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

Repository Citation

Raiford, D. W., Krane, D. E., Doom, T. E., & Raymer, M. L. (2010). Automated Isolation of Translational Efficiency Bias that Resists the Confounding Effect of GC(AT)-Content. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7* (2), 238-250.
https://corescholar.libraries.wright.edu/knoesis/530

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, predominantly, 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 $\mathrm{G}+\mathrm{T}$ 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

- *D.W. Raiford is with the Department of Computer Science, University of Montana, Missoula, MT 59812. E-mail: douglas.raiford@umontana.edu.*
- *D.E. Krane is with the Department of Biological Sciences, Wright State University, Biological Sciences Bldg. 128, Main Campus, 3640 Colonel Glenn Hwy, Dayton, OH 45435. E-mail: dan.krane@wright.edu.*
- *T.E. Doom is with the Department of Computer Science and Engineering, College of Engineering and Computer Science, Wright State University, Russ Engineering Center 331, Main Campus, 3640 Colonel Glenn Hwy, Dayton, OH 45435. E-mail: travis.doom@wright.edu.*
- *M.L. Raymer is with the Department of Computer Science and Engineering, Wright State University, Joshi Research Center, 3640 Colonel Glenn Hwy, Dayton, OH 45435. E-mail: michael.raymer@wright.edu.*

Manuscript received 20 Dec. 2007; revised 23 May 2008; accepted 12 June 2008; published online 19 June 2008.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2007-12-0179.
Digital Object Identifier no. 10.1109/TCBB.2008.65.

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 ($\sim$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

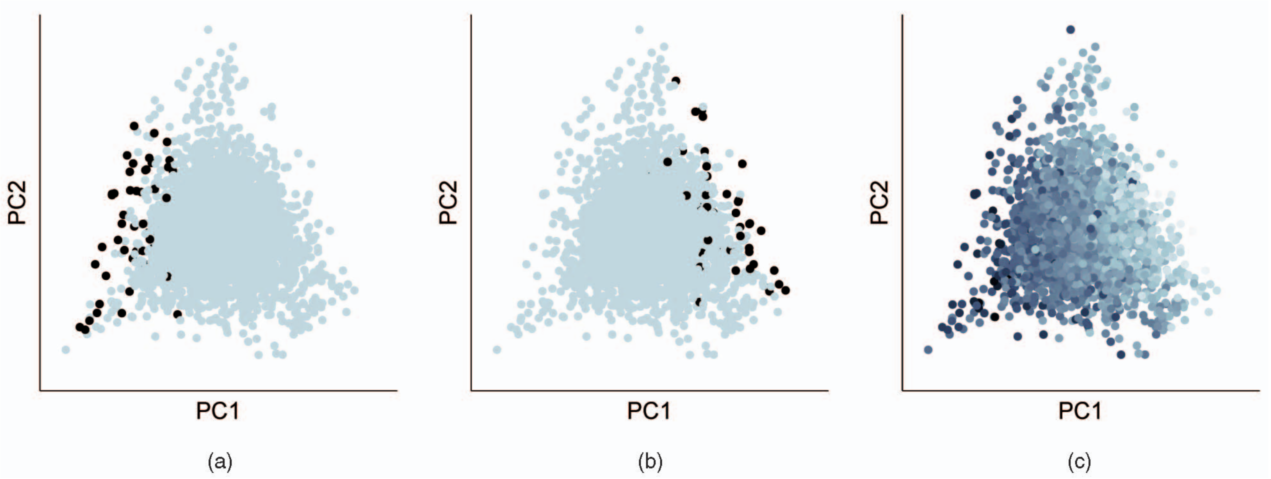


Fig. 1. PCA plots for *Nostoc sp. PCC 7120* genes. (a) SCCI reference set representative of the dominant bias. Each point represents one of the organism's 5,366 genes projected into the vector space identified by the first two principal components (PCA derived). The dark genes comprise the reference set. (b) RPCGs. These are genes known, generally, to be highly expressed. RPCGs are distant from the region identified by the reference set (shown in panel a). This indicates that the dominant bias (strong and consistent use of specific codons) identified by the SCCI algorithm is not representative of translational efficiency bias. (c) Genes shaded by GC-content. Dark genes have lower GC-content (high AT-content). The high AT-content genes are in the same region identified by the algorithm. This indicates that the Carbone et al. algorithm identified AT-content bias. PC1 captures 8.39 percent of the overall variance; PC2 6.22 percent.

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]:

$$f(RS) = \sum_{i=1}^{|RS|} IDX_i, \tag{1}$$

$$f_{max}(|RS|) = \sum_{i=N-|RS|}^{N-1} i, \tag{2}$$

$$f_{norm}(RS) = \frac{f(RS)}{f_{max}(|RS|)}. \tag{3}$$

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

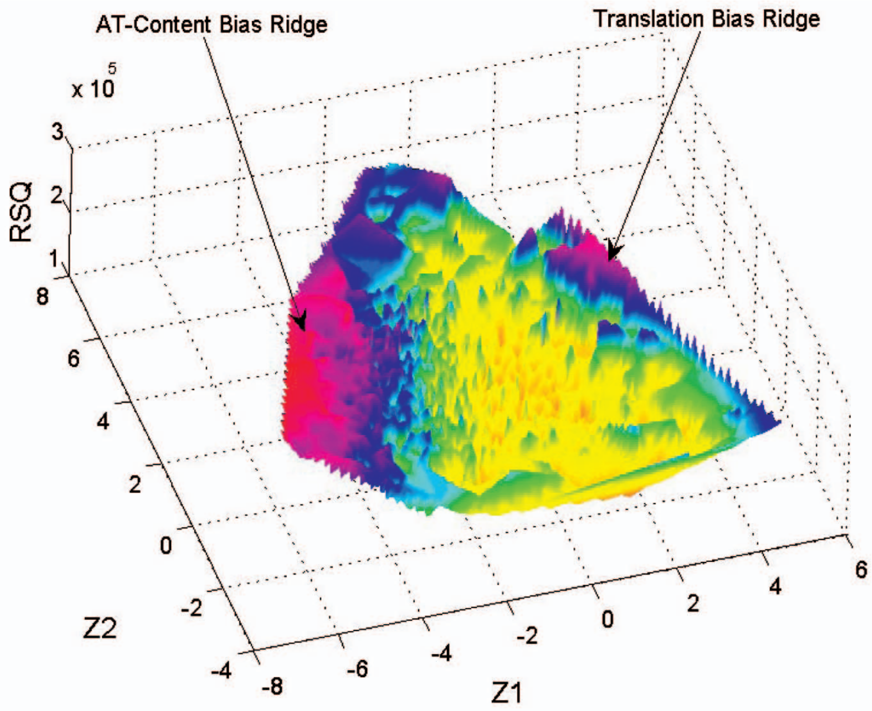


Fig. 2. *Nostoc sp. PCC 7120* RSQ landscape. Surface represents quality scores of reference sets in close proximity to the genes in PCA-derived 2D space (Z_1 and Z_2) depicted along an orthogonal axis (vertical axis). Elevated regions of the RSQ landscape can be thought of as regions from which high-quality reference sets can be drawn.

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

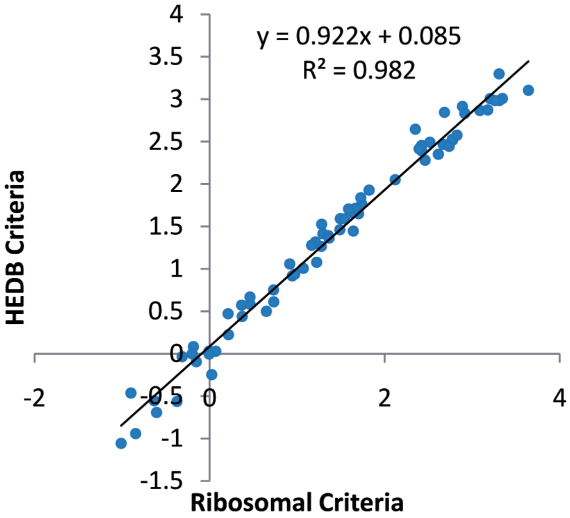


Fig. 3. Relationship between HEDB criteria and ribosomal criteria. The criteria are so closely related that the same threshold is used to determine whether organism is characterized by translational efficiency bias: HEDB criterion $>$ 1.0.

GC-content [10]. Detection is accomplished by measuring the correlation between the GC_3-content (percent of nucleotides in the third codon position that are G or C) of each gene with its CAI/SCCI value. Correlations > 0.7 (< -0.7 for AT-content) are an indication of a genome characterized by GC(AT)-bias.

2.5 Removal of Confounding Effects

While the visualization techniques (PCA and RSQ landscape) can identify situations where confounding factors cause the Carbone et al. algorithm to recognize some bias other than translational efficiency bias, they do not directly address the problem of isolating this bias when such confounding factors exist. We present here a set of techniques for algorithmically isolating translational efficiency bias even in the presence of other confounding biases.

2.5.1 Using A Priori Knowledge of Highly Expressed Genes

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.

2.5.2 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 ($\bar{z} = 0.463$). 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.

2.5.3 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, β (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 α (4) is introduced to regulate the amount of penalty imposed by the β factor, and thus on the mSCCI scores of high GC(AT)-content genes. The scaling constant is organism specific. In biological terms, α 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 α. 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 α:

$$\beta = 1 - \alpha|GC_3(g) - 0.5|, \tag{4}$$

$$w_i' = \beta * w_i. \tag{5}$$

The weight modification factor (β) has a range from 1 downward to a lower limit set by the scaling constant α. The scaling constant can range from 0 (no adjustment or unmodified) to 2. In the case of $\alpha = 2$ β 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 β value is generated for each gene while α remains constant for the entire genome.

