Published in final edited form as:

Methods Mol Biol. 2009 ; 537: 207–232. doi:10.1007/978-1-59745-251-9_10.

CodonExplorer: An Interactive Online Database for the Analysis of Codon Usage and Sequence Composition

Jesse Zaneveld^{*}, Micah Hamady^{*}, Noboru Sueoka, and Rob Knight

Abstract

The analysis of DNA composition and codon usage reveals many factors that influence the evolution of genes and genomes. In this chapter, we show how to use CodonExplorer, a web tool and interactive database that contains millions of genes, to better understand the principles governing evolution at the single gene and whole-genome level. We present principles and practical procedures for using analyses of GC content and codon usage frequency to identify highly expressed or horizontally transferred genes and to study the relative contribution of different types of mutation to gene and genome composition. CodonExplorer's combination of a user-friendly web interface and a comprehensive genomic database makes these diverse analyses fast and straightforward to perform. CodonExplorer is thus a powerful tool that facilitates and automates a wide range of compositional analyses.

Keywords

Database; sequence composition; codon usage; codon adaptation index; gene transfer; genomic islands; web interface

1. Introduction

Different organisms use codons for the same amino acid at different frequencies and this non-randomness can be exploited for a variety of analyses of both theoretical and practical importance. Soon after the genetic code was first elucidated (1,2) it was clear that it was “degenerate” – not in the sense of decaying from a more elaborate form, but in the physicist's sense of having multiple equivalent states. In this case, because there are 64 codons but only 20 amino acids that are universal to all life, some amino acids must be encoded by more than one codon (exceptions to this rule, such as selenocysteine and pyrrolysine, are important but lineage-specific). We might expect these equivalent codons to be used at random, but early studies of nucleotide frequencies (3) suggested that different organisms have extreme biases toward certain synonyms. The availability of the first DNA sequences for protein-coding genes (4–6) confirmed that codon usage is in fact highly non-random and differs dramatically among organisms.

Why should different organisms use codons that have the same meaning with such different frequencies? An innovative analysis of highly-expressed genes by Ikemura et al. showed that genes that are known to be required at high levels, such as ribosomal proteins, have substantially different codon usage from the average gene in the same genome (7). This suggested a strong effect of translation bias on codon usage because the tRNA pool in a given organism would recognize different anticodons with different efficiencies. For example, the GNN anticodon is

^{*}These authors contributed equally to this work

often used to recognize either CNN or UNN at the first position in so-called “wobble pairing” (8,9), but the efficiency of these two pairings can differ substantially. Consequently, highly-expressed genes would be expected to be composed of codons that match the tRNA anticodons that are expressed at the highest levels and/or have the highest copy numbers; this effect has been confirmed in bacteria (10,11), yeast (12), *Drosophila* (13), and many other species. Thus, selection for translational efficiency has been shown to be one important influence on codon usage bias.

Many different measures of the codon usage bias in a given gene have been developed. One of the most popular is Sharp and Li’s codon adaptation index (CAI) (14), which indicates how closely the codon usage of each gene matches the codon usage of a set of known highly expressed genes. The relationship between codon usage and expression has proven useful for optimizing the expression of synthetic genes transferred into a host (reviewed in (15); *see also* Note 1 on the “Synthetic Gene Designer” website).

Selection for translational efficiency is not, however, the only influence on codon usage; the overall composition of the genome also has a strong effect on which codons are chosen. For example, Muto and Osawa were able to show surprisingly strong, linear trends relating the GC content of the whole genome to the GC content at the first, second, and third codon positions in a diverse assortment of 11 organisms (16). Knight et al. confirmed that these trends held true in hundreds of different organisms across all three domains of life and demonstrated that the codon usage of a given organism could be predicted with considerable accuracy simply from a knowledge of these overall trends and the GC content of the individual organisms (17). Consequently, both selection and mutation play a large role in structuring the codon usage of genes and genomes. In fact, multivariate analyses of codon usages typically identify expression and GC content of individual genes as the principal factors structuring composition within a genome (18).

Because codon usage is strongly influenced by mutational processes, codon usage can provide insight into the different mutational processes operating in different genomes (19–21), or in different parts of the same genome where regions of compositional heterogeneity such as isochores exist (22,23). The ratio of different kinds of codons can provide insights into whether deamination or oxidation contributes more to the pattern of codon usage in a specific organism through techniques such as the fingerprint plot and the PR2 bias plot (24–26).

Finally, codon usage and other compositional information can be used to detect horizontal gene transfer (HGT), the movement of genes between different genomes (27). Because different

¹Other tools for codon usage analysis are available online. Examples include the following:

- *Codon W*: A web interface useful for performing correspondence analysis on codon usage values, which can also calculate CAI values and other measures of codon usage bias (38). (<http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html>).
- *Ade4* and *seqinr* packages: seqinr provides a platform for sequence retrieval and analysis using the R statistical package. Ade4 is a package for multivariate statistical analysis and graphical display and allows many graphs related to codon usage to be plotted, including correspondence analysis. The ade4 and seqinr R packages can be run through the Rweb interface (<http://pbil.univ-lyon1.fr/Rweb/Rweb.general.html>). Documentation for ade4 (<http://pbil.univ-lyon1.fr/ADE-4/home.php?lang=eng>) (39) and seqinr (<http://pbil.univ-lyon1.fr/software/seqinr/home.php?lang=eng>) is also available.
- *Genometrics Website*: DNA walks for bacteria, archaea, and several plasmid and bacteriophage genomes are available at the Genometrics website (http://www2.unil.ch/comparativegenometrics/DNA_walk.html) (40).
- *Synthetic gene designer*: This tool, available at <http://www.evolvingcode.net/codon/sgd/index.php>, is designed to help optimize codon usage for heterologous gene expression (41).
- *Codon Usage Tabulated from Genbank (CUTG) database*: (<http://www.kazusa.or.jp/codon/>) Provides codon usage statistics for sequences for 32,775 organisms as well as for sequences derived from bacteriophage, mitochondria, and chloroplasts. Organisms can be browsed or searched by name (42).

organisms differ in the composition of their genes, HGT can be detected, if sufficiently recent (28), by looking for genes of unusual composition (29,30). However, some caution must be exercised in this approach (*See also* Note 2), since highly expressed genes can also show codon bias, and gradients in composition can appear along a genome due to replication-coupled biases (31,32).

2. Program Usage

CodonExplorer (available at <http://bmf.colorado.edu/codonexplorer/>) provides a platform for conducting many diverse analyses of codon usage and nucleotide composition in sequenced genes or genomes. One feature of CodonExplorer is the ease with which millions of sequenced genes and hundreds of genomes can be searched, grouped, and analyzed. By pre-computing statistics from whole-genome databases, using hundreds of hours of CPU time, CodonExplorer is able to provide graphical summaries of vast, complex datasets very rapidly.

