2011

A Genetic Optimization Approach for Isolating Translational Efficiency Bias

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: <https://corescholar.libraries.wright.edu/knoesis>

Part of the [Bioinformatics Commons](#), [Communication Technology and New Media Commons](#), [Databases and Information Systems Commons](#), [OS and Networks Commons](#), and the [Science and Technology Studies Commons](#)

Repository Citation

Raiford, D. W., Krane, D. E., Doom, T. E., & Raymer, M. L. (2011). A Genetic Optimization Approach for Isolating Translational Efficiency Bias. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8 (2), 342-352.
<https://corescholar.libraries.wright.edu/knoesis/529>

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 library-corescholar@wright.edu.

A Genetic Optimization Approach for Isolating Translational Efficiency Bias

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

Abstract—The study of codon usage bias is an important research area that contributes to our understanding of molecular evolution, phylogenetic relationships, respiratory lifestyle, and other characteristics. Translational efficiency bias is perhaps the most well-studied codon usage bias, as it is frequently utilized to predict relative protein expression levels. We present a novel approach to isolating translational efficiency bias in microbial genomes. There are several existent methods for isolating translational efficiency bias. Previous approaches are susceptible to the confounding influences of other potentially dominant biases. Additionally, existing approaches to identifying translational efficiency bias generally require both genomic sequence information and prior knowledge of a set of highly expressed genes. This novel approach provides more accurate results from sequence information alone by resisting the confounding effects of other biases. We validate this increase in accuracy in isolating translational efficiency bias on 10 microbial genomes, five of which have proven particularly difficult for existing approaches due to the presence of strong confounding biases.

Index Terms—Codon usage bias, evolutionary computing and genetic algorithms, miscellaneous, artificial intelligence, computing methodologies, GC-content, strand bias, translational efficiency.

1 INTRODUCTION

CONVENTIONAL wisdom has held that natural selection differentiates among alleles based upon the competitive advantage or disadvantage conferred by mutations that result in changes to the end product of the gene [1]. Recent work, however, has revealed that selective advantage can be present at a much finer level than was previously appreciated. An example of a subtle, genomewide trend that can confer such a selective advantage is the biased usage of codons [2], [3], [4]. Because the genetic code is degenerate (with many amino acids coded for by multiple synonymous codons), there exists the possibility for species to use some codons preferentially, while eschewing others. Microbial organisms can exhibit bias toward increased (or decreased) usage of G and C nucleotides (content bias), bias induced by differential mutation rates along the leading and lagging strands during DNA replication (strand bias), and bias toward codons associated with more common tRNAs (translational efficiency bias). The latter—a tendency to use codons associated with common tRNAs—is assumed to confer more efficient translation of highly expressed genes. The resulting bias has been shown to have a subtle, genomewide effect upon many simple organisms.

The study of translational efficiency codon usage bias has proven to be an important research topic. The degree to which a gene adheres to translational efficiency bias has been shown to correlate with the gene's expression level in many Bacteria [4], [5], [6] and some low-order Eukarya [7], [8]. This trait has been exploited in studies that have used adherence to translational efficiency bias as a proxy for expressivity [9], [10], [11], [12]. Use of translational efficiency bias as a surrogate for expressivity is useful, in part, because experimental methods for obtaining transcript or protein abundance (such as that provided by oligonucleotide microarrays or gel electrophoresis) are relatively expensive in terms of time, materials, and reagent cost. Estimating expressivity using translational efficiency bias may provide an additional benefit in that experimental methods provide measures of transcript or protein abundance, not the underlying production rates. Additionally, knowledge of codon adaptiveness associated with translational efficiency can be employed to maximize the yield of heterologous protein production (genes introduced into a host organism's genome for the purpose of producing recombinant protein) [13], [14]. By modifying the introduced gene to conform to the host organism's codon usage bias, an increase in protein product can be realized [15].

One of the earliest methods for determining adherence to translational efficiency bias did so first by examining highly expressed genes to identify which codons were preferred, and then calculating the frequency of optimal codon usage (FOP) for the genes in the genome [5]. It was shown that FOP was highly correlated with protein expression levels in some Bacteria. This is a binary interpretation of codon preference. Either the codon is identified as optimal, or it is not. Later, the measurement of adherence to translational efficiency bias was refined by introducing gradation to the process. The concept of degree of preference, or the degree to which a codon has adapted, was introduced with the codon adaptation index (CAI) measurement [16].

