lysine sites experienced

nonsynonymous substitutions from AAA to GAA. These sites are now designated as variable sites and are occupied by a minor codon GAA. This would result in an association between "poor codon adaptation" and variable sites that have little to do with translation accuracy. Akashi (1994) was aware of this problem but did not provide a definitive solution.

4 Amino Acid Usage and Translation Elongation Efficiency

There are at least four factors contributing to amino acid usage. The first two are related to selection for translation elongation efficiency, the third related to number of synonymous codons, and the fourth related to genomic mutation bias.

4.1 Factors Related to Selection for Translation Elongation Efficiency

Some amino acids are abundant and energetically cheap to make, i.e., consuming few ATPs in their production, whereas others are rare and energetically expensive, so mass-produced proteins should maximize the use of abundant and cheap amino acids (Akashi and Gojobori 2002). However, such a hypothesis, without considering other factors, often does not produce easily testable predictions. For example, we expect highly expressed proteins to maximize the use of energetically cheap amino acids and avoid the use of the expensive ones. However, many ribosome proteins are highly expressed, yet the need for many of them to bind to the negatively charged mRNA demands the usage of positively charged amino acids such as Lys and Arg that are typically energetically expensive to make in the cell. This would lead to an association between high expression and energetically expensive amino acid, thus confounding the prediction that highly expressed genes should maximize the use of cheap amino acids. Furthermore, amino acid availability changes with environment, and the same amino acid may be manufactured differently with different energy consumption in different organisms. So it is not easy to measure energetic cost of amino acids in different organisms. One could, however, turn the question around and ask how one can characterize energetic costs of amino acids by bioinformatic means. For example, in the ideal situation when all other factors affecting amino acid usage have been controlled for, we may infer that the avoided amino acid is perhaps rare or energenetically expensive to make. This type of inference is of course not very satisfactory and is often derogatively termed the backdoor smuggling approach because one does not present direct evidence for energetic cost.

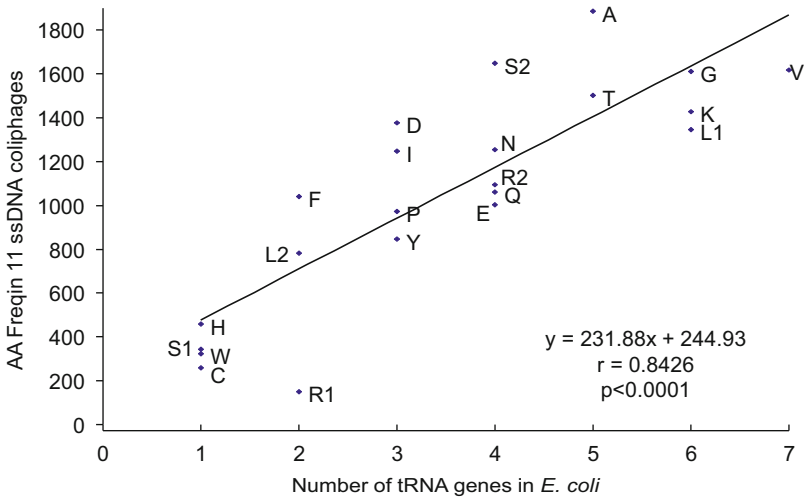

Fig. 9.9 Amino acid usage in single-stranded DNA phages infecting *E. coli* increases with the abundance of isoaccepting tRNA

The other factor related to translation elongation is the tRNA abundance, and one expects mass-produced proteins to use amino acids with many tRNAs to carry them. Designating the proportion of tRNAs carrying amino acid *i* as P_i, and the frequency of amino acid *i* in highly expressed genes as N_i, Xia (1998a) analytically derived an equation with P_i linearly increasing with the square root of N_i. The relationship was well substantiated with data from *E. coli*, *Salmonella typhimurium*, and *Saccharomyces cerevisiae* (Xia 1998a).

Single-stranded DNA (ssDNA) bacteriophages do not carry their own tRNA and depend entirely on the host tRNA pool for decoding their codons. So one would predict that amino acid usage in these phages should be correlated with the abundance of tRNAs in the host cell. This prediction is tested in a study (Chithambaram et al. 2014b) of phages infecting *E. coli*, by using tRNA gene copy number in *E. coli* as a proxy of tRNA abundance (Fig. 9.9). An amino acid carried by more tRNA is used more frequently than another carried by few tRNAs.

4.2 Number of Synonymous Codons

In the lack of any selection, we would expect amino acid usage to increase with the number of synonymous codons (Fig. 9.10). However, this relationship is confounded with the number of tRNAs carrying each amino acid in the cell. If we designate the number of tRNA carrying amino acid *i* as $N_{i.\mathrm{tRNA}}$ and the number of synonymous codons for amino acid *i* as $N_{i.\mathrm{syn\ codon}}$, then amino acid usage depends on both. $N_{i.\mathrm{tRNA}}$ and $N_{i.\mathrm{syn\ codon}}$ are also positively correlated.