2.1. Gene Selection

The first step in performing sequence composition analysis using CodonExplorer is to gather the sequences on which the analysis will be performed. CodonExplorer incorporates multiple convenient methods for selecting nucleotide sequence data.

The first step when using CodonExplorer is to select a set of genes to analyze. There are several different search methods, including by KEGG gene IDs (*see* Note 3), KEGG orthology (KO) groups, or by enzyme commission (EC) ID (*see* Note 4). All genes in one or more genomes (specified by KEGG genome ID or using the NCBI taxonomy) can be conveniently gathered, and a gene screen allows selection of highly expressed genes or putatively horizontally transferred genes within a genome. Finally, many genomic and pathogenicity islands (regions of putatively transferred genes) from the literature have been included and can be easily selected.

Once sets of genes have been selected, they can be sorted, grouped, and edited manually to remove undesired sequences or further refine the set. The KEGG gene IDs for the final set of sequences can be downloaded to allow easy reproduction of the analysis.

2.1.1. Selecting Genes by Gene or Genome—CodonExplorer allows genes to be selected with a list of KEGG genomes, KEGG gene IDs, KEGG gene names, and/or NCBI taxonomy names (Fig. 10.1) (*see* Note 5).

²One disadvantage to compositional methods for HGT detection is that transfer between related or unrelated strains with similar compositional characteristics will not be detected. Conversely, if a sequence characteristic varies within a genome, then classes of genes, such as ribosomal proteins, that have unusual compositions may appear to have been horizontally transferred. Even when the composition of a transferred gene is initially distinct from the composition of the recipient genome, the composition of the gene will drift toward that of the recipient genome over time. Thus, the traces of ancient transfers may be obliterated in sequences that have had time to equilibrate fully (28).

³The Kyoto Encyclopedia of Genes and Genomes (KEGG) is an online database that seeks to organize and integrate information on biological systems from the genomic to the environmental level. As part of this effort, KEGG contains re-annotations of genome sequences taken from RefSeq. Each gene within the genomes re-annotated by KEGG is given a unique KEGG gene id. Similarly, each genome is assigned a KEGG genome id. The KEGG website is <http://www.genome.jp/kegg/>.

⁴For details on EC groups, see the website of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (<http://www.chem.qmul.ac.uk/iubmb/enzyme/>).

⁵Only genes that match all properties in the fields under search by name or ID will be reported. Thus entering “rpsU” in the gene name field and “stm” (the KEGG genome abbreviation for *Salmonella enterica* serovar Typhimurium LT2) will find only the rpsU gene from *Salmonella*, whereas the same search without a name in the genome field will return records for all genes named “rpsU” across KEGG genomes.

Gather genes by KEGG genome: To gather all genes in one or more genomes, a list of KEGG genome IDs (*see* Note 6) may be entered and the genes from each genome will be grouped and sorted according to the selected option (see grouping and sorting data, below).

Gather genes by NCBI taxon name: To gather all genes in an NCBI taxon, enter the NCBI taxon name in the field labeled “Enter NCBI taxon name(s)” and press the “Search for Genes” button (*see* Notes 6, 7).

Gather genes by KEGG ID or gene name: To gather a list of genes by KEGG ID or gene name, enter the IDs or gene names in the appropriate fields. If search terms are entered in multiple fields, only genes in the intersection of terms are returned (i.e., the terms are combined with an “AND” operator).

Example: Searching for KEGG genome “STM” and KEGG gene name “rpsu” returns only the rpsU gene from *Salmonella entericaserovar* Typhimurium LT2; searching for KEGG genome “STM” will return every gene from *Salmonella enterica* serovar Typhimurium LT2.

2.1.2. Selecting Genes by Ortholog, EC, or Gene Region—When a set of a precise set of gene IDs is not known in advance, CodonExplorer provides several convenient ways to select groups of genes sharing particular properties (Fig. 10.2).

Selecting Genes by KEGG Orthology (KO) Group: It is often useful to analyze related genes across several genomes. CodonExplorer facilitates this type of analysis by allowing users to select members of a desired KO group. (*See also* Notes 8, 9.)

Example: To select homologs of recA, enter “K03553” into the KEGG ortholog or KEGG EC ID box, select “KEGG Ortholog” from the “Select type of ID” pull-down menu, and press the “Search for Genes” button at the bottom of the page.

Selecting Genes by Enzymatic Activity: CodonExplorer allows for users to search for genes encoding enzymes that share a common activity, as defined by their KEGG EC ID. To select a group by KEGG EC ID, enter the ID or list of IDs into the “KEGG ortholog or KEGG EC ID” and select “KEGG EC” from the “Select type of ID(s)” menu (*see* Note 10).

Example: To search for α -amylases (EC3.2.1.1), enter EC3.2.1.1 into the KEGG ortholog field and select “KEGG Ortholog” from the “Select type of ID(s)” pull-down menu. Then press the “Search for Genes” button at the bottom of the page.

Selecting Genes by Genomic Region: CodonExplorer allows users to select a range of genes within a genome. To do so, enter the KEGG gene ID for the first and last genes within the

⁶A list of KEGG genome IDs, grouped by phylogeny, can be found at the KEGG website (http://www.genome.jp/kegg/catalog/org_list.html). Similarly, a list of NCBI taxon names can be found at <http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root>.

⁷Due to computational constraints, users are currently allowed to select a maximum of 100 genomes at a time for analysis, so selecting large or heavily sampled lineages (e.g., all Gamma-proteobacteria) for simultaneous analysis using the NCBI taxonomy search is not possible at this time.

⁸Because KO groups may contain multiple members in a single genome (e.g., there are currently two members of K03553 in the *Bacillus cereus* e331 genome), members of a KO group are not always strict orthologs.

⁹If the KO number for a particular group of genes is unknown, but an example of the group is known, one can search for the known gene by name or ID (*see* above); click “View” (in the view genes column) for a genome containing the gene name or ID, then follow the link under the “description” column to the KEGG gene description page. This description includes the KO group to which the gene belongs.

¹⁰Currently only individual EC IDs are allowed, and not IDs for entire EC groups. However, if all EC IDs within a group are desired, each EC category in the class can be entered on a separate line.

region in the “Enter KEGG gene range” field and select “Forward Only,” “Reverse Only,” or “Both Directions” from the “Gene region direction” pull-down menu (*see* Note 11).

