2010

Automated Isolation of Translational Efficiency Bias that Resists the Confounding Effect of GC(AT)-Content

Douglas W. Raiford  
*Wright State University - Main Campus*

Dan E. Krane  
*Wright State University - Main Campus, dan.krane@wright.edu*

Travis E. Doom  
*Wright State University - Main Campus, travis.doom@wright.edu*

Michael L. Raymer  
*Wright State University - Main Campus, michael.raymer@wright.edu*

Follow this and additional works at: [http://corescholar.libraries.wright.edu/knoesis](http://corescholar.libraries.wright.edu/knoesis)

Part of the [Bioinformatics Commons](http://corescholar.libraries.wright.edu/knoesis), [Communication Technology and New Media Commons](http://corescholar.libraries.wright.edu/knoesis), [Databases and Information Systems Commons](http://corescholar.libraries.wright.edu/knoesis), [OS and Networks Commons](http://corescholar.libraries.wright.edu/knoesis), and the [Science and Technology Studies Commons](http://corescholar.libraries.wright.edu/knoesis)

**Repository Citation**

[http://corescholar.libraries.wright.edu/knoesis/530](http://corescholar.libraries.wright.edu/knoesis/530)

---

This Article is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu.
Automated Isolation of Translational Efficiency Bias That Resists the Confounding Effect of GC(AT)-Content

Douglas W. Raiford, Dan E. Krane, Travis E. Doom, and Michael L. Raymer

Abstract—Genomic sequencing projects are an abundant source of information for biological studies ranging from the molecular to the ecological in scale; however, much of the information present may yet be hidden from casual analysis. One such information domain, trends in codon usage, can provide a wealth of information about an organism’s genes and their expression. Degeneracy in the genetic code allows more than one triplet codon to code for the same amino acid, and usage of these codons is often biased such that one or more of these synonymous codons are preferred. Detection of this bias is an important tool in the analysis of genomic data, particularly as a predictor of gene expressivity. Methods for identifying codon usage bias in genomic data that rely solely on genomic sequence data are susceptible to being confounded by the presence of several factors simultaneously influencing codon selection. Presented here is a new technique for removing the effects of one of the more common confounding factors, GC(AT)-content, and of visualizing the search-space for codon usage bias through the use of a solution landscape. This technique successfully isolates expressivity-related codon usage trends, using only genomic sequence information, where other techniques fail due to the presence of GC(AT)-content confounding influences.

Index Terms—Codon usage bias, GC-content, strand bias, translational efficiency.

1 INTRODUCTION

Most molecular evolutionary analyses presume that the principal factor upon which natural selection acts is the drive for specific, functional, and stable proteins. Thus, selective pressure driving genomic change acts, predominately, at the level of the translated amino acid. However, selection also acts at a much finer scale, driving the selection of individual nucleotides, even when the changes induced have no direct consequence to the protein product of a gene [1], [2]. Some of the forces that have been suggested include translational efficiency [3], [4], [5], [6], [7], and mutational forces that introduce biased GC-content [8], [9], [10] or biased strand content (replication induced increased G + T concentration on the leading strand) [10], [11]. Additionally, genes introduced into a genome through horizontal gene transfer (HGT) will retain the bias composition of the source organism until the process of amelioration causes the genes to conform to the target genome’s codon usage bias [12]. Analysis of codon usage can help determine which genes are candidates for having been horizontally transferred [13].

The primary selective advantage enjoyed by genes exhibiting a bias that promotes translational efficiency is imparted by the preferential usage of codons associated with more abundant tRNAs [5], [14]. Because highly expressed genes tend to adhere more closely to this bias [4], [6], adherence can be used as a predictor of expressivity. Use of these preferred codons also results in the realization of an additional benefit—an increase in the accuracy of translation [15].

In some cases, multiple biases can coexist within an organism’s genome [9], [10]. When this occurs, the translation-driven codon usage bias (if it exists) can be obscured, making gene expression levels difficult to predict. Several approaches have been employed to identify and measure codon usage biases [5], [8], [16], [17], [18], [19], [20]. Some methods, such as Sharp and Li’s codon adaptation index [19], require prior knowledge of a set of genes known to be highly expressed. Others, such as the one put forward by Carbone et al. [20] (a purely algorithmic method of determining codon adaptation index), attempt to identify the bias using coding sequence information only. In cases where the intent is to identify the bias associated with translational efficiency, algorithms that take the latter approach (using sequence information only) can be confounded by other biases that exist within the target genome (e.g., GC or strand bias) [10], [20], [21].

We present a new technique for visualizing the search-space for codon usage bias along with an automated procedure for isolating the translational efficiency bias that resists the confounding effects of GC(AT)-content bias. Following the description of how to generate the visualization of the search-space, we present a new technique for
removing the effects of GC(AT)-content (confounding bias) that can be implemented using genomic sequence information only. The efficacy of our method is demonstrated in Section 3 where we present the outcome of our approach to isolating translational efficiency bias in organisms previously shown to be confounded by highly skewed GC(AT)-content [10]. The technique is compared to traditional methods of calculating CAI by examining how they correlate with expression data as determined by microarray experimentation. Additionally, we provide new results on organisms previously shown to lack confounding factors [10] in order to demonstrate that the our approach is effective under nonconfounded conditions as well.