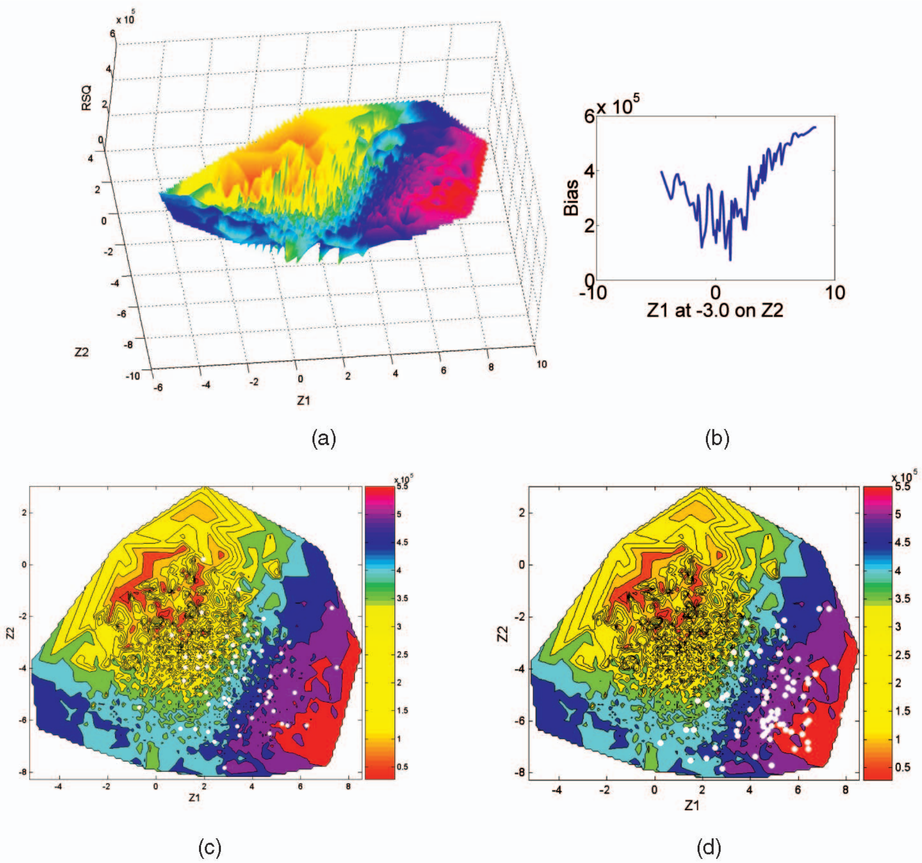


Fig. 4. (a) *Streptomyces coelicolor A3(2)* is an example of a genome where GC-content constitutes the dominant bias. The translational efficiency bias ridge is merged or subsumed into and by the GC-content ridge. To help understand the surface of the RSQ landscape, a 2D slice that is horizontal with 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 organism's α is determined iteratively by running the mSCCI algorithm with various α's and employing a golden section search algorithm [31] to determine the α 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 α's that generate reference sets with quality scores (3) of less than 0.900 are discarded regardless of achieved HEDB criterion.

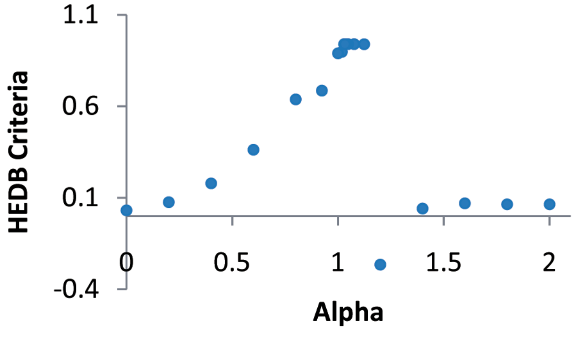


Fig. 5. *Chlorobium tepidum TLS* HEDB criteria at various alpha values. Additional data points are shown near the apex. This is the result of the golden section search.

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

TABLE 1
Criteria for Sharp and Li, Carbone et al., and mSCCI Runs on GC-Confounded Organisms

	CAI_{HEDB}		SCCI		mSCCI	
Organism Name	**HEDB**	**Content**	**HEDB**	**Content**	**HEDB**	**Content**
Aeropyrum pernix K1	0.865	0.810	0.730	0.898	0.931	0.795
Caulobacter crescentus CB15	1.381	0.825	0.915	0.839	1.505	0.798
Chlorobium tepidum TLS	0.478	0.896	-0.006	0.937	0.940	0.734
Halobacterium sp. NRC-1	0.141	0.937	0.215	0.825	0.430	0.702
Methanopyrus kandleri AV19	0.953	0.930	0.372	0.959	0.992	0.912
Neisseria meningitidis MC58	1.489	0.508	-0.185	0.750	1.568	0.516
Neisseria meningitidis Z2491	1.498	0.481	-0.314	0.733	1.519	0.474
Pseudomonas aeruginosa PAO1	0.523	0.771	-0.899	0.844	1.337	0.605
Pyrobaculum aerophilum str. IM2	0.111	0.501	0.070	0.900	0.239	0.059
Ralstonia solanacearum GMI1000	0.667	0.851	-0.199	0.887	0.923	0.790
Streptomyces coelicolor A3(2)	1.089	0.857	0.466	0.873	1.179	0.838
Thermoplasma acidophilum	0.600	0.717	-0.005	0.880	0.344	0.701
Xanthomonas campestris str. ATCC 33913	1.015	0.858	0.461	0.874	1.305	0.790

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_{HEDB} 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 $> .7$ indicating that the dominant bias is GC-Content (and not translational efficiency bias).

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 α) unless it outperformed the Sharp and Li version in some way. To determine whether this is the case, predicted expression orderings are compared to expression data from microarray experiments.