Example: To select genes in either orientation between STM1379 and STM1422 in the *Salmonella enterica* serovar Typhimurium genome (i.e., the Salmonella Pathogenicity Island-2 region), enter “STM1379” in the “Gene Start” field, “STM 1422” in the “Gene End” field, “STM” in the KEGG genome ID field, select “Either direction” in the “Gene region direction” pull-down menu, and press the “Search for Genes” button at the bottom of the page.

2.1.3. Selecting Genes in Genomic Islands or by Gene Screen

Selecting Genes in Pre-Annotated Genomic Islands: Many regions of genes present in a particular genome but absent in closely related lineages have been described in the literature (33,34). Frequently these regions, often called “genomic islands,” have unusual dinucleotide compositions or contain mobility genes such as integrases or transposases (34). Thus, it is suggested that genomic islands are the product of HGT. A subset of these genomic islands called “pathogenicity islands” also contains pathogenicity determinants, making them particularly interesting.

CodonExplorer allows users to select genomic islands that have been described in the literature by selecting a genomic island from the “Select genomic island” pull-down menu (Fig. 10.3). The genomic islands used in Hsiao et al.’s 2005 analysis (34) are included, as are many other islands collected manually from the literature (*see also* Note 12).

Selecting Highly Expressed or Putatively Transferred Genes: For codon usage or other compositional property analyses, it is often useful to explore variation in highly expressed genes (e.g., those encoding ribosomal proteins) or genes that may have entered the genome by HGT.

To select such a set of genes in CodonExplorer, enter the KEGG genome name in the “KEGG genome” in the gene screen section, then select “Ribosomal + highly expressed KO groups,” “Putative Aliens,” or “Not putative aliens” from the pull-down menu (Fig. 10.3). The putative and aliens and non-alien sets are defined using a Markov Model method described by Nakamura et al. (30). However, like any compositional method, the Nakamura method should not be considered definitive evidence of transfer. See also the discussion below about the benefits and limitations of compositional methods for detecting gene transfer.

2.2. Output Options for Sorting and Grouping Selected Genes

By default, genes selected for analysis in CodonExplorer are grouped together by genome and sorted in descending order of the number of selected genes in each genome. However, it is sometimes useful to group and sort genes in different ways.

Grouping selected genes—CodonExplorer allows genes to be grouped by KEGG genome name, KEGG ortholog ID, KEGG EC ID, or (when NCBI taxon names were entered in the search) NCBI Taxonomy name (Fig. 10.4). To change how genes are grouped, enter your search

¹¹In some analyses it is desirable to separate genes on each strand of the DNA. One can do so in CodonExplorer by using the region selector for the entire genome, choosing only genes in either “forward” or “reverse” orientation, then repeating the analysis using genes in the other orientation.

¹²The boundaries of many genomic islands are not unambiguously defined and the gene ranges given in the pull-down menu should not be considered definitive. In cases where a better gene range for a gene has been determined (e.g., by comparative or compositional analysis), literature islands can be expanded or reduced based upon any extra information possessed by the user by searching for the predefined island, noting the first and last genes, and then using the Search by Gene Region feature to define a new region.

terms as described above, then select the appropriate option from the “Group matching genes by” pull-down menu before pressing the “Search for Genes” button.

Sorting selected genes—CodonExplorer also allows groups of genes to be sorted by the number of genes in the group, the group ID, or the group description (Fig. 10.4). Select the desired option from the “Sort gene groups by” pull-down menu under the “Output options” heading.

2.3. Search Results – Viewing, Editing, and Saving Sets of Genes

After searching for genes on which to perform an analysis, determining how those genes should be sorted and grouped, and pressing the “Search for Genes” button at the bottom of the page, the genes will be displayed at the top of the page (Fig. 10.5) and sorted and grouped according to the selected options (*see* Note 13). From this screen it is possible to view, edit, or save gene groups.

View genes—Each gene group can be viewed by pressing the “View” link to the right of the group description. This will bring up a window listing the genome, KEGG gene ID, KEGG gene name, and a description for each gene (Fig. 10.6). Clicking the “description” link will bring up the KEGG description page for that gene.

Include or exclude gene groups from analyses—To include or exclude one or more gene groups from the analyses, check or uncheck the check box to the left of each gene group’s ‘Group ID’.

Save genes—The KEGG primary gene IDs for genes in each group can be saved locally as a text file by clicking the “Save Genes” link next to that group.

2.4. Search Results – Selecting Analyses

Once one or more groups of genes have been selected, CodonExplorer allows many compositional analyses to be performed easily. In general, one or more analyses are selected, after which pressing the “Generate selected graph(s)” will cause a new window to appear displaying the desired plots. There are dozens of different analyses available, including comparisons of GC content at different codon positions, measures of codon usage bias, amino acid usage, and a Markov model method for HGT detection (Fig. 10.7). Custom scatter plots can also be constructed plotting many of these properties against one another, or against other properties such as length or hydrophobicity. Finally, a Monte Carlo method allows determination of the statistical significance of apparent differences between the properties of selected groups of genes and the genomes that contain them. For ease of use, we have classified these diverse approaches into the broad categories of “Codon Usage,” “GC Content,” and “Custom Graphs.”

2.4.1. Codon Usage Analyses

Codon Fingerprint Plots: Fingerprint plots were introduced by Sueoka (20) to measure mutational biases within and between genomes. The y -axis represents the frequency of A at position 3 relative to the frequency of A and T at that position ($A_3/(A_3+T_3)$), while the x -axis represents $G_3/(G_3 + C_3)$. Each circle represents a different amino acid, with the radius of the circle proportional to the relative frequency of that amino acid (19). There are several options when generating fingerprint plots. The *Quartets* option generates a codon fingerprint plot using

¹³If searches for genes or genomes repeatedly fail, it is sometimes useful to click the “Clear Search Results” button at the bottom of the page to ensure that all previous entries have been cleared.

the four codon blocks (e.g., the CCN block coding for Proline, where N stands for any of the four nucleotides A, C, G, or T). These blocks are Leucine4, Valine, Alanine, Threonine, Proline, Serine4, Glycine, and Arginine4 (the 4 refers to the four-codon block of an amino acid that has six codons in total). The **Split Quartets** option generates a codon fingerprint plot using each of the six split codon blocks (e.g., the GAN block coding for Aspartic or Glutamic acid). These blocks include Phenylalanine/Leucine2, Isoleucine/Methionine, Histidine/Glutamine, Asparagine/Lysine, Aspartic Acid/Glutamic Acid, and Serine/Arginine2. The Tyrosine/Stop and Cysteine/Tryptophan/Stop partial blocks are not included on this plot.