2 METHODS

2.1 Codon Usage Bias

The Sharp and Li codon adaptation index measurement [19] assigns a value known as the relative adaptiveness (or weight) to each codon. The weight for each codon is derived from the set of genes that is determined experimentally to be the most highly expressed genes of the genome. This is known as the reference set. The weight for each codon is determined by normalizing the count for that codon in the reference set by the count of its maximal sibling (codon with the maximum count within the same synonymous codon family). A geometric mean is taken of the weights associated with the codons in a gene to calculate its CAI.

Carbone et al. removed the need for prior knowledge of gene expressivity from the CAI calculation process [20]. Theirs is a two-step approach that works first to identify the reference set of genes (using a greedy algorithmic approach) and then calculates the CAI score for each gene based upon this reference set. Identification of the reference set is performed by assigning a precise mathematical definition to reference set membership and then searching for the genes that match this definition. Carbone et al. define a reference set as a small set of genes (~1 percent of genome) characterizing a bias to which its (the reference set’s) adherence is stronger than all other genes in the genome. The search algorithm (for the reference set) is iterative in nature, beginning with the entire genome as the reference set and iterating until a reference set of approximately 1 percent is achieved. To prevent confusion, when the term CAI is used it will be in reference to values derived using the traditional Sharp and Li approach. When the Carbone et al. method is utilized we will employ the terminology “self-consistent codon index (SCCI),” as described in [22]. Self-consistency refers to the definitional condition that the identified reference set adhere more strongly to the bias (which the reference set itself defines, hence self-consistent) than all other genes in the genome.

The SCCI algorithm described above is designed to isolate the dominant bias within a genome. The dominant bias is not necessarily driven by translational efficiency. A major contribution of this work is to provide a technique (modified SCCI—mSCCI) that specifically searches for translational efficiency bias while preserving the algorithm’s independence from the need for prior knowledge of a set of highly expressed genes. Before examining this resolution, we introduce two visualization techniques: one to demonstrate that multiple biases can coexist in a genome and that the SCCI algorithm finds the dominant one, and another technique for observing where in the codon usage space good solutions to the search for self-consistent reference sets can be found.

2.2 Principal Components Analysis (PCA)

Using Principal Components Analysis (PCA) [23], the 59D frequency (relative synonymous codon usage or RSCU [24]) vectors for the genes (64 codons less start, stop, and the two amino acids with only one codon) can be reduced to two dimensions. Fig. 1 illustrates how PCA can be used to depict the possible presence of confounding biases. The figure reveals that the dominant bias (in this case, AT-content) may overshadow translational efficiency bias. While the resulting 2D space shows the relationship among genes in terms of their codon usage, this 2D projection does not facilitate recognition of where good solutions to the search for reference sets reside. A more direct visualization of competing biases can be realized by extending this space into a third dimension that is based on the quality of reference sets in the specified region of codon usage. The resulting 3D space can be viewed as a reference set quality landscape, or RSQ landscape.

2.3 RSQ Landscape

In order to build this RSQ landscape, proposed reference sets representative of biases in different regions of the PCA-derived 2D codon usage plots are generated by gathering the nearest neighbors to each gene into discrete reference sets. This is accomplished by first determining the euclidean distance between the RSCU vector of each gene and that of every other gene. Once these distances are computed, a set of local neighbors is constructed for each gene. This set consists of the 1 percent of genes nearest to the gene in question based on the RSCU distance. The RSQ landscape is assembled by determining a quality score for each gene’s neighboring set and using this as a value in the third dimension, orthogonal to the 2D PCA-derived plane. A surface is then constructed using these quality scores.

To calculate the quality score for gene $i$, SCCI scores are first assigned to all genes where the weights are defined by the reference set comprised of the nearest neighbor genes (1 percent of genes in genome) to gene $i$. The genome is then sorted by this SCCI value, and the degree to which gene $i$’s reference set rises to the top of the sorted list is assessed ((1) through (3)). A self-consistent reference set is defined as a small set of genes (1 percent of genome) characterizing a bias to which its adherence is stronger than all other genes in the genome. Intuitively, the closer a proposed reference set’s behavior is to the definition of a reference set (how self-consistent it is), the higher its quality can be said to be. In (1), $RS$ is the reference set, $|RS|$ is the size of the reference set, $N$ is the number of genes within the genome, and $IDX$ is the index of the proposed reference set gene $i$ in the sorted list of all genes. An ideal reference set (one that matches, exactly, the mathematical definition of a self-consistent reference set) will rise to the top of this list and is represented by (2). This measure will assign a score (from 0 to 1) to a reference set. To put this in context, the Carbone et al. algorithm is designed
to locate reference sets with perfect quality scores. The score is used in the construction of the RSQ landscape, but it can also be viewed as a measure of the quality of a proposed reference set. This measure of quality is known as a fitness score in evolutionary computation techniques, and is used in the construction of fitness landscapes for the purpose of search-space analysis [25], [26], [27]. This approach is, in turn, based upon the work of biologists in the field of genetics and evolutionary theory [28]:

\[
f(RS) = \sum_{i=1}^{[RS]} IDX_i,
\]

\[
f_{\text{max}}([RS]) = \sum_{i=N-[RS]}^{N-1} i,
\]

\[
f_{\text{norm}}(RS) = \frac{f(RS)}{f_{\text{max}}([RS])}.
\]

The surface of the RSQ landscape is constructed by creating a regularly spaced grid of points within the PCA determined 2D plot space. A triangle-based linear interpolation method (based on a Delaunay triangulation using the quickhull algorithm for convex hulls) is used to determine an associated value at each grid point [29]. This value is an aggregation of the nearby gene quality scores. The surface is then rendered orthogonal to the 2D plot of RSCU data projected onto the first two principal components.