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

TABLE 2
Criteria for Sharp and Li, Carbone et al., and mSCCI Runs on AT-Confounded Organisms

	CAI_{HEDB}		SCCI		mSCCI	
Organism Name	**HEDB**	**Content**	**HEDB**	**Content**	**HEDB**	**Content**
Buchnera aphidicola str. Sg	-0.013	-0.756	-0.554	-0.762	-0.182	-0.517
Campylobacter jejuni NCTC 11168	0.838	-0.714	-0.094	-0.786	-0.066	-0.778
Fusobacterium nucleatum ATCC 25586	0.933	-0.739	0.500	-0.781	0.500	-0.781
Leptospira interrogans L1-130 chromosome I	0.627	-0.302	-0.941	-0.877	0.937	-0.073
Mycoplasma genitalium G-37	-0.596	-0.808	-1.056	-0.937	0.475	0.011
Mycoplasma pneumoniae M129	0.895	0.135	-0.246	-0.887	0.760	0.325
Mycoplasma pulmonis UAB CTIP	0.939	-0.650	0.573	-0.778	0.573	-0.778
Nostoc sp. PCC 7120	0.713	-0.232	-0.692	-0.915	1.418	0.095
Rickettsia conorii str. Malish 7	0.829	-0.874	0.612	-0.866	0.948	-0.787
Sulfolobus solfataricus P2	0.971	-0.872	0.917	-0.877	0.972	-0.734
Sulfolobus tokodaii str. 7	0.858	-0.914	0.938	-0.883	0.948	-0.882
Thermoanaerobacter tengcongensis MB4	-0.107	-0.417	-0.562	-0.853	-0.004	-0.001
Thermoplasma volcanium GSS1	1.042	-0.070	0.472	-0.796	0.472	-0.796

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_{HEDB} 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 $< -.7$ indicating that the dominant bias is AT-Content (and not translational efficiency bias).

(Table 5) in order to show mSCCI's performance on nonconfounded organisms.

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.

3.5 Distance between Biases and α

In Section 2, α (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 α 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 α'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 α values. Fig. 8 illustrates that there is a positive correlation between α and bias distance ($r^2 = 0.48,\ p < 0.05$), supporting the given biological interpretation of α. 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

TABLE 3
Criteria for Sharp and Li, Carbone et al., and mSCCI Runs on Unconfounded Organisms

	CAI_{HEDB}		SCCI		mSCCI	
Organism Name	**HEDB**	**Content**	**HEDB**	**Content**	**HEDB**	**Content**
Agrobacterium tumefaciens str. C58	1.545	0.738	1.650	0.700	1.879	0.657
Archaeoglobus fulgidus DSM 4304	1.059	0.768	1.266	0.771	1.232	0.763
Aquifex aeolicus VF5	1.513	0.574	1.526	0.491	1.540	0.483
Bacillus halodurans C-125	2.630	-0.366	2.915	-0.431	2.972	-0.411
Bacillus subtilis subsp. subtilis str. 168	2.992	-0.504	3.298	-0.460	3.304	-0.446
Bifidobacterium longum NCC2705	1.347	0.851	1.415	0.823	1.606	0.765
Brucella melitensis 16M chromosome I	1.597	0.726	1.592	0.716	1.746	0.676
Brucella suis 1330 chromosome I	1.691	0.751	1.764	0.732	1.843	0.702
Chlamydia muridarum Nigg	1.069	-0.510	1.078	-0.512	1.339	-0.307
Chlamydophila pneumoniae AR39	1.320	-0.775	1.392	-0.798	1.392	-0.798
Chlamydophila pneumoniae J138	1.498	-0.758	1.661	-0.757	1.695	-0.736
Clostridium acetobutylicum ATCC 824	1.453	-0.728	1.715	-0.676	1.699	-0.678
Clostridium perfringens str. 13	1.356	-0.343	1.706	-0.250	2.050	-0.176
Corynebacterium glutamicum ATCC 13032	2.462	0.499	2.646	0.456	2.653	0.454
Deinococcus radiodurans R1 chromosome 1	1.386	0.811	1.590	0.746	1.676	0.727
Escherichia coli K12	1.974	0.427	2.469	0.323	2.498	0.315
Escherichia coli O157:H7	2.046	0.476	2.518	0.405	2.597	0.383
Escherichia coli O157:H7 EDL933	2.004	0.509	2.493	0.439	2.586	0.404
Haemophilus influenzae 86-028NP	1.718	-0.466	2.051	-0.432	2.124	-0.390
Lactococcus lactis Il1403	2.389	-0.286	2.845	-0.204	2.939	-0.170
Listeria innocua Clip11262	1.824	-0.530	1.839	-0.511	1.935	-0.488
Listeria monocytogenes EGD-e	1.886	-0.542	1.929	-0.519	1.964	-0.498
Mesorhizobium loti MAFF303099	1.542	0.895	1.317	0.902	1.500	0.881
Mycobacterium leprae TN	1.294	0.844	1.279	0.858	1.372	0.811
Oceanobacillus iheyensis HTE831	1.602	-0.654	1.462	-0.697	1.831	-0.613
Pasteurella multocida str. Pm70	1.884	-0.369	2.351	-0.405	2.351	-0.405
Pyrococcus abyssi GE5	1.164	0.533	1.446	0.583	1.463	0.530
Rickettsia prowazekii str. Madrid E	1.138	-0.782	1.007	-0.784	1.035	-0.741
Salmonella typhimurium LT2	1.902	0.500	2.465	0.377	2.629	0.322
Shewanella oneidensis MR-1	2.586	-0.124	3.008	-0.211	3.014	-0.204
Shigella flexneri 2a str. 2457T	2.066	0.309	2.525	0.242	2.586	0.216
Staphylococcus aureus Mu50	2.112	-0.602	2.389	-0.508	2.425	-0.508
Staphylococcus aureus MW2	2.212	-0.556	2.454	-0.474	2.489	-0.470
Staphylococcus aureus N315	2.167	-0.588	2.415	-0.504	2.467	-0.500
Streptococcus agalactiae 2603V/R	2.390	-0.488	2.867	-0.410	2.915	-0.381
Streptococcus agalactiae NEM316	2.529	-0.499	3.006	-0.442	3.057	-0.411
Streptococcus mutans UA159	1.903	-0.606	2.281	-0.545	2.301	-0.539
Streptococcus pneumoniae R6	2.486	-0.283	2.834	-0.287	2.832	-0.286
Streptococcus pneumoniae TIGR4	2.488	-0.253	2.873	-0.262	2.873	-0.262
Streptococcus pyogenes MGAS315	2.565	-0.416	2.983	-0.418	2.983	-0.418
Streptococcus pyogenes MGAS8232	2.620	-0.412	2.983	-0.419	2.987	-0.416
Synechocystis sp. PCC 6803	0.960	0.671	1.363	0.588	1.385	0.579
Vibrio cholerae N16961 chromosome I	2.441	0.020	3.105	-0.170	3.109	-0.168
Yersinia pestis CO92	2.152	0.122	2.445	0.120	2.461	0.121
Yersinia pestis KIM	2.216	0.084	2.576	0.041	2.579	0.047