Three graph types are available for codon fingerprint blocks in CodonExplorer. First, the **Combined** option plots all genes in the selected gene groups into a single fingerprint plot. This is useful when for summarizing large data sets, such as whole genomes, DNA strands within a genome, or genomic islands. Second, the **Each Gene** option allows for a finer grained analysis by plotting each gene within the set of genes selected for analysis separately (*but see* Note 14). Finally, the **Gradient** option plot displays a combined fingerprint plot for genes within each 10% interval of GC content. Thus, depending on the number of genes selected, and the range of GC content represented by those genes, only some of the ten possible ranges of GC content will be displayed. An example of constructing and interpreting a gradient codon fingerprint plot is given in **Section 3**.

CAI versus P₃ Plots: As discussed above, GC content has an important influence on codon usage bias (Guy and Sueoka, manuscript in preparation). This influence can be illustrated by plotting CAI values against P₃ GC content (see the examples section for some plots of CAI vs. P₃ in organisms with very different genomic GC contents).

PR2 Bias Plots: Chargaff's second parity rule (PR2) states that within each DNA strand, the frequency of A $\approx$ T and the frequency of G $\approx$ C (35). Parity Rule bias plots were developed to examine deviations from PR2 (26). Selecting the PR2 bias plot option generates a set of PR2 bias plots for the selected gene set.

2.4.2. GC Content Plots—GC content at different codon positions is used to study mutation and selection pressures. Because GC content varies dramatically between but not within bacterial lineages (36), unusual GC content may indicate transfer. Thus unusual GC content has often been used as an indicator of gene transfer, although there are important limitations to this approach (see the example on HGT detection below). CodonExplorer allows the generation of several graphs that allow users to explore the GC content of genes and genomes.

P₁₂ versus P₃ GC: Useful for contrasting the effects of mutation and selection for amino acid usage (P₁ and P₂) versus codon usage (P₃). The *x*-axis represents the GC content at the third position within each codon in the selected group of genes, and the *y*-axis represents the GC content at either the first or second position in each codon in the selected group of genes. (P₁ and P₂ GC contents are plotted as separate series.)

P₁₂ versus P₃: Plots the combined GC for nucleotides at the first or second position in each gene against the GC content for nucleotides at the third position in that gene. The dotted diagonal line represents the expectation if all positions were under identical mutation and selection pressure.

¹⁴When plotting codon fingerprints, it is important to remember that the smaller the dataset, the larger the expected visual difference in the plots due to chance. While this is not likely to be a major problem at the whole genome level, it can be an important factor when analyzing codon fingerprints for each gene individually, or when using the P₃ gradient option. It is a good idea to test the significance of an observed visual difference in codon fingerprint plots using Monte Carlo analysis (see the section on Monte Carlo Histograms, above).

P₁₂₃ versus GC: Plots the first, second, and third position GC content (as separate series) against the overall GC content of the gene. Thus each point represents the P₁, P₂, or P₃ GC content of a gene, versus the total GC content of that gene.

P₃ versus All Amino Acids Plots: This plot allows the user to view a plot of the amino acid content of each gene within the group of selected genes against the P₃ GC content of that gene. It should be noted that since this approach divides the data contained in each gene into 20 parts (by recording P₃ information for usage of each amino acid for each gene), these plots may be subject to large amounts of noise when examining small datasets. (See Note 14 for a similar caution regarding gradient fingerprint plots).

2.4.3. Custom Graphs

Custom scatter plot: In addition to allowing the generation of many predefined plots, CodonExplorer also allows users to construct custom scatter plots using a variety of compositional statistics.

Values that can be plotted against one another on either axis include the CAI (as well as several proposed variations on it), the third position GC content (P₃), gene length, predicted peptide hydrophobicity, and the horizontal transfer index of Nakamura et al. (30) (Table 10.1).

Histograms: Histograms can be constructed of CAI values, GC content at the third codon position (P₃), gene length, or peptide hydrophobicity.

Codon Usage Table: Often it is useful to report a table of codon usages for a single gene or a collection of genes, e.g., to identify frequently used codons (37). This can be done in CodonExplorer by selecting the “Codon Usage Table” option. The output is a table (Fig. 10.8) that lists the frequency of each codon per thousand codons in the dataset, followed by the absolute number of occurrences of the codon in the selected set of genes.

The total number of genes and codons in the data set are listed at the top of the table, and the overall GC content and the average GC content at positions 1, 2, and 3 are listed at the bottom of the table. This format is similar to that of the CUTG database (See also Note 1).

Monte Carlo Histogram Plots: It is often desirable to compare the properties of some subset of the genes in a genome to the properties of the genome as a whole. For example, one may wish to determine whether the CAI values for genes in a proposed genomic island are significantly lower than those for the genome as a whole. However, visually comparing histograms of CAIs for the island and for the genome can lead to a false inference of difference simply because a small set of genes will tend to have more variation in compositional properties than the genome as a whole. Thus it is useful to be able to test whether observed differences are significant or if they could be explained simply by the reduced size of the examined sub-sample of the genome.

CodonExplorer provides a Monte Carlo method for evaluating the significance of differences between a set of genes and the genome as a whole. The procedure involves generating many (10, 100, or 1,000, at the user's option) random sets of a number of genes equal to those in the selected subset (chosen individually or as a block of contiguous genes). The mean value of the property is then recorded for each randomly chosen set of genes. Finally a figure is generated showing a histogram of the average values for each random set (Fig. 10.9). The blue circle (dark grey here) represents the mean value of the chosen property for these random sets. The red circle represents the average of the chosen property for the query set. Finally, the light red bars represent a histogram of the values of the chosen property in the query set. Comparing the location of the average of the query set (red circle) against the histogram of average values

for equally sized random sets (blue bars) allows the user to determine if the set of genes that they selected varies from other random equally sized gene sets more than would be expected from chance. The caption at the top of the Monte Carlo histogram figure lists the number of random gene sets constructed, how many of those random gene sets had a higher value for the chosen property than the query set, and the probability (as assessed using the one-tailed *t*-test) that the query set average could be as different as it was from a normal distribution fit to the distribution of random set averages by chance.

2.5. Starting a New Search

To start over and select a new set of genes for analysis, press the “new search” button at the top right of the analysis screen (your old set of genes will not be saved).

3. Examples

Many classic analyses of codon usage and DNA sequence composition can quickly and easily be recapitulated in CodonExplorer. This section provides a step by step guide to using CodonExplorer to examine the effects of selection, mutation, and horizontal transfer on codon usage.