Fig. 2 depicts a landscape dominated by two regions of the codon usage space where high-quality reference sets can be found. Their corresponding proximity to the locations of the SCCI reference set (representative of the bias induced by high AT-content) and the ribosomal protein coding genes (RPCGs) is a strong indication that one ridge corresponds to high-quality AT-bias solutions and the other to high-quality translational efficiency bias solutions.

2.4 Determination of Whether SCCI is Confounded

Multiple selective and mutational forces act simultaneously to influence codon usage [9], [10]. Figs. 1 and 2 visually depict this phenomenon. In these situations, translational efficiency bias may or may not be the dominating trend. In the presence of multiple, coexisting biases, it is useful to be able to determine whether the bias discovered by the Carbone et al. algorithm is that of translational efficiency, or of some other, confounding bias.

2.4.1 Ribosomal Criterion

Carbone et al. have developed a number of useful measures for determining the nature of the dominating bias identified by their algorithm [10]. One measure of how well a trend identified by SCCI captures the translational efficiency bias for a particular organism is known as the ribosomal criterion. This criterion is based on the degree to which the RPCGs (which may be assumed to be highly expressed) are found in the upper region of a sorted list of genes (by SCCI value). As employed by Carbone et al., it can be concluded that their algorithm has identified the translational efficiency bias for an organism when the average SCCI value for the organism’s RPCGs is greater than one standard deviation above the mean SCCI value for the organism’s genome. The average of the CAI/SCCI scores for RPCGs in standard normal form is used to define the ribosomal criterion. A genome characterized by translational efficiency bias will have a high ribosomal criterion.

2.4.2 HEDB Criterion

A design requirement of the mSCCI algorithm presented here is that it be able to isolate translational efficiency bias using sequence information only. Traditional methods of identifying translational efficiency bias required both sequence information and prior knowledge of a set of highly expressed genes. RPCGs, which are known to be highly expressed, can be used for this purpose. Our approach will
also require a set of presumed highly expressed genes in order to guide the search for a self-consistent reference set indicative of translational efficiency bias (versus the dominant bias). We identify these genes via sequence similarity thereby making the algorithm fully automated. We employ a database of proteins associated with genes known to be highly expressed (high-expression database or HEDB) to help identify a set of genes that are likely to be highly expressed in the target genome. The proteins chosen for the database are ribosomal proteins, elongation factors, and RNA polymerase subunits (excluding the sigma subunit). These were chosen because they are known generally to be highly expressed, and they tend to exhibit high overall average CAI/SCCI scores in well-characterized genomes (extensively studied genomes whose dominant bias is known to be translational efficiency, data not shown).

For each gene in the target genome, an unfiltered blast search is performed [30] using the corresponding protein as a query against the HEDB. The query protein is considered to be a homolog of a protein in the database if they exhibit 60 percent identity. The 60 percent threshold was chosen empirically to balance false positives with false negatives. The database currently contains proteins from 66 organisms, none of which are used as target genomes in this study. In this way, independence between those genomes being analyzed and the HEDB is maintained. The organisms used to build the HEDB are drawn from 25 different bacterial taxonomical subclasses, or groups, in order to achieve a class-wide representative sampling.

Any genes considered homologous to a database-identified highly expressed gene are placed in a list of HEDB genes for the target genome. Similar to Carbone et al.’s ribosomal criterion, a standard normal average is calculated using CAI/SCCI scores for these genes and is referred to as the HEDB criterion. Due to the strong correlation that exists between HEDB and ribosomal criterion (Fig. 3), the same threshold is employed ($\bar{z} > 1$, or the average CAI/SCCI score for HEDB genes being one standard deviation above the genome’s average).

### 2.4.3 Content Criterion

Carbone et al. developed the content criterion [10] to determine whether the organism’s bias is influenced by...
characterized by GC(AT)-bias.

The technique described here (of using HEDB genes as a more representative of the most highly expressed genes).

In organisms exhibiting distinct biases (such as In the absence of actual expression data, an accepted method for isolating translational efficiency bias is through the use of a set of known highly expressed genes as a reference set and employing the Sharp and Li method for determining CAI scores. Since we have now identified the genes within the target genome that have homologs in the HEDB, these genes can be utilized for this purpose. The problem is that these genes are not necessarily the most highly expressed in the genome; they are simply among the more highly expressed genes. For this reason, the bias captured by their sequence data may not be the best representation for translational efficiency bias. Thus, we use our mSCCI algorithm to search for a reference set that is more representative of the most highly expressed genes. The technique described here (of using HEDB genes as a reference set) will be used as a benchmark for comparing the performance of the various bias isolation methods.

Local Search

In organisms exhibiting distinct biases (such as Nostoc sp. PCC 7120, Fig. 2), a local search could be employed to identify the translational efficiency bias. Carbone et al. employed such a methodology (a random search seeded with genes in a localized area of interest) [20], though not specifically for the purposes of isolating translational efficiency bias. Local search methods work well when distinct bias ridges are present. Unfortunately, the factors influencing codon usage trends do not always operate primarily on disjoint gene sets. When codon usage biases act in tandem on similar sets of genes, the ridges in the RSQ landscape can intersect, overlap, or merge. Because both biases occur in close proximity (in the codon usage space), the gene ranking (by SCCI) is close to that generated by translational efficiency bias alone, and if confounded, will often have ribosomal criterion close to (but not above) the threshold of 1. This is an especially challenging confounding condition. Streptomyces coelicolor A3(2) is an organism that exhibits a merged bias. It can be seen that the RPCGs are on the ascending slope of the ridge associated with the dominant bias (Fig. 4c). As expected, Streptomyces coelicolor A3(2) has a ribosomal criterion well above zero ($z = 0.463$). This implies that the dominant bias (GC) will cause a ranking of genes (by SCCI score) similar to that produced by a true translational efficiency bias, but dissimilar enough to be considered confounded.