- D.W. Raiford is with the Department of Computer Science, University of Montana, 32 Campus Dr., Missoula, MT 59812.
E-mail: Douglas.Raiford@umontana.edu.
- D.E. Krane is with the Department of Biological Sciences, Wright State University, 3640 Colonel Glenn Highway, Dayton, OH 45435.
E-mail: dan.krane@wright.edu.
- T.E.W. Doom and M.L. Raymer are with the Department of Computer Science and Engineering, Wright State University, 3640 Colonel Glenn Highway, Dayton, OH 45435.
E-mail: {travis.doom, michael.raymer}@wright.edu.

Manuscript received 24 July 2008; revised 31 Dec. 2008; accepted 3 Feb. 2009; published online 11 Feb. 2009.

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2008-07-0133. Digital Object Identifier no. 10.1109/TCBB.2009.24.

Fig. 1. Reference set placement in a sorted list of genes for an unconfounded (*Escherichia coli* K12) and a confounded (*Campylobacter jejuni* NCTC 11168) organism. In this context, confounded indicates an organism whose dominant codon usage bias is something other than translational efficiency. Horizontal lines indicate location of reference set genes. Gene ranking indicates position within a listing of genes sorted according to CAI score. Note the large number of reference set genes for the confounded organism that fall in the lower (fourth) quartile of the sorted (by CAI) list of genes. Even the unconfounded organism has a gene in the lower portion of the third quartile. The extent to which reference set genes fall in these lower quartiles is an indication of the amount of error in the expression-prediction ability of the computed CAI scores. (a) Unconfounded. (b) Confounded.

As a testament to their effectiveness and ease of implementation, these types of methods (methods like FOP and CAI) have remained in use, largely unmodified, since their conception. These tally-based methods rely upon the uniform conformance of highly expressed genes to the underlying translational efficiency bias in order to identify the degree of preference for each codon.

However, the adherence of highly expressed genes to translational efficiency bias is far from uniform. As an example, on average, only 58 percent of the highly expressed genes for *Saccharomyces cerevisiae* prefer the codon that is preferred "overall." This lack of uniformity can result in a reduced ability to predict expressivity. Fig. 1 shows the predicted placement of the most highly expressed genes in a sorted list of all the genes in two genomes (one known to have translational efficiency bias as the dominant bias and one thought to exhibit little or no translational efficiency bias but strong GC-content and strand biases). The degree to which the highly expressed genes fall in the lower regions of the sorted lists is an indication of prediction error. This error is particularly pronounced in organisms with genomes that exhibit codon usage trends dominated by a bias other than that of translational efficiency bias (Fig. 1b). Throughout the rest of this paper we refer to these organisms as confounded. In Section 3, specific criteria are presented for determining whether an organism is confounded or not. All tally-based translational efficiency bias identification methods (methods that use the codon counts in a set of highly expressed genes to determine codon preference) are susceptible to the

confounding effects of other biases (such as content or strand bias) simultaneously influencing codon selection.

The direct search-based approach described here resists these confounding effects. Through the use of this methodology, translational efficiency bias can be identified in genomes previously thought not to exhibit this bias. Also, improvements in expression prediction can be realized for organisms where other biases confound the isolation process. The effectiveness of the approach is tested on 10 genomes. Five of these genomes are known to be dominant for either GC- or AT-content (and in two cases, also dominant for strand bias) and are believed to exhibit weak or nonexistent translational efficiency bias. The remaining genomes are dominant for translational efficiency bias [17]. The performance of the approach is assessed by comparing the resultant CAI score gene rankings with that of experimentally determined expression data.

2 METHODS

The direct search-based approach operates similarly to the traditional CAI technique [16]. The traditional method uses a tally-based approach to determine the degree of preference, or weight, for each codon by examining a set of highly expressed genes. A codon with a high degree of usage in the highly expressed genes is assigned a relatively large weight. The direct search approach, instead, searches for weights that explain the high adherence scores of the highly expressed genes. In both approaches, these weights are then used to measure the adherence of the rest of the genes in the genome.