3.1. Codon Usage as an Indicator of Highly Expressed Genes

Histograms of CAI values, and plots of CAI against GC content at the third codon position (P_3), can provide insight into selection for translational efficiency. Monte Carlo techniques are useful in assessing the significance of differences in CAI or P_3 /CAI between a subset of genes and the genome as a whole. To generate the plots of CAI versus P_3 shown in Figs. 10.10 and 10.11 requires analysis of all genes in the *Salmonella enterica* serovar Typhimurium LT2 genome (to generate panel A and C in Fig. 10.10), followed by an analysis of highly expressed genes in *Salmonella* (to generate panels B and D in Fig. 10.10 and panels A and B in Fig. 10.11).

To select all genes in the *Salmonella enterica* serovar Typhimurium LT2 genome for analysis, enter “stm” (see Note 6) in the “Enter KEGG genome abbreviation(s)” field. Then press the “Search for Genes” button at the bottom of the page. A new screen will appear showing the group of all genes in *Salmonella*, and several available analyses. Check the “scatter plot” check box by the “CAI vs. P_3 ” header (to select the analysis in Fig. 10.10 A), then check the “CAI” check box to the right of the “Histograms” header (to select the analysis in Fig. 10.10 C). Finally, press the “Generate selected graph(s)” button at the bottom of the screen. After a brief wait, the images in Fig. 10.10 A and C will appear in a new window. (Right-click on the figures and select “Save image as...” to save them to disk as .png images.)

When you finish examining and/or saving the figures, close the figure window, and press the “new search button” at the top of the analysis screen to start a new search. On the search screen, locate the “Gene screen” header (Fig. 10.3), enter “stm” into the field to the left of the “KEGG genome abbreviation” header, and select “Ribosomal proteins + highly expressed Kos” from the pull-down menu. Press the “Search for Genes” button at the bottom of the screen.

When the analysis screen appears, check the boxes to select a “CAI vs. P_3 ” scatter plot and a CAI histogram as before (to generate Fig. 10.10 B and D). Finally, to generate Fig. 10.11 A, check the “generate Monte Carlo Histograms for the selected gene group(s)” checkbox, then select “CAI (two g)” from the “Compositional measure” pull-down menu, and the desired number of replicates (in this case 100) from the “Number of replicates(n)” pull-down menu. Set the “Sample reference Set” pull-down menu to “individual genes” (see Note 15). Then click the “Generate Selected Graph(s)” button. After a brief pause, the images in Figs. 10.10 B, D, and Fig. 10.11 A should appear in a new window. After examining and (if desired) saving

the images to disk, return to the select analysis screen and uncheck all boxes other than the Monte Carlo histogram box. Then set the “Compositional measure” pull-down menu to “P₃/CAI.” Once again press “Generate selected graph(s)” at the bottom of the screen. The image from Fig. 10.11 B should appear in a new window.

Along with mutation pressure, selection for translational efficiency is believed to be an important force driving codon usage. Thus, those genes that are more highly expressed are expected to have unusually high CAI values for their GC content. Ribosomal proteins are highly expressed (although not all have high CAIs), as are elongation factors and some membrane factors (27). Highly expressed genes can often be detected by plotting CAI versus P₃ (Fig. 10.10).

The trends in CAI versus P₃ plots or histograms of CAI can be confirmed using the Monte Carlo histogram feature (Fig. 10.11). This is useful because small sets of genes frequently display apparently unusual codon usage or sequence compositional properties simply due to the effects of a small sample size. Thus, comparing against randomly chosen gene sets of the same size can be used to demonstrate or refute the statistical significance of an apparent difference in the plots.

3.2. Codon Usage as an Indicator of Mutation Pressure

CodonExplorer can be used to generate various P₃ versus CAI plots useful for studying the effect of mutation pressure on codon usage. To generate the P₃ versus CAI plots shown in Fig. 10.12, enter the KEGG abbreviations for the genomes desired (*see* Note 6) in the “Enter KEGG genome abbreviation(s)” field, entering one genome abbreviation per line. (When generating Fig. 10.12, “mge,” “eco,” and “sco” were entered on separate lines to search for all genes in *Mycoplasma genitalium*, *Escherichia coli* K-12 MG1655, and *Streptomyces coelicolor*, respectively.) Then press the “Search for Genes” button. A new screen will appear listing the genes found by the search, and available analyses. To generate the plots in Fig. 10.12 click the “scatter plot,” “heat-map,” and “contour plot” checkboxes next to the “CAI vs. P₃” header. Make sure the “Combine Data Sets” option is set to “No,” then click the “Generate Selected Graph(s)” button. After a brief wait, figures corresponding to each of the nine panels in Fig. 10.12 will be generated.

Another type of plot useful for analyzing the effects of mutation pressure is the fingerprint plot (19). To generate the fingerprint plot shown in Fig. 10.13, first ensure that you are on the gene search page (if you are on the select analysis page press the “new search” button at the top of the page to start a new search). Search for all genes in the human genome by entering the KEGG genome id for *Homo sapiens* (hsa) in the “Enter KEGG genome abbreviation(s)” field (*see* Note 6). Then press “Search for Genes.” A new screen will appear displaying the selected genes and various options for analysis. Select the “combined” and “gradient” checkboxes next to the “Fingerprint plots” header in the codon usage section (*see* Fig. 10.7). Finally, click the “Generate Selected Graph(s)” button at the bottom of the page. A combined fingerprint plot as well as fingerprint plots for genes in each 10% bin of third codon position GC content (P₃) will be generated. The combined fingerprint plot for *Homo sapiens* (Fig. 10.13 A) as well as the fingerprint plots for the 70–80% (Fig. 10.13 B), 80–90% (Fig. 10.13 C), and 90–100% GC content (Fig. 10.13 D) bins for this genome are shown in Fig. 10.13.

¹⁵Comparing a contiguous block of putative HGT genes to individual random genes using the Monte Carlo tool may bias the analysis, since genes are not distributed randomly on the genome. Some classes of genes, such as those encoding ribosomal proteins, are both unusual in composition and clustered together in the genome. Thus it is usually better to compare a contiguous set of genes (e.g., a putative genomic island) to other similarly sized contiguous blocks of genes.

The scatter plot of CAI versus P_3 illustrates the variable effect of GC content on CAI in different lineages (Fig. 10.12 A–C). In the *Mycoplasma genitalium* genome, which possesses a low GC content, there exists a strong negative correlation ($r^2 = 0.86$) between the position three GC content and CAI, whereas in the high-GC *Streptomyces coelicolor* CAI and GC content are positively correlated ($r^2 = 0.73$).