Modified SCCI Algorithm

By making a minor modification to the SCCI algorithm, we can be assured that the bias identified is more likely to be that of translational efficiency. The modification of the SCCI algorithm is a direct result of the exploratory data analysis enabled by the bias visualization techniques described previously.

The SCCI algorithm is designed to find the dominant bias. If the dominant bias is GC-content, then the algorithm can be modified to give lower SCCI scores to genes whose GC$_3$-content deviates from balanced usage. This directs the reference set search away from high GC(AT)-content regions. To avoid confusion, these scores will be described as mSCCI scores. The modification allows the discovery of the presence of translational efficiency secondary bias (and does not inhibit the search when translational efficiency is the primary bias). Recall that the CAI/SCCI score for a gene is calculated as the geometric mean of the weights associated with each codon used in that gene. The mSCCI algorithm multiplies each codon-associated weight by a factor, $\beta$ (5), that is inversely proportional to the gene’s deviation from balanced GC$_3$-use. The result is a reduction in the mSCCI score for genes that do not exhibit balanced GC$_3$-content. The degree to which high/low GC-content genes should be penalized depends upon the interplay of competing biases in codon usage for an organism. The scaling constant $\alpha$ (4) is introduced to regulate the amount of penalty imposed by the $\beta$ factor, and thus on the mSCCI scores of high GC(AT)-content genes. The scaling constant is organism specific. In biological terms, $\alpha$ can be thought of as representing the degree to which the biases (translational efficiency bias and the confounding bias) work in concert ($\alpha = 0$), or are at odds with one another ($\alpha = 2$). Biases in close proximity work in tandem on a gene’s codon usage. The ranking of genes when sorted according to their adherence to the two biases will be similar. These organisms will require smaller values for $\alpha$. Biases that are far apart in the codon usage space are in opposition to each other. The gene orderings (when sorted by mSCCI and SCCI) will be opposite (or nearly so) for the two biases. These organisms will have large values for $\alpha$:  

$$ \beta = 1 - \alpha GC_3 (g) - 0.5, \quad (4) $$  

$$ u'_i = \beta w_i. \quad (5) $$  

The weight modification factor ($\beta$) has a range from 1 downward to a lower limit set by the scaling constant $\alpha$. The scaling constant can range from 0 (no adjustment or unmodified) to 2. In the case of $\alpha = 2$ $\beta$ goes to zero when a gene with 100 percent GC(AT)$_3$-bias is encountered. In these situations, the weights are set to 0.01, which is the same value used in the SCCI algorithm when a particular codon exhibits no codon usage in a reference set. A different $\beta$ value is generated for each gene while $\alpha$ remains constant for the entire genome.
The organism’s $\alpha$ is determined iteratively by running the mSCCI algorithm with various $\alpha$’s and employing a golden section search algorithm [31] to determine the $\alpha$ associated with the greatest achievable HEDB criterion (Fig. 5). By maximizing the HEDB criterion, the reference set search is directed away from high GC(AT)-content regions and toward the region of the highest translational efficiency bias. The $\alpha$’s that generate reference sets with quality scores (3) of less than 0.900 are discarded regardless of achieved HEDB criterion.

![Fig. 4](image)

Fig. 4. (a) *Streptomyces coelicolor* A3(2) is an example of a genome where GC-content constitutes the dominant bias. The translational efficiency bias ridge is merged or subsumed into and by the GC-content ridge. To help understand the surface of the RSQ landscape, a 2D slice that is horizontal with $Z1$ (slice made at $-3$ on $Z2$) is provided in (b). Evidence of this can be seen in (c) and (d) where RPCGs (white points) are shown to be on the ascending slope of the dominant ridge, below the identified reference set. The reference set is as defined by the SCCI algorithm (GC-bias). Local search methodologies have difficulty isolating such a bias due to localized interference from the GC-bias. (a) *Streptomyces coelicolor A3(2)* RSQ landscape. (b) Two-dimensional slice of RSQ landscape. (c) Location of RPCGs. (d) Location of reference set genes.

The modified weights are used only during the phase of the algorithm that searches for the reference set. Using the adjusted weights to calculate the final mSCCI scores for the genome may introduce unnecessary bias in the gene ranking. After locating the reference set using modified weights, the final gene ranking is produced using unadjusted mSCCI scores in the traditional manner. In the case of *Streptomyces coelicolor* A3(2), the HEDB criterion goes from 0.466 (SCCI algorithm) to 1.179 (mSCCI algorithm). The method of using HEDB genes as a reference set achieves a HEDB criterion of 1.089, while the traditional Sharp and Li technique (reference set of actual most highly expressed genes) is 0.713.

The mSCCI algorithm also works for organisms characterized by separate GC(AT) and translational efficiency bias ridges. When the mSCCI method is utilized for *Nostoc* sp. *PCC 7120*, an HEDB criterion of 1.418 is attained. The HEDB criterion is $-0.692$ using the SCCI algorithm. Additionally, the content criterion (correlation of SCCI with GC-content) is $-0.915$, indicating that the bias identified by the algorithm is that of high AT-content. The content criterion drops in magnitude to $0.095$ using the mSCCI algorithm indicating that the identified bias is no longer representative of AT-content.