2.1 Reference Sets

Both the traditional CAI calculation method and the direct search-based approach described here rely upon the identification of a set of genes (the reference set) that is determined experimentally (or known, generally) to comprise the most highly expressed genes of the genome. The direct search-based approach uses as a reference set the most highly expressed genes as determined from experimental microarray expression data (Section 2.8) or as determined through the use of a high-expression database (HEDB, Section 2.4). In the case of the experimentally determined highly expressed (EDHE) genes, the size of the reference set is set to 1 percent of the genes in the genome. This size was chosen because it roughly corresponds to the size of reference sets proposed by the authors of [6], [7], [16], [18], [19], [20]. In the case of the HEDB-determined reference set, the size is set by the number of hits returned from the database. These reference sets and reference set sizes are the same for both the traditional CAI and the genetic algorithm methods when they are compared.

2.2 CAI

The Sharp and Li [16] CAI approach assigns a value known as the relative *adaptiveness* (or weight) to each codon. The weight for each codon is derived from the codon usage frequencies of genes in the reference set. These frequencies are normalized by the average usage for the set of synonymous codons related to the codon in question. This normalized frequency is known as relative synonymous

codon usage (RSCU) [21], (1). The RSCU and the weight of a codon w_{ij} are:

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}}, \quad (1)$$

$$\text{Normalized } RSCU = w_{ij} = \frac{RSCU_{ij}}{RSCU_{i(max)}} = \frac{X_{ij}}{X_{i(max)}}. \quad (2)$$

X_{ij} is the count of the codon within the subset of genes known as the reference set, and $X_{i(max)}$ is the count of its sibling (siblings are codons from the family of codons that code for a given amino acid) with the highest count (the maximal sibling will have a weight of one). w_{ij} is the weight w of the i th codon for the j th amino acid. To determine the CAI of a particular gene, the geometric mean of the normalized RSCU value (or weight) for each codon in the gene is calculated (3). Here, L is the length of the gene (in codons)

$$CAI(g) = \sqrt[L]{\prod_{i=1}^L w_i}. \quad (3)$$

2.3 Ribosomal Criterion

To determine whether an identified bias is, specifically, translational efficiency, it is necessary to show that highly expressed genes adhere strongly to that bias. Carbone et al. [17] have developed a measure that can help determine whether an identified bias is representative of translational efficiency. The measure is known as the *ribosomal criterion*. This criterion is based on the degree to which the ribosomal protein coding genes (which may be assumed to be highly expressed) are found in the upper region of a list of genes sorted by CAI value. Carbone et al. conclude that the identified bias is that of translational efficiency when the average CAI value for the organism's ribosomal protein coding genes (RPCGs) is greater than one standard deviation above the mean CAI value for the organism's genome

$$z = \frac{(CAI(r) - \overline{CAI})}{\sigma_{CAI}}, \quad (4)$$

where $CAI(r)$ is the CAI score of a given ribosomal protein coding gene. The average of the z-scores for all RPCGs (4), $\bar{z}$, is the ribosomal criterion. A genome characterized by translational efficiency bias would be expected to have a high $\bar{z}$.

2.4 High-Expression Database

Recently, sequenced genomes may not have identified ribosomal protein coding genes. To allow for this contingency, we have developed a database of highly expressed proteins (HEDB) that can be used to identify homologs within the target genome [22]. 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 scores in well-characterized genomes (extensively studied genomes whose dominant bias is known to be translational efficiency, data not shown).

An unfiltered blast search is performed [23] to identify probable highly expressed genes in a target genome. This

search is performed for each gene in the target genome, using the corresponding protein as a query against the HEDB. The query protein is considered to be a homolog of database proteins with 60 percent identity. The database currently contains proteins from 66 organisms. The organisms used to build the HEDB are drawn from 25 different bacterial taxonomical subclasses, or groups, in order to achieve a classwide representative sampling. The database contains no proteins from organisms used as target genomes in this study to guarantee independence between the genomes being analyzed and the HEDB.

2.5 Strength Criterion

A characteristic of a genome with a strong, dominant bias of translational efficiency is that the most highly expressed genes (the reference set) tend to be much more highly biased than the rest of the genome. The presence of this condition (of a strongly biased reference set) can be further evidence that a detected bias is that of translational efficiency bias. The CAI weights associated with the reference set are an indication of the strength of this bias. As an example, consider the strongest possible bias that could be exhibited. The maximal sibling will always have a weight assignment of one. The strongest possible bias would yield weights of zero for the rest of the siblings. Weaker biases would have nonmaximal sibling weights closer to one. Thus, one measure of the strength of the bias is indicated by how far the weight vector (vector comprised of the 64 codon-associated weights) deviates from balanced usage. Another, perhaps more useful, measure is how much more biased the weight vector is from the overall bias of the organism. Carbone et al. [17] developed such a criterion for measuring bias strength (5)

$$d(w^G, w^S) = \frac{\sum_{i=1}^{64} |w_i(G) - w_i(S)|}{2}. \quad (5)$$

The overall bias of the organism is obtained by extracting the weight vector generated using the entire genome as the reference set (w^G). This is compared to the weight vector of the smaller reference set containing only highly expressed genes (w^S). The result can be thought of as measuring the number of codon families where the preferred codon is different in two biases.