One difficulty when examining scatter plots of CAI versus P_3 is that it can be difficult to determine the relative density of points in different regions of a scatter plot when the number of data points is very large. To help solve this problem, CodonExplorer allows generation of both contour plots (Fig. 10.12 D–F), and heat maps (Fig. 10.12 G–I). These plots more clearly illustrate the relative density of different regions.

The fingerprint plot in Fig. 10.13 illustrates that groups of codons containing different nucleotides at the second codon position tend to change in amino/keto bias together (20). The partial gradient fingerprint plot (Fig. 10.13 B–D) illustrates that although codons for different amino acids respond differently to changes in GC content, groups of codons with the same second position nucleotide tend to respond similarly.

3.3. Codon Usage and Nucleotide Sequence Composition as Indicators of Horizontal Gene Transfer

Monte Carlo techniques are useful for testing the statistical significance of differences in codon usage or nucleotide sequence composition between putatively transferred sets of genes and the genome as a whole. To generate the Monte Carlo histograms in Fig. 10.14, first, select the genomic island of interest from the “select genomic island” pull-down menu (Fig. 10.3), or enter the first and last gene in the island of interest, using the “Search by Gene Region” option (Fig. 10.2; see Program Usage, above). Then press the “Search for Genes” button at the bottom of the page. Next, on the search results/analysis page, select the “generate Monte Carlo Histograms for the selected gene group(s)” checkbox, the desired compositional property from the “compositional measure” pull-down menu (for Fig. 10.14, “CAI (two g)” was selected), and the desired number of replicates (in this case 100) from the “Number of replicates(n)” pull-down menu. Make sure the “Contiguous blocks” option is selected from the “Sample reference Set” pull-down menu (see Note 15) and then click the “Generate Selected Graph(s)” button.

Because optimal codon usage varies in divergent microbial lineages, it has been proposed that genes that have been transferred from one genome to another will have patterns of codon usage more similar to the genome from which they were transferred than to the genome in which they currently reside (27) (Guy and Sueoka, submitted). Thus, unusually low CAI values (e.g., as shown for the SPI-2 pathogenicity island and the Fels-1 prophage in the genome of *Salmonella enterica* serovar Typhimurium LT 2 displayed in the Monte Carlo histograms in Fig. 10.14) are consistent with the hypothesis that these regions were acquired by HGT. However, these values on their own do not prove that such a transfer occurred.

Although an unusual sequence composition for a gene (or group of genes) is often taken as an indication of horizontal transfer, other factors may also explain unusual codon usage. Thus, while unusual codon usage may suggest the possibility of lateral transfer, the hypothesis of transfer should be confirmed using additional methods for HGT detection, such as discordance between gene trees and organismal trees, lack of orthologs in related lineages, unusual dinucleotide usage, or the presence of nearby mobility genes.

Conversely, horizontally transferred genes may, over time, alter their composition and codon usage to match that of the genome in which they reside (28). Thus a lack of unusual codon usage or nucleotide composition does not rule out the possibility of recent horizontal transfer

from a lineage with similar properties, nor does it exclude the possibility of an ancient transfer that is no longer detectable using these methods.

Acknowledgments

This work was supported in part by the Biophysics and SCR training grants T32GM08759 and T32GM065103 from NIH. CodonExplorer is hosted on the W.M. Keck Foundation Bioinformatics Facility at the University of Colorado, Boulder.

References

1. Nirenberg MW, Matthaei JH. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci USA* 1961;47:1588–602. [PubMed: 14479932]
2. Soll D, Ohtsuka E, Jones DS, Lohrmann R, Hayatsu H, Nishimura S, Khorana HG. Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. *Proc Natl Acad Sci USA* 1965;54:1378–85. [PubMed: 5325653]
3. Sueoka N. Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harb Symp Quant Biol* 1961;26:35–43. [PubMed: 13918160]
4. Efstratiadis A, Kafatos FC, Maniatis T. The primary structure of rabbit beta-globin mRNA as determined from cloned DNA. *Cell* 1977;10:571–85. [PubMed: 558827]
5. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;74:5463–7. [PubMed: 271968]
6. Sonneborn TM. Nucleotide sequence of a gene: first complete specification. *Science* 1965;148:1410. [PubMed: 17738135]
7. Ikemura T, Ozeki H. Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harb Symp Quant Biol* 1983;47(Pt 2):1087–97. [PubMed: 6345068]
8. Crick FH. Codon – anticodon pairing: the wobble hypothesis. *J Mol Biol* 1966;19:548–55. [PubMed: 5969078]
9. Agris PF, Vendeix FA, Graham WD. tRNA's wobble decoding of the genome: 40 years of modification. *J Mol Biol* 2007;366:1–13. [PubMed: 17187822]
10. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 1981;151:389–409. [PubMed: 6175758]
11. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* 1981;146:1–21. [PubMed: 6167728]
12. Ikemura T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* 1982;158:573–97. [PubMed: 6750137]
13. Moriyama EN, Powell JR. Codon usage bias and tRNA abundance in *Drosophila*. *J Mol Evol* 1997;45:514–23. [PubMed: 9342399]
14. Sharp PM, Li WH. The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987;15:1281–95. [PubMed: 3547335]
15. Kane JF. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr Opin Biotechnol* 1995;6:494–500. [PubMed: 7579660]
16. Muto A, Osawa S. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA* 1987;84:166–9. [PubMed: 3467347]
17. Knight RD, Freeland SJ, Landweber LF. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol* 2001;2:RESEARCH0010. [PubMed: 11305938]