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_{HEDB} 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.

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

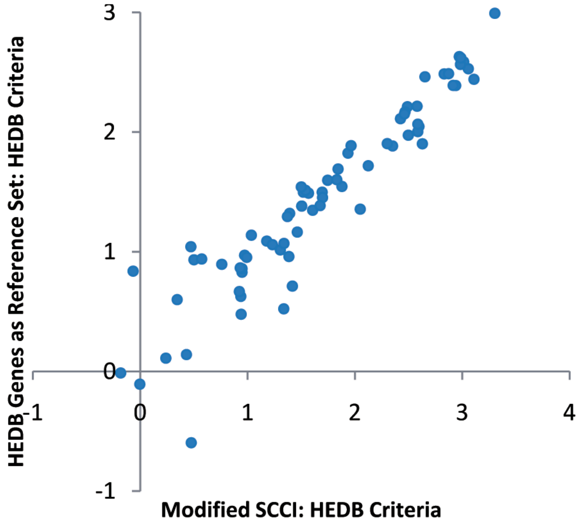


Fig. 6. Association between HEDB criteria when using mSCCI and HEDB genes as reference set. Note that most of the genomes that can be isolated using mSCCI are also the genomes that can be isolated when using HEDB genes as a reference set.

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
Spearman Rank Correlation Coefficients between Microarray Expression Data and $\mathrm{CAI}_{S\&L}/\mathrm{CAI}_{HEDB}/\mathrm{SCCI}/\mathrm{mSCCI}$ Values as Determined by Various Methods: Confounded Organisms

Organism	$\mathrm{CAI}_{\mathrm{S\&L}}$	$\mathrm{CAI}_{\mathrm{HEDB}}$	mSCCI	SCCI
Caulobacter crescentus CB15	0.280	0.306	0.311	0.284
Halobacterium sp. NRC-1	0.165	0.086	0.199†	0.151
Nostoc sp. PCC 7120	0.046	0.150	0.275†‡	-0.269
Pseudomonas aeruginosa PAO1	0.347	0.303	0.381†	0.204
Streptomyces coelicolor A3(2)	0.131	0.148	0.151	0.130

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_{HEDB} indicates CAI values are determined using HEDB genes as a reference set. $\mathrm{CAI}_{S\&L}$ 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_{HEDB}. ‡ mSCCI r_S value significantly more positive than that of $\mathrm{CAI}_{S\&L}$. 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
Spearman Rank Correlation Coefficients between Microarray Expression Data and $\mathrm{CAI}_{S\&L}/\mathrm{CAI}_{HEDB}/\mathrm{SCCI}/\mathrm{mSCCI}$ Values as Determined by Various Methods: Nonconfounded Organisms

Organism	$\mathrm{CAI}_{\mathrm{S\&L}}$	$\mathrm{CAI}_{\mathrm{HEDB}}$	mSCCI	SCCI
Bacillus subtilis subsp. subtilis str. 168	0.159	0.121	0.135	0.131
Shewanella oneidensis MR-1	0.286	0.323	0.296	0.295
Chlamydophila pneumoniae AR39	0.230	0.228	0.213	0.213
Escherichia coli K12	0.501	0.496	0.506	0.505
Lactococcus lactis Il1403	0.403	0.417	0.417	0.414

All Spearman rank correlation coefficients are significant $(p < 0.05)$. CAI_{HEDB} indicates CAI values are determined using HEDB genes as a reference set. $\mathrm{CAI}_{S\&L}$ 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.

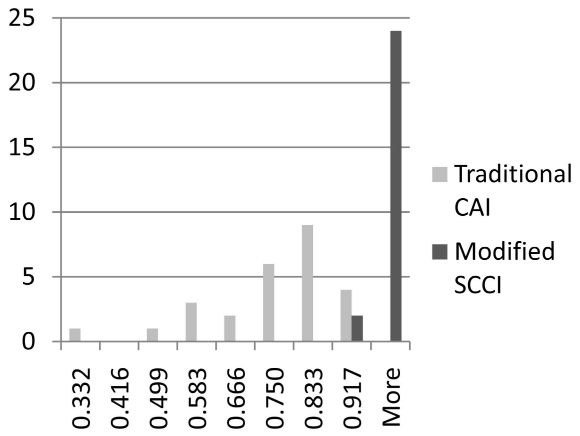


Fig. 7. Distribution of quality scores for the confounded organism reference sets when the traditional (Sharp and Li) approach is utilized (with HEDB genes as the reference set) and when the mSCCI technique is employed.

translational efficiency bias when traditional methods do not (HEDB criterion $>$ 1 using mSCCI; $\leq$ 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).