Consider a case of a binary weight vector (each family has a single weight of one and the rest zero) where the major codons change for all families between G and S . This yields a summation of the differences (Manhattan distance) of 40. Dividing by two brings the number back to the number of changed major codons (20). In other words, (5) is equal to one-half the Manhattan distance.

A genome characterized by translational efficiency bias should have a relatively large strength measure. Carbone et al. [17] set the criterion for an organism's bias to be considered strong at $d(w^G, w^S) > 8$, but they note that not all organisms characterized by translational efficiency bias meet this criterion. Instead, they use it in conjunction with the ribosomal criterion to differentiate weak and strong translational efficiency bias.

2.6 Direct Search-Based Approach

The direct search-based approach differs from the traditional method in the way in which the weights are determined.

Instead of assigning weights based upon the codon usage of the reference set, a genetic algorithm performs a computational search for an optimal set of weights. The search looks for a set of weights that optimizes the placement of the reference set in a sorted listing of genes according to their adherence to the bias defined by the weights.

A genetic algorithm is an optimization technique modeled after nature's approach to maximizing a species' fitness for survival and reproduction [24], [25]. The algorithm simulates natural selection by maintaining a population of individual members, each of which is a proposed solution to the given search problem. The search for the best solution proceeds by selecting members from the population to produce offspring that combine traits from the parents. A fitness score is assigned to each proposed solution and poor solutions are removed from the population in a process that mimics natural selection. Offspring in nature vary from their parents through recombination of the parents' DNA (also known as crossover) and through the occurrence of genetic mutations. In order to imitate this behavior, the GA must have a representation of the solution that is amenable to these operations (crossover and mutation).

2.6.1 Representation

The search being performed is for a set of weights. These weights are represented as a 64-dimensional (one for each codon) vector that contains real-valued weights. The GA will search for weights that produce high CAI scores in high-expression genes and lower scores for the rest of the genes in the genome.

2.6.2 Selection of Parents

Identification of the parents for the purposes of recombination is performed using tournament selection [26]. Candidates are chosen at random until a tournament size of three is achieved (tournament size of three was chosen with the intent to impose a small degree of selective pressure so as to avoid premature convergence to local optima). The member with the best fitness within this group of three is the tournament winner and is selected as a parent. Once two parents are identified, crossover (Section 2.6.3) is employed to generate a child chromosome. This is followed by a mutation operation (Section 2.6.4) to add additional variation into the population.

2.6.3 Crossover Operator

Crossover between two population members with more simple representations (such as bit strings) can be achieved by choosing the contents from one of the parents at each gene (bit) location, and inserting it into the appropriate position of the offspring's bit string. While this is a viable approach in real-valued vectors, there are other, often more effective, techniques in real vector spaces [27]. Simply populating the offspring with entries from each parent can cause wild, unpredictable jumps in the search space. Instead, if a point is chosen that is somewhere between the two chromosomal vectors, such as a midpoint or a weighted midpoint, then crossover can become a more meaningful search operation in many real-valued vector spaces.

The GA described here employs geometric crossover as described in [28] where the entry in the offspring's real-valued vector is the geometric mean of the two parents. This

method was chosen because the geometric mean favors the smaller weight (the conservative choice in terms of choosing most preferred codons). This is due to the "fundamental inequality" (the geometric mean is never greater than the arithmetic mean). Additionally, this is similar to the way in which CAI is determined; it is the geometric mean of the weights associated with the codons in each gene.