18. Gupta SK, Ghosh TC. Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. *Gene* 2001;273:63–70. [PubMed: 11483361]
19. Sueoka N. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Gene* 1999;238:53–8. [PubMed: 10570983]
20. Sueoka N. Wide intra-genomic G+C heterogeneity in human and chicken is mainly due to strand-symmetric directional mutation pressures: dGTP-oxidation and symmetric cytosine-deamination hypotheses. *Gene* 2002;300:141–54. [PubMed: 12468095]
21. Sueoka N. Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci USA* 1988;85:2653–7. [PubMed: 3357886]
22. Bernardi G. The vertebrate genome: isochores and evolution. *Mol Biol Evol* 1993;10:186–204. [PubMed: 8450755]
23. Costantini M, Clay O, Auletta F, Bernardi G. An isochore map of human chromosomes. *Genome Res* 2006;16:536–41. [PubMed: 16597586]
24. Sueoka N. Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. *J Mol Evol* 1999;49:49–62. [PubMed: 10368434]
25. Lobry JR, Sueoka N. Asymmetric directional mutation pressures in bacteria. *Genome Biol* 2002;3:RESEARCH0058. [PubMed: 12372146]
26. Sueoka N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* 1995;40:318–25. [PubMed: 7723058]
27. Karlin S, Mrazek J, Campbell AM. Codon usages in different gene classes of the *Escherichia coli* genome. *Mol Microbiol* 1998;29:1341–55. [PubMed: 9781873]
28. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 1997;44:383–97. [PubMed: 9089078]
29. Groisman EA, Sturmoski MA, Solomon FR, Lin R, Ochman H. Molecular, functional, and evolutionary analysis of sequences specific to *Salmonella*. *Proc Natl Acad Sci USA* 1993;90:1033–7. [PubMed: 8430070]
30. Nakamura Y, Itoh T, Matsuda H, Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* 2004;36:760–6. [PubMed: 15208628]
31. Lobry JR. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene* 1997;205:309–16. [PubMed: 9461405]
32. Faith JJ, Pollock DD. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 2003;165:735–45. [PubMed: 14573484]
33. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* 2000;54:641–79. [PubMed: 11018140]
34. Hsiao WW, Ung K, Aeschliman D, Bryan J, Finlay BB, Brinkman FS. Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet* 2005;1:e62. [PubMed: 16299586]
35. Rudner R, Karkas JD, Chargaff E. Separation of microbial deoxyribonucleic acids into complementary strands. *Proc Natl Acad Sci USA* 1969;63:152–9. [PubMed: 4980029]
36. Sueoka N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA* 1962;48:582–92. [PubMed: 13918161]
37. Sharp PM, Devine KM. Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do ‘prefer’ optimal codons. *Nucleic Acids Res* 1989;17:5029–39. [PubMed: 2762118]
38. Peden, JF. PhD Thesis. Department of Genetics, University of Nottingham; Nottingham, UK: 1999. Analysis of codon usage.
39. Thioulouse J, Chessel D, Dolédec S, Olivier JM. ADE-4: a multivariate analysis and graphical display software. *Stat Comput* 1996;7:75–83.
40. Roten CA, Gamba P, Barblan JL, Karamata D. Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. *Nucleic Acids Res* 2002;30:142–4. [PubMed: 11752276]

41. Wu G, Bashir-Bello N, Freeland SJ. The synthetic gene designer: a flexible web platform to explore sequence manipulation for heterologous expression. *Protein Expr Purif* 2006;47:441–5. [PubMed: 16376569]
42. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from the international DNA sequence databases. *Nucleic Acids Res* 1997;25:244–5. [PubMed: 9016546]
43. Sharp PM, Li WH. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for ‘rare’ codons. *Nucleic Acids Res* 1986;14:7737–49. [PubMed: 3534792]
44. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;157:105–32. [PubMed: 7108955]

CodonExplorer

[Help](#) | [About](#) | [New Search](#) | [Log Off](#)

Enter one or more search criteria below. (Queries are combined and the intersection is returned when multiple criteria are entered.)

Your search results are currently limited to **100** species and **50000** genes. [Contact us if you need to analyze a larger dataset]

Search by Name or ID:

Enter KEGG genome abbreviation(s) - Help	<input type="text"/>
Enter KEGG primary gene ID(s) - Help	<input type="text"/>
Enter KEGG gene name(s) - Help	<input type="text"/>
Enter NCBI taxon name(s) - Help	<input type="text"/>

Fig. 10.1.
Selecting genes by name or ID.

Search by Ortholog or Enzyme Commission ID:

Select type of ID(s) - Help	KEGG Ortholog ▾
KEGG ortholog or KEGG EC ID(s)- Help	<input type="text"/>

Search by Gene Region (Gene Range or Known Genomic Island): Please note that only one of the following search criteria can be specified at a time.

Gene region direction - Help	Either Direction ▾
Enter KEGG gene range - Help	<input type="text"/> KEGG genome abbreviation
	<input type="text"/> Gene start
	<input type="text"/> Gene stop
	<input type="text"/> 0 Number genes plus/minus ends of region

Fig. 10.2.
Selecting genes by KEGG orthology (KO) group, enzyme commission (EC) ID, or gene region.

Select genomic island - [Help](#)

-OR-

Gene Screen:

Specify gene screen - [Help](#)

Fig. 10.3.
Genomic islands and gene screen.

Output Options:

Group matching genes by - Help	KEGG genome name ▼	
Sort gene groups by - Help	Number of genes in group ▼	Descending order ▼

Search for Genes	Clear Search Fields
------------------	---------------------

Fig. 10.4.
Output options.

Your search returned **18575** genes from **5** species. Click [here](#) for general search results help, or click on any one of the help links below for more more specific help.

Select Gene Group(s): (The selected groups are used when generating the graphs below.)

☒	Group ID	Genes	Group Description	View Genes	Save IDs
☒	KEGG Genome: ATU	5402	Agrobacterium tumefaciens c58 (u.washington/dupont)	View	Save
☒	KEGG Genome: BCL	4096	Bacillus clausii ksm-k16	View	Save
☒	KEGG Genome: BHA	4066	Bacillus halodurans c-125	View	Save
☒	KEGG Genome: MGE	484	Mycoplasma genitalium g-37	View	Save
☒	KEGG Genome: STM	4527	Salmonella typhimurium lt2	View	Save

Fig. 10.5.

Search results: all genes in genomes “ATU, MGE, STM, BHA, BCL.”

Displaying detailed information on **3 genes**.

Click on the gene description to view more gene detail.

Genome	Gene ID	Gene Name	Description
ATU	Atu3123	rpsU	30S ribosomal protein S21
ATU	Atu3637	rpsU	30S ribosomal protein S21
ATU	Atu4064	rpsU	30S ribosomal protein S21

Click [here](#) to close this window.

Fig. 10.6.

“View Genes” for rpsU group from ATU are shown.

Select Analyses: (You may select one or more analyses from the groups listed below.)

Codon Usage:

Fingerprint plots - Help	Codon blocks: Quartets ☐ Combined ☐ Each gene ☐ Gradient
PR2 Bias - Help	☐ PR2 Bias
Codon Usage Table - Help	☐ Codon Usage Table

GC Content:

P12 versus P3 - Help	☐ Scatter Plot ☐ Contour ☐ Heatmap
P12 versus P3GC - Help	☐ Scatter Plot
P123 versus GC - Help	☐ Scatter Plot ☐ Contour ☐ Heatmap
P3 versus all Amino Acids - Help	☐ P3 versus all 20 amino acids
CAI versus P3 - Help	☐ Scatter Plot ☐ Contour ☐ Heatmap

Custom Graphs:

Custom Scatter Plot - Help	☐ Generate Custom Scatter Plot
	P3 versus CAI (two g)
Histograms - Help	☐ CAI ☐ P3 ☐ Length ☐ Hydrophobicity
	☐ Generate Monte Carlo histogram for selected genes set(s) (Query Set (qs))
	Compositional Measure: CAI (two g)
Monte Carlo Histograms - Help	Reference Gene Set (rgs): Any gene in genome
	Sample Reference Set Genes as: Individual genes
	Number of replicates (n): 100 replicates

Fig. 10.7.
Selecting analyses in CodonExplorer.