The implementation used to calculate all versions of the SCCI algorithm is our own. We chose not to utilize available
tools (such as CAIJava [32]) due to the need for control over weight values on a gene-by-gene basis.

2.6 Microarray Expression Data
Perhaps the best way to evaluate CAI/SCCI/mSCCI calculation methods is by comparing their generated CAI/SCCI/mSCCI values to experimentally derived expression quantities. Protein expression data is not widely available; however, the extensive use of oligonucleotide microarrays has made mRNA abundance data readily accessible. With the exception of *Nostoc*, all expression data were retrieved from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus [33] (accession numbers GSE4896, GDS1469, GSE2983, GSE3876, GSE2667, GSE4617, GSE7070, GDS1099, and GSE2823). For *Nostoc*, expression data were retrieved from Wünschiers’ Hydrogen Database (HyDaBa) which focuses on gene-expression data from the filamentous nitrogen fixing cyanobacterium *Nostoc PCC 7120* [34]. In dual channel experiments, the results from the reference channel are utilized (no treatment or pre). For *Escherichia coli K12*, those trials using glucose as the carbon supply were used. When raw data were provided, background was subtracted from signal and user determined flags and thresholds were accepted. For preprocessed data, genes listed as absent are removed from consideration.

3 RESULTS

3.1 Modified SCCI Results
The Sharp and Li (using HEDB genes as a reference set), Carbone et al., and mSCCI techniques were applied to many of the same organisms studied in [10]. The organisms are grouped into three categories: 13 organisms confounded by GC-content (Table 1), 13 organisms confounded by AT-content (Table 2), and another 45 organisms believed to have no confounding biases (Table 3). All sequence data were obtained from the NCBI [35]. The mSCCI algorithm is able to isolate translational efficiency bias in 6 of the 13 GC-confounded organisms. Translational efficiency bias in one of the AT-confounded organisms can be isolated yielding a total of seven organisms that the mSCCI algorithm can disambiguate. All seven have exponential growth phases and are relatively easy to cultivate which are indicators of the probable existence of translational efficiency bias.

3.2 HEDB Genes as Reference Set
When the performance of mSCCI is compared to that of the Sharp and Li method (using HEDB genes as the reference set), very similar results are obtained (at least in terms of attained HEDB criteria) (Fig. 6). This being the case, it would hardly be worth the added computational complexity of calculating mSCCI (multiple iterations to identify the correct CAI) unless it outperformed the Sharp & Li approach with HEDB genes as a reference set. SCCI indicates SCCI values are determined using Carbone et al. mSCCI indicates SCCI values are determined using modified SCCI. All SCCI determined HEDB criteria for the organisms in this table are < 1 with content criteria > .7 indicating that the dominant bias is GC-Content (and not translational efficiency bias).

3.3 Comparison to Microarray Expression Data
We are particularly interested in organisms whose dominant bias is GC(AT)-content that we were able to demonstrate also are characterized by translational efficiency bias. It is believed that these organisms will provide the most opportunity for improvement in expression prediction using mSCCI. We were able to locate data for four such organisms, plus one additional organism (*Halobacterium sp. NRC-1*) that showed some translational efficiency bias, though not the requisite HEDB criterion of one standard deviation above the average SCCI for the organism (Table 4). *Halobacterium* has an HEDB Criterion of 0.430 (when using mSCCI, Table 1). Data for an equal number of organisms whose dominant bias is translational efficiency are included
In the confounded-organism analysis, all instances of mSCCI exhibit Spearman rank correlation coefficients ($r_S$) between mSCCI values and expression levels that are more positive than that achieved by the Sharp and Li method. Using a signs test, this is enough to infer that mSCCI generates solutions that are more correlated with microarray data ($p = 0.031$). Of the five confounded organisms, three (Nostoc, Halobacterium, and P. aeruginosa) exhibit mSCCI $r_S$ values that are significantly more positive ($p < 0.05$) than those attained using the Sharp and Li method (using HEDB genes as a reference set). Significance was determined using a two-tailed Fisher z-transform [36] using $1.06$ in the numerator of the variance calculation due to $r_S$ being nonparametric [37]. One of the organisms (Nostoc) even exhibits a Spearman rank correlation that is significantly more positive than the traditional Sharp and Li method (where the reference set is the set of genes whose actual expressivity is the greatest for the organism). The Carbone et al. method identifies the dominant bias, and the dominant bias for these organisms is known to be that of GC(AT)-content, ensuring that all the Carbone et al. correlations will be weak or negative.

None of the five organisms whose dominant bias is translational efficiency (Table 5) have statistically different $r_S$ values between any of the four tested methods for determining CAI/SCCI scores.

### 3.4 Reference Set Quality

The quality score (3) can give insights into just how well a reference set captures the underlying codon usage bias. Fig. 7 shows that the quality scores attained when using mSCCI on confounded organisms tend to be higher than that of the HEDB reference sets when using the traditional approach.