2.6.4 Mutation Operator

Mutation is achieved by slightly perturbing each weight within the weight vector. The new weight is a randomly generated real number normally distributed about the current weight. Equation (7) demonstrates the generation of a mutated weight w'_i from the original weight w_i . N is a function that generates a random number that is normally distributed about a mean (w) with a given standard deviation (σ). i is the index that identifies which of the 64 dimensions in the vector is associated with the weight in question. The mutation is self-limiting to a value that is no greater than one. For the GA described here, the standard deviation of the random normal probability distribution function is greater in the early generations and less during the late generations. This is related to the process of simulated annealing [29], [30]. This slow decrease in σ allows for greater "jumps" early in the search process and smaller steps during the latter stages. For the purposes of this GA, the standard deviation (σ) for a given generation ($gen_{current}$) is determined as in (8), where gen_{final} is the number of generations that the GA will be allowed to run, σ_0 is the starting standard deviation (the standard deviation at generation zero), and b is a parameter that changes the shape of the curve described by the function. The GA described here uses a b parameter of one and an initial standard deviation of 0.3. The b value of one is chosen to generate a simple linear decrease in the degree of mutation with respect to generation number. A beginning standard deviation of 0.3 is chosen to distribute the initial mutation-introduced variance normally with room for mutations of up to three standard deviations. The weight scale is from zero to one; this allows three standard deviations of size 0.3 to fit within this range

$$w'_{i_{proposed}} = N(w, \sigma), \quad (6)$$

$$w'_i = \begin{cases} 1, & \text{if } w'_{i_{proposed}} > 1, \\ \varepsilon, & \text{if } w'_{i_{proposed}} \leq 0, \\ w'_{i_{proposed}}, & \text{otherwise,} \end{cases} \quad (7)$$

$$\sigma = \sigma_0 \times \left(1 - \frac{gen_{current}}{gen_{final}}\right)^b. \quad (8)$$

2.6.5 Repair Operator

A repair operator is employed to ensure that the weights derived using the selected mutation and crossover operators are meaningful and comparable to those derived using traditional methods. The maximal sibling must have a weight of one and all other siblings must have weights less than or equal to one and greater than zero. Crossover (a geometric mean) can never increase a weight above one. Likewise, the mutation operator is self-limiting to no greater than one and

greater than zero (7). An ε value (in (7)) of 0.0001 was chosen to allow for relatively small weight values. While the crossover and mutation operators will never generate weights greater than one, they may cause a maximal sibling to have a value less than one. Therefore, a repair operator was developed that sets the maximum weight in each codon family to one, and it increases the weight for each family member an amount proportional to the increase of the maximal sibling. Equation (9) represents the repair performed to weight w_i at each dimension i . The original weight for the maximal sibling is $w_{max_sib_i}$. The maximal sibling weight is normalized to one. All other weights will be increased (or will remain the same) when divided by this value (as all weights are less than or equal to one)

$$w'_i = \frac{w_i}{w_{max_sib_i}}. \quad (9)$$

2.6.6 Fitness Function

The goal of the GA is the discovery of a set of weights that produces a ranked listing of genes (by CAI score) that matches the expectation that highly expressed genes will be highly ranked in such a listing. The technique devised to measure this property is rank fitness. The genes are sorted by CAI score and an index is assigned to all genes. The gene with the smallest CAI value is assigned an index of 0 while the gene with greatest CAI value is assigned an index of $N - 1$, where N is the number of genes in the genome. The indexes of the reference set genes are summed ($IDXs$ in (10)) and this number is normalized by the maximum attainable rank fitness value (11) yielding a value between 0 and 1 representing the degree to which the reference set rises to the top of a sorted (by CAI score) list of genes (12)

$$f(RS) = \sum_{i=1}^{|RS|} IDX_i, \quad (10)$$

$$f_{max}(|RS|) = \sum_{i=N-|RS|}^{N-1} i, \quad (11)$$

$$f_{norm}(RS) = \frac{f(RS)}{f_{max}(|RS|)}. \quad (12)$$

2.6.7 Initializing the Population

The population weights begin with values that are defined using the traditional method (CAI, Section 2.2) and then mutated following a normal distribution with a 0.01 standard deviation. This value is chosen to introduce a fairly small amount of variance into the initial population. This results in the search process beginning in the neighborhood of the weights as determined by the Sharp & Li method [16].