Combined graphs for STM

Codon Table

4527 gene(s) (1429204 codon(s))

Fields: [triplet] [frequency: per thousand] ([number])

UUU 23.2 (33210)	UCU 7.2 (10281)	UAU 17.1 (24424)	UGU 4.8 (6886)
UUC 15.2 (21684)	UCC 10.1 (14442)	UAC 11.5 (16412)	UGC 6.6 (9481)
UUA 13.2 (18828)	UCA 6.2 (8905)	UAA 1.9 (2685)	UGA 1.0 (1412)
UUG 12.3 (17581)	UCG 9.5 (13566)	UAG 0.3 (433)	UGG 15.2 (21686)
CUU 11.8 (16828)	CCU 7.1 (10216)	CAU 13.2 (18928)	CGU 18.5 (26451)
CUC 10.4 (14901)	CCC 6.9 (9910)	CAC 9.5 (13624)	CGC 23.1 (33072)
CUA 4.9 (7000)	CCA 5.7 (8215)	CAA 12.7 (18132)	CGA 3.6 (5096)
CUG 53.4 (76262)	CCG 24.6 (35095)	CAG 30.9 (44132)	CGG 6.9 (9912)
AUU 29.3 (41863)	ACU 6.7 (9556)	AAU 17.8 (25426)	AGU 7.3 (10467)
AUC 24.3 (34721)	ACC 23.3 (33328)	AAC 20.1 (28667)	AGC 17.4 (24937)
AUA 5.3 (7582)	ACA 5.7 (8187)	AAA 34.8 (49731)	AGA 2.3 (3344)
AUG 27.3 (39072)	ACG 18.8 (26910)	AAG 11.2 (15996)	AGG 1.6 (2335)
GUU 15.4 (22053)	GCU 12.7 (18144)	GAU 31.6 (45185)	GGU 17.3 (24715)
GUC 18.2 (25970)	GCC 29.1 (41591)	GAC 20.2 (28927)	GGC 35.3 (50413)
GUA 11.3 (16134)	GCA 12.9 (18500)	GAA 35.0 (50085)	GGA 8.7 (12443)
GUG 25.2 (35972)	GCG 42.4 (60613)	GAG 20.6 (29478)	GGG 12.0 (17169)

Coding GC 53.98% 1st letter GC 62.31% 2nd letter GC 41.60% 3rd letter GC 58.03%

Fig. 10.8.

The codon usage table for all genes in the *Salmonella enterica* serovar Typhimurium LT2 genome, as output by CodonExplorer.

Fig. 10.9.
A Monte Carlo histogram in CodonExplorer.

Fig. 10.10.

Ribosomal proteins, and proteins from other orthology groups believed to be highly expressed, tend to have higher CAI values than most genes (C + D), and are outliers on a plot of CAI versus P₃ (A + B). In this case, the *Salmonella enterica* serovar Typhimurium genome is plotted.

Fig. 10.11.

The ribosomal gene coding for the small subunit protein rpsU has significantly high ($p = 0.00000897$) CAI values relative to other genes in *Salmonella enterica* serovar Typhimurium (A), and significantly lower ($p = 0.0111$) P3/CAI values (B).

Fig. 10.12.

Several views of CAI versus P3 available in CodonExplorer. The genomes chosen reflect low (mge), mid (eco), and high (sco) values of genomic GC content. GC content strongly influences CAI; the two are not independent. Thus, analysis of expression and HGT are connected (Guy and Sueoka, submitted). Heat maps and contour plots show trends that are hard to see in the scatter plot. Seeing which areas are denser allows easier interpretation (e.g., than graphing in Excel and being unable to relate the distribution of points and the regression line).

Fig. 10.13.
Fingerprint plot of all human genes, similar to that in (20).

Fig. 10.14. Monte Carlo histogram of CAI values for the *Salmonella enterica* serovar Typhimurium prophage Fels-1, and the Salmonella Pathogenicity island 2 (SPI-2)

Genes in both the STM Fels-1 prophage (A) and SPI-2 (B) show significantly ($p = 0.0223$ and 0.01 , respectively) lower average CAI values than the genome as a whole.

Table 10.1

Compositional properties calculated by CodonExplorer. Arbitrary combinations of these values may be plotted against one another using the “Custom scatter plot” option

Name	Equation/reference
CAI* – the codon adaptation index	$CAI = CAI_{obs} / CAI_{max}$ $CAI_{obs} = \left(\prod_{k=1}^L RSCU_k \right)^{(1/L)}$ $CAI_{max} = \left(\prod_{k=1}^L RSCU_{kmax} \right)^{(1/L)}$ $RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}}$
	(14,43)
P ₃ – The third position GC content	$P_3 = \frac{G_3 + C_3}{A_3 + T_3 + C_3 + G_3}$
P ₃ /CAI	P ₃ /CAI (<i>see</i> equations for P ₃ and CAI, above)
Length	Gene length
Hydrophobicity	Predicted peptide hydrophobicity (44)
Horizontal transfer index (HTI)	$Pr(COD_i F) = \frac{Pr(F COD_i) Pr(COD_i)}{\sum_{m=1}^6 Pr(F COD_m) Pr(COD_m) + Pr(F NON) Pr(NON)}$ <p>(<i>m</i> = 1,2,3,4,5,6) The probability that a gene was produced by a Markov model matching in-frame coding genes in a genome (rather than other frames of coding genes or non-coding sequence). Genes with low HTIs and low ribosomal HTIs (<i>see</i> below) have been proposed as transferred (30)</p>
Ribosomal HTI	The posterior probability that a gene matches the model for in-frame ribosomal proteins in a genome, rather than intergenic regions separating ribosomal genes or the non-coding frames of ribosomal proteins. Calculated as the HTI above, but using ribosomal proteins or intergenic regions separating ribosomal proteins to build the coding and non-coding models (30)
Putative alien	A putatively alien gene identified using the scheme of Nakamura et al. (30)
Amino acid frequencies	The frequency of individual amino acids, or a group of amino acids within a gene. Available options include A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y, (D+E), (K + R + H), and (L + I + V + M)

* The equation shown for CAI is the traditional equation (14); the other CAI options are variants of this measure that may be useful for specific analyses.