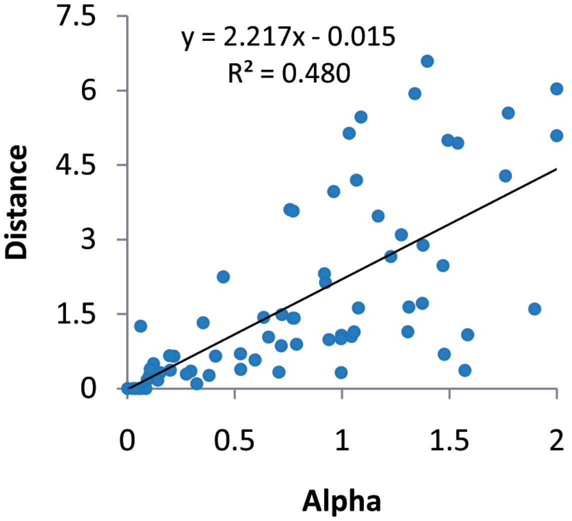
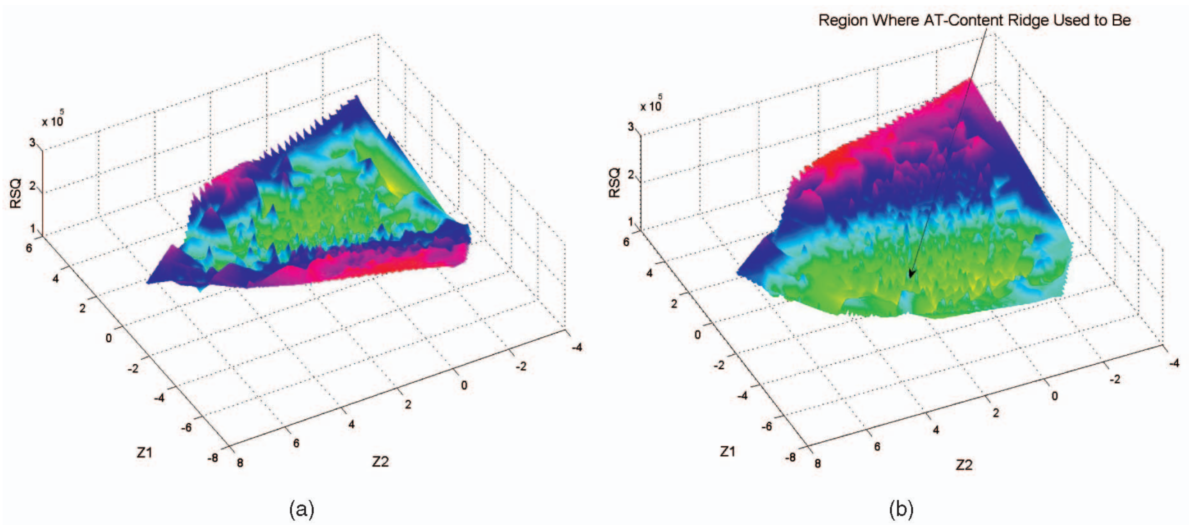


Fig. 8. Relationship between α (degree to which two biases work in concert) and distance between biases ($r^2 = 0.48,\ p < 0.05$). Organisms that have confounding biases that are near the translational efficiency bias exhibit low α's while those that are far apart have large α's. Euclidean distance between 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).

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



(a)

(b)

Fig. 9. Side-by-side comparison of *Nostoc sp. PCC 7120* RSQ landscapes. (a) Unadjusted and (b) adjusted for GC-content. Note that the region that corresponds to high AT-content (that previously dominated the landscape) is depressed in the adjusted landscapes and no longer confounds the search for translational efficiency bias.

isolates translationalefficiency 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

[1] M. Bulmer, "The Selection-Mutation-Drift Theory of Synonymous Codon Usage," *Genetics,* vol. 129, no. 3, pp. 897-907, Nov. 1991.

[2] S. Varenne, J. Buc, R. Lloubes, and C. Lazdunski, "Translation is a Non-Uniform Process. Effect of tRNA Availability on the Rate of Elongation of Nascent Polypeptide Chains," *J. Molecular Biology,* vol. 180, no. 3, pp. 549-576, Dec. 1984.

[3] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pavé, "Codon Catalog Usage and the Genome Hypothesis," *Nucleic Acids Research,* vol. 8, no. 1, pp. r49-r62, http://nar.oupjournals.org/cgi/content/abstract/9/1/r43, 1981.

[4] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier, "Codon Catalog Usage is a Genome Strategy Modulated for Gene Expressivity," *Nucleic Acids Research,* vol. 9, no. 1, pp. r43-74, http://nar.oupjournals.org/cgi/content/abstract/9/1/r43, 1981.

[5] T. Ikemura, "Correlation between the Abundance of *Escherichia coli* Transfer RNAs and the Occurrence of the Respective Codons in Its Protein Genes," *J. Molecular Biology,* vol. 146, pp. 1-21, 1981.

[6] M. Gouy and C. Gautier, "Codon Usage in Bacteria: Correlation with Gene Expressivity," *Nucleic Acids Research,* vol. 10, no. 22, pp. 7055-7074, http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=6760125, 1982.