The GA population size is set at 100 with the number of offsprings also set at 100. Population management follows that described by Bäck et al. [31] as $(\mu + \lambda)$ where λ offsprings are generated from μ parents and combined into a single population in which both parents and children must compete for survival. A degree of elitism is introduced as in [32] by maintaining a certain number of the best members of the combined population in the next generation. We set this elitism parameter to 10 percent. The

TABLE 1
Confounding Biases

Organism	GC	AT	Strand
<i>Campylobacter jejuni</i> NCTC 11168		X	X
<i>Halobacterium sp. NRC-1</i>	X		
<i>Nostoc sp. PCC 7120</i>		X	
<i>Pseudomonas aeruginosa</i> PAO1	X		X
<i>Streptomyces coelicolor</i> A3(2)	X		

All listed organisms have as one of their dominant biases either GC- or AT-content. Additionally, two of the organisms are also dominant for strand bias. Results taken from [17].

remaining 90 slots in the next generation's population are selected using a fitness proportional (roulette) selection technique [33], where the more fit the individuals in the combined population are, the more likely they are to be drawn into the following generation. The elitism setting of 10 percent was chosen arbitrarily with the intent of maintaining at least a sampling of the very best solutions from one generation to the next.

2.7 Selection of Target Genomes

The organisms in this study were chosen to demonstrate that the direct search-based methodology can isolate translational efficiency bias (if it exists) in organisms whose genomes are dominant for either translational efficiency bias or some other, confounding bias. The organisms were drawn from those listed in a previous study that identified their dominant biases [17]. Additionally, the genomes chosen have experimental expression data available (Section 2.8). The search-based approach is validated on 10 organisms. Nine of the 10 organisms are Bacteria while one (*Halobacterium sp. NRC-1*) is an Archaeon. Five of the genomes used for validation are known to be dominant for biases other than translational efficiency (Table 1) and, heretofore, were thought to exhibit weak or nonexistent translational efficiency bias. The remaining five genomes in this study are thought to be dominant for translational efficiency bias.

2.8 Microarray Data

The overall objective of the search-based approach is to identify a weight vector that captures a bias in codon usage associated with translational efficiency. Such a bias can be expected to correlate well with the expressivity of the genes in the organism's genome. An independent estimator of the expressivity of the genes is needed to measure the effectiveness of the direct search-based approach in finding a high-quality set of weights. This allows comparison with other methods in terms of ability to identify a bias that correlates well with gene expressivity. Snapshots of mRNA expression levels, such as those obtained using microarray techniques, are one such estimator for the expressivity of each gene.

Expression data are retrieved from the NCBI Gene Expression Omnibus [34] (accession numbers GSE5400, GSE4896, GDS1469, GSE2983, GSE2667, GSE7070, GDS1099, GSE2823, and GSE3876). For *Nostoc*, expression data are instead retrieved from Wünschiers' Hydrogen Database (HyDaBa) which focuses on gene expression data from the filamentous nitrogen-fixing cyanobacterium *Nostoc PCC 7,120* [35]. In dual-channel experiments, the results from

TABLE 2
Measurements for Genetic Algorithms (Unconfounded Organisms)

Organism	CAI _{opt} HEDB				CAI _{opt} EDHE			
	$r_s \mu$	$r_s \sigma$	RF μ	RF σ	$r_s \mu$	$r_s \sigma$	RF μ	RF σ
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	0.146	0.004	0.978	0.001	0.149	0.004	0.978	0.001
<i>Chlamydomonas reinhardtii</i> AR39	0.291	0.008	0.923	0.002	0.290	0.006	0.922	0.002
<i>Escherichia coli</i> K12	0.462	0.007	0.934	0.001	0.461	0.006	0.934	0.001
<i>Lactococcus lactis</i> I1403	0.360	0.013	0.971	0.001	0.357	0.009	0.971	0.001
<i>Shewanella oneidensis</i> MR-1	0.331‡	0.010	0.966	0.000	0.325	0.008	0.967	0.001

CAI_{opt}EDHE: Results for determining CAI values using the GA that optimizes weights by maximizing the placement of experimentally determined highly expressed genes. CAI_{opt}HEDB: GA results that maximize the placement of HEDB genes. $r_s \mu$: Average Spearman rank correlation coefficient over 10 runs. $r_s \sigma$: Standard deviation for Spearman rank correlation coefficients. RF μ : Average rank fitness values over 10 runs (10). RF σ : Standard deviation for rank fitness values. All Spearman rank correlation coefficients are significant ($p < 0.05$). ‡: r_s value significantly more positive than that of CAI EDHE (traditional approach, Table 5). Significance between r_s values determined using a two-tailed Fisher z-transform [37] with 1.06 in the numerator of the variance calculation due to r_s being nonparametric [38].

the reference channel are utilized (no treatment or pre). For *Escherichia coli* K12, those trials using glucose as the carbon supply are used. When raw data are provided, background is subtracted from signal and user-determined flags and thresholds are accepted. For preprocessed data, genes listed as absent are removed from consideration.