#### Table 2

<table>
<thead>
<tr>
<th>Organism Name</th>
<th>CAI&lt;sub&gt;HEDB&lt;/sub&gt;</th>
<th>HEDB</th>
<th>Content</th>
<th>SCCI</th>
<th>HEDB</th>
<th>Content</th>
<th>mSCCI</th>
<th>HEDB</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buchnera aphidicola str. Sq</td>
<td>-0.013</td>
<td>-0.756</td>
<td></td>
<td>-0.554</td>
<td>-0.762</td>
<td></td>
<td>-0.182</td>
<td>-0.517</td>
<td></td>
</tr>
<tr>
<td>Campylobacter jejuni NCTC 11618</td>
<td>0.838</td>
<td>-0.714</td>
<td></td>
<td>-0.094</td>
<td>-0.786</td>
<td></td>
<td>-0.066</td>
<td>-0.778</td>
<td></td>
</tr>
<tr>
<td>Fusobacterium nucleatum ATCC 25586</td>
<td>0.933</td>
<td>-0.739</td>
<td></td>
<td>0.500</td>
<td>-0.781</td>
<td></td>
<td>0.500</td>
<td>-0.781</td>
<td></td>
</tr>
<tr>
<td>Leptospira interrogans L1-130 chromosome I</td>
<td>0.627</td>
<td>-0.302</td>
<td></td>
<td>-0.941</td>
<td>-0.877</td>
<td></td>
<td>0.937</td>
<td>-0.073</td>
<td></td>
</tr>
<tr>
<td>Mycoplasma genitalium G-37</td>
<td>-0.596</td>
<td>-0.808</td>
<td></td>
<td>-1.056</td>
<td>-0.937</td>
<td></td>
<td>0.475</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>Mycoplasma pneumoniae M129</td>
<td>0.895</td>
<td>0.135</td>
<td></td>
<td>-0.246</td>
<td>-0.887</td>
<td></td>
<td>0.760</td>
<td>0.325</td>
<td></td>
</tr>
<tr>
<td>Mycoplasma pulmonis UAB CTIP</td>
<td>0.939</td>
<td>-0.650</td>
<td></td>
<td>0.573</td>
<td>-0.778</td>
<td></td>
<td>0.573</td>
<td>-0.778</td>
<td></td>
</tr>
<tr>
<td>Nostoc sp. PCC 7120</td>
<td>0.713</td>
<td>-0.232</td>
<td></td>
<td>-0.692</td>
<td>-0.915</td>
<td></td>
<td>1.418</td>
<td>0.095</td>
<td></td>
</tr>
<tr>
<td>Rickettsia conorii str. Malish 7</td>
<td>0.829</td>
<td>-0.874</td>
<td></td>
<td>0.612</td>
<td>-0.866</td>
<td></td>
<td>0.948</td>
<td>-0.787</td>
<td></td>
</tr>
<tr>
<td>Sulfolobus solfataricus P2</td>
<td>0.971</td>
<td>-0.872</td>
<td></td>
<td>0.917</td>
<td>-0.877</td>
<td></td>
<td>0.972</td>
<td>-0.734</td>
<td></td>
</tr>
<tr>
<td>Sulfolobus tokodaii str. 7</td>
<td>0.858</td>
<td>-0.914</td>
<td></td>
<td>0.938</td>
<td>-0.883</td>
<td></td>
<td>0.948</td>
<td>-0.882</td>
<td></td>
</tr>
<tr>
<td>Thermoanaerobacter tengcongensis MB4</td>
<td>-0.107</td>
<td>-0.417</td>
<td></td>
<td>-0.562</td>
<td>-0.853</td>
<td></td>
<td>-0.004</td>
<td>-0.001</td>
<td></td>
</tr>
<tr>
<td>Thermoplasma volcanoan GSS1</td>
<td>1.042</td>
<td>-0.070</td>
<td></td>
<td>0.472</td>
<td>-0.796</td>
<td></td>
<td>0.472</td>
<td>-0.796</td>
<td></td>
</tr>
</tbody>
</table>

HEDB Criterion: values > 1 indicate that the organism is characterized by translational efficiency bias. Content Criterion: values > 0.7 indicate that the organism is characterized by GC-content bias (< -0.7 for AT-content). CAI<sub>HEDB</sub> indicates CAI values are determined using Sharp & Li approach with HEDB genes as a reference set. SCCI indicates SCCI values are determined using Carbone et al. SCCI. mSCCI indicates SCCI values are determined using modified SCCI. All SCCI determined HEDB criteria for the organisms in this table are < 1 with content criteria < -1 indicating that the dominant bias is AT-Content (and not translational efficiency bias).

#### 3.5 Distance between Biases and $\alpha$

In Section 2, $\alpha$ (4) was described as a measure of the degree to which the two biases work in concert, or in opposition, to one another. A simple method for verifying this concept is to examine the relationship between $\alpha$ and the euclidean distance between the reference sets representing the two biases. The distance between reference sets is calculated between the center points of the two reference-set clouds, in the 59-dimension codon usage space (RSCU). One would expect biases in close proximity to work in concert and to exhibit low $\alpha$'s. Alternatively, those biases that are far apart should result in opposing gene orderings (when the genes are ranked according to adherence to each bias) and larger $\alpha$ values. Fig. 8 illustrates that there is a positive correlation between $\alpha$ and bias distance ($r^2 = 0.48$, $p < 0.05$), supporting the given biological interpretation of $\alpha$. The biases being examined are those identified by the SCCI algorithm (the dominant bias) versus the bias found by mSCCI (translational efficiency bias).

#### 3.6 RSQ Landscape after Adjusting for GC-Content Bias

Fig. 9 depicts a side-by-side comparison of the RSQ landscape for Nostoc sp. PCC 7120 when both the SCCI algorithm and the mSCCI algorithm are utilized in their generation. Note that the region that previously dominated the landscape has a comparatively reduced elevation. This allows the discovery of the now dominant translational efficiency ridge by the mSCCI algorithm.