[7] P.M. Sharp, E. Cowe, D.G. Higgins, D.C. Shields, K.H. Wolfe, and F. Wright, "Codon Usage Patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*: A Review of the Considerable Within-Species Diversity," *Nucleic Acids Research,* vol. 16, no. 17, pp. 8207-8211, Sept. 1988.

[8] F. Wright, "The "Effective Number of Codons" Used in a Gene," *Gene,* vol. 87, pp. 23-29, 1990.

[9] R.J. Grocock and P.M. Sharp, "Synonymous Codon Usage in *Pseudomonas aeruginosa* PA01," *Gene,* vol. 289, noS. 1/2, pp. 131-139, May 2002.

[10] A. Carbone, F. Képès, and A. Zinovyev, "Codon Bias Signatures, Organization of Microorganisms in Codon Space, and Lifestyle," *Molecular Biology and Evolution,* vol. 22, no. 3, pp. 547-561, http://dx.doi.org/10.1093/molbev/msi040, Mar. 2005.

[11] B. Lafay, A.T. Lloyd, M.J. McLean, K.M. Devine, P.M. Sharp, and K.H. Wolfe, "Proteome Composition and Codon Usage in Spirochaetes: Species-Specific and DNA Strand-Specific Mutational Biases," *Nucleic Acids Research,* vol. 27, no. 7, pp. 1642-1649, Apr. 1999.

[12] R. Jain, M.C. Rivera, J.E. Moore, and J.A. Lake, "Horizontal Gene Transfer in Microbial Genome Evolution," *Theoretical Population Biology,* vol. 61, no. 4, pp. 489-495, June 2002.

[13] S. Waack, O. Keller, R. Asper, T. Brodag, C. Damm, W.F. Fricke, K. Surovcik, P. Meinicke, and R. Merkl, "Score-Based Prediction of Genomic Islands in Prokaryotic Genomes Using Hidden Markov Models," *BMC Bioinformatics,* vol. 7, p. 142, http://dx.doi.org/10.1186/1471-2105-7-142, 2006.

[14] P.M. Sharp and W.H. Li, "An Evolutionary Perspective on Synonymous Codon Usage in Unicellular Organisms," *J. Molecular Evolution,* vol. 24, no. 1/2, pp. 28-38, 1986.

[15] J. Precup and J. Parker, "Missense Misreading of Asparagine Codons as a Function of Codon Identity and Context," *J. Biological Chemistry,* vol. 262, no. 23, pp. 11351-11355, Aug. 1987.

[16] D.C. Shields and P.M. Sharp, "Synonymous Codon Usage in *Bacillus Subtilis* Reflects Both Translational Selection and Mutational Biases," *Nucleic Acids Research,* vol. 15, no. 19, pp. 8023-8040, http://nar.oupjournals.org/cgi/content/abstract/15/19/8023, 1987.

[17] M.A. Freire-Picos, M.I. Gonzalez-Siso, E. Rodriguez-Belmonte, A.M. Rodriguez-Torre, E. Ramil, and M.E. Cerdan, "Codon Usage in *Kluyveromyces lactis* and in Yeast Cytochrome C-Encoding Genes," *Gene,* vol. 139, pp. 43-49, 1994.

[18] S. Kanaya, Y. Yamada, Y. Kudo, and T. Ikemura, "Studies of Codon Usage and tRNA Genes of 18 Unicellular Organisms and Quantification of *Bacillus subtilis* tRNAs: Gene Expression Level and Species-Specific Diversity of Codon Usage Based on Multivariate Analysis," *Gene,* vol. 238, pp. 143-155, 1999.

[19] P.M. Sharp and W.H. Li, "The Codon Adaptation Index—A Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications," *Nucleic Acids Research,* vol. 15, pp. 1281-1295, 1987.

[20] A. Carbone, A. Zinovyev, and F. Kepes, "Codon Adaptation Index as a Measure of Dominating Codon Bias," *Bioinformatics,* vol. 19, no. 16, pp. 2005-2015, http://bioinformatics.oupjournals.org/cgi/content/abstract/19/16/2005, 2003.

[21] A.C. McHardy, A. Pühler, J. Kalinowski, and F. Meyer, "Comparing Expression Level-Dependent Features in Codon Usage with Protein Abundance: An Analysis of "Predictive Proteomics"," *Proteomics,* vol. 4, no. 1, pp. 46-58, 2004.

[22] A. Carbone, "Computational Prediction of Genomic Functional Cores Specific to Different Microbes," *J. Molecular Evolution,* vol. 63, no. 6, pp. 733-746, http://dx.doi.org/10.1007/s00239-005-0250-9, Dec. 2006.

[23] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *J. Educational Psychology,* vol. 24, pp. 417-441, 1933.

[24] P. Sharp, T. Tuohy, and K. Mosurski, "Codon Usage in Yeast: Cluster Analysis Clearly Differentiates Highly and Lowly Expressed Genes," *Nucleic Acids Research,* vol. 14, no. 13, pp. 5125-5143, http://nar.oupjournals.org/cgi/content/abstract/14/13/5125, 1986.

[25] A. Ghosh, S. Tsutsui, and S. Tusutsui, *Advances in Evolutionary Computing.* Springer, 2003.

[26] V.K. Vassilev, T.C. Fogarty, and J.F. Miller, "Smoothness, Ruggedness and Neutrality of Fitness Landscapes: From Theory to Application," *Advances in Evolutionary Computing: Theory and Applications,* pp. 3-44, 2003.