2.9 Sequence Data

The sequence information required for calculating weights and the corresponding CAI values are derived from annotated whole genome files located in the complete microbial genomes database on the NCBI Web site (<http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>).

2.10 Implementation and Hardware

Genetic algorithms lend themselves to a high degree of parallelism [36]. A simple technique is to manage the population on a central computing node (parent selection, crossover, mutation, etc.) and to compute fitness concurrently on satellite nodes. This is the strategy employed in our implementation. The GA manager and fitness calculator components described here are written in C++. The jobs are presented to the remote hosts in the form of real-valued weight vectors sent over a socket connection while the results are returned over the same socket connections in the form of “long float” data types (64-bit IEEE floating-point standard format).

3 RESULTS

In order to demonstrate that the search-based approach operates effectively in situations where experimental expression data is present and where it is not, two sets of results were generated; one where the placement of experimentally determined highly expressed genes (CAI_{opt} EDHE) were maximized, and one that maximized the placement of genes identified using the HEDB (CAI_{opt} HEDB).

For each of these approaches (CAI_{opt} EDHE and CAI_{opt} HEDB), two sets of results were generated; one for organisms dominant for translational efficiency bias (ribosomal criterion > 1 when isolated using traditional methods), and another for organisms dominant for either GC- or AT-content, and/or strand bias (as determined by

Carbone et al. [17]; Table 1). Additionally, these confounded organisms exhibit weak or nonexistent translational efficiency bias when traditional techniques are employed (confounded; ribosomal criterion ≤ 1 and strength criterion ≤ 8 when isolated using traditional methods; Figs. 3 and 4).

3.1 Comparison to Experimental Expression Data

In order to compare the performance of the search-based and traditional methods, Spearman rank correlations between bias adherence scores and transcript abundance data are computed for the GA and traditionally generated CAI scores. The two direct search-based result sets are very similar to one another (Tables 2 and 3). For confounded organisms, the results achieve Spearman rank correlations between CAI and transcript abundance that are significantly more positive than either of the Sharp and Li methods (using either experimentally determined highly expressed genes or HEDB genes) for four out of five organisms (Table 3). The search-based results are also more positive for the fifth organism, though not significantly so. For unconfounded organisms, the Spearman rank correlations generally exhibit no statistical difference. There are two exceptions; one where a Sharp and Li approach performs significantly better, and one where a direct search-based approach performs significantly better. CAI HEDB performs significantly better than either search-based approach for *L. lactis*, and CAI_{opt}EDHE performs significantly better than CAI EDHE for *S. oneidensis* (Tables 2 and 4). Significance between r_s values is determined using a two-tailed Fisher z-transform [37] with 1.06 in the numerator of the variance calculation due to r_s being nonparametric [38].

The performance of the direct search-based approach on the genome for *C. jejuni* is particularly noteworthy. For the other organisms, the traditional method of calculating CAI, while confounded (ribosomal criteria ≤ 1 and dominant for content and/or strand bias), nevertheless exhibits a significantly positive correlation with experimentally determined expression data. *C. jejuni* exhibits no such significant trend when traditional CAI is employed. The search-based method, however, is able to disambiguate the translational efficiency bias from the high AT-content and strand biases and yields a ribosomal criterion > 1 and a significantly

TABLE 3
Performance Measurements for Genetic Algorithms (Confounded Organisms)

Organism	CAI _{opt} EDHE				CAI _{opt} HEDB			
	$r_s \mu$	$r_s \sigma$	RF μ	RF σ	$r_s \mu$	$r_s \sigma$	RF μ	RF σ
<i>Campylobacter jejuni</i> NCTC 11168	0.093†‡	0.017	0.970	0.003	0.088†‡	0.014	0.964	0.008
<i>Halobacterium sp.</i> NRC-1	0.388†‡	0.017	0.861	0.007	0.393†‡	0.010	0.868	0.004
<i>Nostoc sp.</i> PCC 7120	0.324†‡	0.030	0.968	0.003	0.324†‡	0.020	0.968	0.002
<i>Pseudomonas aeruginosa</i> PAO1	0.444†‡	0.007	0.968	0.002	0.439†‡	0.007	0.967	0.004
<i>Streptomyces coelicolor</i> A3(2)	0.149	0.006	0.743	0.003	0.146	0.004	0.747	0.002