#### 4 Discussion

The presence of competing biases (such as GC-content) can make the discovery of translational efficiency bias problematic for automated algorithms (those using sequence information only). Previous work has indicated that multiple biases can coexist in a genome [9], [10]. With the use of...
our RSQ landscape and mSCCI techniques, we have shown how to observe and, in many cases, remove the effects of GC(AT)-content bias from the translational efficiency bias discovery process, using sequence information alone. While our results demonstrate a disambiguation technique for genomes confounded by GC(AT)-content usage trends, these methods should be equally applicable to any other well-characterized confounding bias. It would simply be a matter of determining an appropriate method for establishing β (4) that sets the amount of penalty to impose on genes that adhere more strongly to this confounding bias. As GC(AT)-content appears to be the most common example of a confounding bias (40 organisms studied in [10] had a dominant bias that was something other than translational
efficiency, 35 of these were dominant for GC(AT)-content), it made sense to begin with the removal of this GC(AT)-content as a confounding factor.

The mSCCI algorithm is able to correct for a dominating GC- or AT-bias in several (7 of 26) tested organisms (Tables 1 and 2). The organisms for which this is true exhibit a dominant GC- or AT-content bias and an HEDB criterion less than one as determined by the SCCI algorithm. For these organisms, the mSCCI algorithm identifies a reference set and subsequent gene ranking (by mSCCI) that is more highly correlated with experimentally determined expression data than the traditional Sharp and Li technique (Table 4) (though it should be noted that the Sharp and Li approach was never intended to be utilized on genomes that do not exhibit clear translational bias). When an organism’s dominant bias is that associated with translational efficiency (and without the confounding effects of GC-content), the mSCCI algorithm achieves results that are comparable with [not significantly different from \((p > 0.05)\)] that of the Sharp and Li technique (Table 5), making the mSCCI algorithm a reasonable choice in all situations. The confounded organisms that the traditional Sharp and Li approach are able to isolate tend to be the same that the mSCCI algorithm can isolate (five organisms isolated by both; seven total organisms isolated by mSCCI; six total organisms isolated by Sharp and Li technique).

The results for two of the organisms warrant closer examination: *P. aeruginosa* and *Nostoc*. For these two genomes, the traditional Sharp and Li approach to isolating translational efficiency bias yields HEDB criteria of less than one (0.523 and 0.713, respectively), while mSCCI yields HEDB criteria of greater than one (1.337 and 1.418, respectively) (Tables 1 and 2). Both of these organisms have been shown to grow competitively (*P. aeruginosa* [9] and *Nostoc* [38]), leading to the expectation that they will exhibit translational efficiency bias. Care should be taken in drawing conclusions in cases where mSCCI uncovers

---

### TABLE 4

<table>
<thead>
<tr>
<th>Organism</th>
<th>CAI&lt;sub&gt;SKL&lt;/sub&gt;</th>
<th>CAI&lt;sub&gt;HEDB&lt;/sub&gt;</th>
<th>mSCCI</th>
<th>SCCI</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Caulobacter crescentus</em> CB15</td>
<td>0.280</td>
<td>0.306</td>
<td>0.311</td>
<td>0.284</td>
</tr>
<tr>
<td><em>Halobacterium sp. NRC-1</em></td>
<td>0.165</td>
<td>0.086</td>
<td>0.199†</td>
<td>0.151</td>
</tr>
<tr>
<td><em>Nostoc sp. PCC 7120</em></td>
<td>0.046</td>
<td>0.150</td>
<td>0.275‡</td>
<td>-0.69</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em> PA01</td>
<td>0.347</td>
<td>0.303</td>
<td>0.381§</td>
<td>0.204</td>
</tr>
<tr>
<td><em>Streptomyces coelicolor</em> A3(2)</td>
<td>0.131</td>
<td>0.148</td>
<td>0.151</td>
<td>0.130</td>
</tr>
</tbody>
</table>

All Spearman rank correlation coefficients are significant \((p < 0.05)\). All \(r_S\) values for modified SCCI are more positive than the \(r_S\) values for any of the other three techniques. CAI<sub>HEDB</sub> indicates CAI values are determined using HEDB genes as a reference set. CAI<sub>SKL</sub> indicates CAI values are determined using most highly expressed genes (using microarray expression data) as a reference set. mSCCI indicates SCCI values are determined using modified SCCI. SCCI indicates SCCI values are determined using Carbone et al. SCCI. † mSCCI \(r_S\) value significantly more positive than that of CAI<sub>HEDB</sub>. ‡ mSCCI \(r_S\) value significantly more positive than that of CAI<sub>SKL</sub>. Significance between \(r_S\) values determined using a two-tailed Fisher z-transform [36] with 1.06 in the numerator of the variance calculation due to \(r_S\) being non-parametric [37].