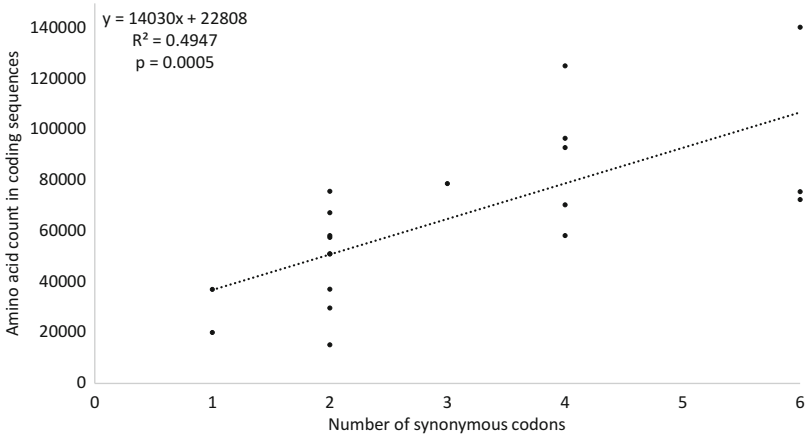

Fig. 9.10 Amino acid count in all coding sequences in *E. coli* I12 (NC_000913) increases with number of synonymous codons

4.3 Genomic Mutation Bias

E. coli genomes have roughly equal nucleotide frequencies. A more AT-rich or GC-rich genome would tend to have more AT-rich or GC-rich codon and their encoded amino acids. For example, AT-rich genomes in bacterial pathogens tend to have many more lysine (encoded by AAA and AAG) than less AT-rich genomes (Xia and Palidwor 2005). This is highly visible even with mild difference in genomic AT content. For example, yeast (*Saccharomyces cerevisiae*) is only mildly AT-rich (0.3090, 0.1917, 0.1913, and 0.3080 for A, C, G, and T, respectively), but the yeast clearly uses more amino acids encoded by AT-rich codons and fewer amino acid encoded by GC-rich codons (Table 9.10).

In summary, amino acid usage (U) is a function of four factors:

$$U = F\left(E, N_{\text{tRNA}}, N_{\text{syncodon}}, \text{GC}\%\right) \tag{9.14}$$

where E is energetic cost, N_{tRNA} and N_{syncodon} have been defined before, and GC% is genomic GC% reflecting mutation bias. One needs to include all these factors in a model in order to reach a reasonable understanding of the determinants of amino acid usage.

Postscript

I usually will share Simpson's Paradox with students after lecturing on the joint effect of translation initiation and elongation on protein production. If we do not take translation initiation into consideration, we may arrive at a wrong conclusion that

Table 9.10 Amino acid usage in *E. coli* K12 (NC_000913) and *S. cerevisiae* (NC_001133-NC_001148) coding sequences

AA	Codon	*E. coli*	Yeast	*E. coli%*	Yeast%
Ala	*GCT,GCC,GCA,GCG*	*125,332*	*160,810*	*9.5527*	*5.4966*
Arg	*CGT,CGC,CGA,CGG,AGA,AGG*	*72,502*	*130,068*	*5.5260*	*4.4458*
Asn	**AAT,AAC**	**51,075**	**179,836**	**3.8929**	**6.1469**
Asp	GAT,GAC	67,349	171,072	5.1333	5.8473
Cys	TGT,TGC	15,188	37,093	1.1576	1.2679
Gln	CAA,CAG	58,360	115,741	4.4481	3.9561
Glu	GAA,GAG	75,786	191,267	5.7763	6.5376
Gly	*GGT,GGC,GGA,GGG*	*96,701*	*145,433*	*7.3705*	*4.9710*
His	CAT,CAC	29,751	63,505	2.2676	2.1706
Ile	**ATT,ATC,ATA**	**78,845**	**191,677**	**6.0095**	**6.5516**
Leu	TTG,TTA,CTT,CTC,CTA,CTG	140,571	277,988	10.7142	9.5017
Lys	**AAA,AAG**	**57,620**	**214,842**	**4.3917**	**7.3434**
Met	ATG	37,093	60,672	2.8272	2.0738
Phe	**TTT,TTC**	**51,131**	**129,516**	**3.8972**	**4.4269**
Pro	*CCT,CCC,CCA,CCG*	*58,293*	*128,177*	*4.4430*	*4.3811*
Ser	TCT,TCC,TCA,TCG,AGT,AGC	75,661	263,096	5.7668	8.9927
Thr	ACT,ACC,ACA,ACG	70,494	173,084	5.3730	5.9161
Trp	TGG	20,060	30,387	1.5290	1.0386
Tyr	**TAT,TAC**	**37,134**	**98,746**	**2.8303**	**3.3752**
Val	GTT,GTC,GTA,GTG	93,061	162,642	7.0930	5.5592

Amino acids encoded by AT-rich codons are in bold, and those encoded by GC-rich codons are italicized

Table 9.11 Success rate (in percentage) of two surgical treatments for removing kidney stone: "all open procedure" (AOS) or percutaneous nephrolithotomy (PN), taken from Table 2 of Charig et al. (1986)

Size	AOS	PN
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Pooled	78% (273/350)	83% (289/350)

Values in parenthesis are in the format of "Number of successes/number of patients treated." Kidney stone size (Size) is discretized into two categories as in the original paper

codon usage bias contributes little to the rate of protein synthesis, as did by Kudla et al. (2009). Simpson's Paradox, illustrated with data in Table 9.11, presents a similar case in which one would reach a wrong conclusion when one factor is ignored.

Charig et al. (1986) summarized their findings in the abstract on the basis of the last row of "Pooled" data, stating that "Success was achieved in 273 (78%) patients after open surgery, 289 (83%) after percutaneous nephrolithotomy." A reader would have thought that AOS is worse (78% success rate) than PN (83% success rate). However, taking kidney stone size into consideration allows us to immediately reach

an opposite (and correct) conclusion, i.e., AOS is better than PN for both small stones (93% vs. 87%) and large stones (73% vs 69%). We also note that both AOS and PN have much higher success rate for small stones than for large stones. Patients treated with PN had mostly small stones and patients treated with AOS had mostly large stones. It is this association between PN and small stone that leads to the misleading conclusion that PN is better than AOS when kidney stone size is ignored.