[27] E.D. Weinberger, "Correlated and Uncorrelated Fitness Landscapes and How to Tell the Difference," *Biological Cybernetics,* vol. 63, pp. 325-336, 1990.

[28] S. Wright, "The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution," *Proc. Sixth Int'l Congress on Genetics,* pp. 355-366, 1932.

[29] C.B. Barber, D.P. Dobkin, and H. Huhdanpaa, "The Quickhull Algorithm for Convex Hulls," *ACM Trans. Math. Software,* vol. 22, no. 4, pp. 469-483, 1996.

[30] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic Local Alignment Search Tool," *J. Molecular Biology,* vol. 215, pp. 403-410, 1990.

[31] W. Press, S. Teukolsky, W.T. Vetterling, and B. Flannery, *Numerical Recipes in C, the Art of Scientific Computing,* second ed. Cambridge Univ. Press, 1999.

[32] A. Zinovyev and A. Carbone, "CAIJava," *Calculates Codon Frequencies and CAI-Values of All Genes,* http://www.ihes.fr/\%7Ematerials/CAIJava.java, 2002.

[33] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, "NCBI GEO: Mining Tens of Millions of Expression Profiles-Database and Tools Update," *Nucleic Acids Research,* vol. 35, database issue, pp. D760-D765, http://www.ncbi.nlm.nih.gov/geo/, http://dx.doi.org/10.1093/nar/gkl887. Jan. 2007.

[34] R. Wünschiers and H. Eckes, *HyDaBa Hydrogen Database: Nostoc,* http://www.hydaba.uni-koeln.de/index.php, Apr. 2005.

[35] NCBI, Nat'l Center for Biotechnology Information, http://www.ncbi.nih.gov/, May 2005.

[36] R.A. Fisher, "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika,* vol. 10, no. 4, pp. 507-521, May 1915.

[37] E.C. Fieller, H.O. Hartley, and E.S. Pearson, "Tests for Rank Correlation Coefficients. I," *Biometrika,* vol. 3, no. 4, pp. 470-481, Dec. 1957.

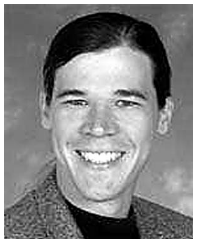
[38] A. Otero and M. Vincenzini, "Extracellular Polysaccharide Synthesis by Nostoc Strains as Affected by n Source and Light Intensity," *J. Biotechnology,* vol. 102, no. 2, pp. 143-152, Apr. 2003.

[39] M. dos Reis, R. Savva, and L. Wernisch, "Solving the Riddle of Codon Usage Preferences: A Test for Translational Selection," *Nucleic Acids Research,* vol. 32, no. 17, pp. 5036-5044, http://dx.doi.org/10.1093/nar/gkh834, 2004.

[40] P. Puigbò, I.G. Bravo, and S. Garcia-Vallvé, "E-CAI: A Novel Server to Estimate an Expected Value of Codon Adaptation Index (ECAI)," *BMC Bioinformatics,* vol. 9, p. 65, http://dx.doi.org/10.1186/1471-2105-9-65, 2008.

Douglas W. Raiford received the BS, MS, and PhD degrees in computer science from Wright State University, Dayton, Ohio, in 2002, 2005, and 2008, respectively. He is with the Department of Computer Science and Engineering, Wright State University. His research interests include bioinformatics, investigating global forces involved in genomic and proteomic evolution, codon usage bias, evolutionary computation, and pattern recognition.

Dan E. Krane received the BS degree in biology and chemistry from John Carroll University, Cleveland, Ohio, in 1985 and the PhD degree in biochemistry from Pennsylvania State University in 1990. He pursued postdoctoral research at Washington University and Harvard University before accepting a faculty appointment in the Department of Biological Sciences, Wright State University (WSU), Dayton, Ohio, in 1993. His research interests include the areas of molecular evolution and population genetics. Since 1991, he has been testifying as an expert witness in more than 60 criminal trials in which DNA evidence has been presented. He is a cofounder of Forensic Bioinformatics (www.bioforensics.com), a consulting company that generates automated reviews of forensic DNA profile evidence for hundreds of court cases each year. He is also the lead author of the first textbook coauthored by a biologist and computer scientist that is specifically designed to make bioinformatics accessible to undergraduates and prepare them for more advanced work.



Travis E. Doom received the BS degree in computer science and mathematics from Bowling Green State University, Bowling Green, Ohio, in 1992 and the MS and PhD degrees from Michigan State University, East Lansing, in 1994 and 1998, respectively. He is currently an associate professor in the Department of Computer Science and Engineering, Wright State University (WSU), Dayton, Ohio. His research interests include design automation, computational biology, optimization theory, and engineering education. He is an associate member of WSU's Biomedical Sciences PhD program faculty and a co-director of WSU's bioinformatics research group. He is a senior member of the IEEE.

Michael L. Raymer received the BS degree in computer science from Colorado State University, Fort Collins, in 1991 and the MS and PhD degrees from Michigan State University, East Lansing, in 1995 and 2000, respectively. He is currently an associate professor in the Department of Computer Science and Engineering, Wright State University (WSU), Dayton, Ohio. He is an associate member of WSU's Biomedical Sciences PhD program faculty and a co-director of WSU's bioinformatics research group. His research interests include protein structure and function, bioinformatics, evolutionary computation, and pattern recognition. He has coauthored one of the first bioinformatics textbooks developed for undergraduate students. He is a member of the IEEE.