### TABLE 5

<table>
<thead>
<tr>
<th>Organism</th>
<th>CAI&lt;sub&gt;SKL&lt;/sub&gt;</th>
<th>CAI&lt;sub&gt;HEDB&lt;/sub&gt;</th>
<th>mSCCI</th>
<th>SCCI</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Bacillus subtilis</em> subsp. <em>subtilis</em> str. 168</td>
<td>0.159</td>
<td>0.121</td>
<td>0.135</td>
<td>0.131</td>
</tr>
<tr>
<td><em>Shewanella oneidensis</em> MR-1</td>
<td>0.286</td>
<td>0.323</td>
<td>0.296</td>
<td>0.295</td>
</tr>
<tr>
<td><em>Chlamyphila pneumoniae</em> AR39</td>
<td>0.230</td>
<td>0.228</td>
<td>0.213</td>
<td>0.213</td>
</tr>
<tr>
<td><em>Escherichia coli</em> K12</td>
<td>0.501</td>
<td>0.496</td>
<td>0.506</td>
<td>0.505</td>
</tr>
<tr>
<td><em>Lactococcus lactis</em> IL1403</td>
<td>0.403</td>
<td>0.417</td>
<td>0.417</td>
<td>0.414</td>
</tr>
</tbody>
</table>

All Spearman rank correlation coefficients are significant \((p < 0.05)\). CAI<sub>HEDB</sub> indicates CAI values are determined using HEDB genes as a reference set. CAI<sub>SKL</sub> indicates CAI values are determined using most highly expressed genes (using microarray expression data) as a reference set. mSCCI indicates SCCI values are determined using modified SCCI. SCCI indicates SCCI values are determined using Carbone et al. SCCI.
translational efficiency bias when traditional methods do not (HEDB criterion > 1 using mSCCI; ≤ 1 using traditional approach). One should verify that the organism in question would be expected to exhibit translational efficiency bias (i.e., is a fast grower and easily cultivatable) or compare the adherence scores to experimentally determined expression data.

The mSCCI algorithm works by repeatedly searching for reference sets using various values of α until a reference set with the best HEDB criterion is achieved. A useful exercise is to consider why this approach can outperform (on confounded organisms) the more direct Sharp and Li method of simply utilizing these putative highly expressed genes as a reference set (under the assumption that they represent the genes that are composed of, predominantly, the most translationally efficient codons). One explanation could center on just how well the reference set captures the underlying translational efficiency bias. When CAI values are calculated in the traditional way (Sharp and Li using the most highly expressed genes as a reference set), genes are found with CAI values greater than the lowest reference set CAI score that are not part of the reference set. As an example, when the most highly expressed genes for E. coli are used as a reference set, there are 1,073 nonreference set genes with better CAI scores than the lowest reference set CAI score. It may be that, particularly in these confounded organisms, the mSCCI algorithm allows for the discovery of a “better” reference set—a reference set that is more self-consistent and that better represents the expected high placement of the HEDB genes in the overall SCCI gene ranking. This could be allowing the discovery of weight values that more accurately reflect the adaptiveness of the codons. The quality score (3) of a reference set (degree to which the reference set rises to the top of a sorted list of genes) is a direct measurement of its self-consistency characteristic, and Fig. 7 shows that the reference sets identified by mSCCI have generally higher quality scores than the Sharp and Li approach when HEDB genes are used as the reference set (in organisms with GC(AT)-content as the dominant bias).

A good topic for further study is how to improve the quality score of the translational efficiency reference set even further. The mSCCI algorithm brings the search into the general neighborhood of translational efficiency bias, but until a quality score of 1 is achieved, the perfectly self-consistent reference set has not been discovered. This can have a significant impact on the results. Our investigations have shown improvements of as little as 0.07 in RSQ yield improvements in ribosomal criteria of more than a full standard deviation (data not shown). Additionally, the HEDB should be expanded to include proteins from all sequenced prokaryotic organisms.

Others have investigated the effects of GC-content on expected bias scores [8], [39], [40], though these analyses did not explore whether measuring deviation from such expected values could enhance prediction of translational efficiency bias. The study by dos Reis et al. [39] used a correlation between their transfer RNA adaptation index and deviation from expected effective number of codons to make a determination of whether the underlying bias was that of translational efficiency bias. And Puigbò et al. used a ratio of observed CAI to their expected CAI (eCAI) to determine similarity with a predetermined reference set, but did not specifically examine whether such a ratio would improve expression level prediction.

Our novel approach to visualizing the RSQ landscape is useful in gaining insights into the bias composition of an organism’s genome. While it is computationally intensive to generate the topography, it can be constructive when interpreting results. The RSQ landscape results shown here tend to demonstrate the effectiveness of the mSCCI algorithm. When a landscape is generated using the mSCCI algorithm, the ridges associated with the GC(AT)-content bias are diminished to a state where they no longer dominate the RSQ landscape (Fig. 9). The mSCCI algorithm

---

**Fig. 7. Distribution of quality scores for the confounded organism reference sets when the traditional (Sharp and Li) approach is utilized (with HEDB genes as the reference set) and when the mSCCI technique is employed.**

**Fig. 8. Relationship between α (degree to which two biases work in concert) and distance between biases ($r^2 = 0.48, p < 0.05$). Organisms that have confounding biases that are near the translational efficiency bias exhibit low α’s while those that are far apart have large α’s. Euclidean distance between center points of reference sets is used to represent distance between biases. The biases being examined are those identified by the SCCI algorithm (the dominant bias) versus the bias found by mSCCI (translational efficiency bias).**
isolates translational efficiency without requiring a priori knowledge of the set of the most highly expressed genes. Translational efficiency bias is not necessarily present in all organisms; however, in those genomes where it is present, mSCCI can be used to automatically isolate translational efficiency bias regardless of which bias is dominant (GC(AT)-content or translational efficiency). The results shown here indicate that it is particularly appropriate to use mSCCI when GC(AT)-content is a confounding influence on the search for translational efficiency bias (Table 4).

REFERENCES