CAI_{opt}EDHE: Results for determining CAI values using the GA that optimizes weights by maximizing the placement of experimentally determined highly expressed genes. CAI_{opt} HEDB: GA results that maximize the placement of HEDB genes. $r_s \mu$: Average Spearman rank correlation coefficient over 10 runs. $r_s \sigma$: Standard deviation for Spearman rank correlation coefficients. RF μ : Average rank fitness values over 10 runs (10). RF σ : Standard deviation for rank fitness values. All Spearman rank correlation coefficients are significant ($p < 0.05$). †: r_s value significantly more positive than that of CAI HEDB. ‡: r_s value significantly more positive than that of CAI EDHE. (traditional approach, Table 5). Significance between r_s values determined using a two-tailed Fisher z-transform [37] with 1.06 in the numerator of the variance calculation due to r_s being nonparametric [38].

positive Spearman rank correlation with transcript abundance data (Table 3).

3.2 Search Process

A fitness history (Fig. 2) shows that dramatic improvements in rank fitness occur very quickly in the search process. As the GA converges on the final solution, the other population members begin approaching similar rank fitness values to the best in the population (average, median, and worst population members begin approaching the best population member's fitness).

3.3 Comparison of Summary Data

As the GA works by maximizing rank fitness for a given reference set, it stands to reason that the rank fitness values will be greater for the search-based approach than for the traditional method of determining CAI. Less obvious is the effect that this search strategy would have on ribosomal criterion and bias strength. Figs. 3 and 4 demonstrate that the direct search-based approach identifies a bias that is both stronger and that exhibits higher ribosomal criterion

than the biases identified by traditional methods. Comparisons are made between techniques that use the same reference set genes, though other combinations yield similar results (data not shown).

To summarize the effectiveness of the direct search-based approach in the dimensions of strength criterion, ribosomal criterion, rank fitness, and Spearman rank correlation coefficient, we perform principal components analysis [39], [40], and then project the original data onto the first two principal components (Figs. 5 and 6). There is clear separation between the search-based and CAI data for confounded organisms providing visual support for the improved performance of the direct search-based method in all of these dimensions.

A series of t-tests are employed to determine whether the search-based criteria are statistically greater than those of traditional CAI (Table 6). All p -values are less than 0.05 except those associated with tests between r_s values for unconfounded organisms. The p -values for the r_s values are

TABLE 4
Performance Measurements for
Traditional CAI (Unconfounded Organisms)

Organism	CAI EDHE		CAI HEDB	
	r_s	RF	r_s	RF
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	0.159	0.730	0.121	0.958
<i>Chlamydomonas reinhardtii</i> AR39	0.230	0.754	0.228	0.877
<i>Escherichia coli</i> K12	0.501	0.903	0.496	0.910
<i>Lactococcus lactis</i> IL1403	0.403	0.935	0.417†‡	0.930
<i>Shewanella oneidensis</i> MR-1	0.286	0.919	0.323	0.942

CAI EDHE: CAI determined using the traditional Sharp and Li approach using experimentally determined highly expressed genes as a reference set. CAI HEDB: CAI determined using the traditional Sharp and Li approach using HEDB genes as a reference set. r_s : Spearman rank correlation coefficient with experimentally determined expression data. RF: Rank fitness value (10). All Spearman rank correlation coefficients are significant ($p < 0.05$). †: r_s value significantly more positive than that of CAI_{opt} EDHE (Table 2). ‡: r_s value significantly more positive than that of CAI_{opt} HEDB. Significance between r_s values determined using a two-tailed Fisher z-transform [37] with 1.06 in the numerator of the variance calculation due to r_s being nonparametric [38].

Fig. 2. Best, average, median, and worst fitness values for each generation of a typical GA run drawn from one for *P. aeruginosa*.

Fig. 3. Comparison of strength criterion (5) and ribosomal criterion (4) for confounded organisms between GA ($CAI_{opt} EDHE$) that determines CAI values by optimizing weights in a way that maximizes location of experimentally determined highly expressed genes and traditional CAI ($CAI EDHE$) using those same genes as a reference set. Each point represents an organism. GA data are the means of 10 runs. Bars represent one standard deviation of the means of the GA runs.

Fig. 4. Comparison of strength criterion (5) and ribosomal criterion (4) for confounded organisms between GA ($CAI_{opt} HEDB$) that determines CAI values by optimizing weights in a way that maximizes the location of HEDB genes and traditional CAI ($CAI HEDB$) using those same genes as a reference set. Each point represents an organism. GA data are the means of 10 runs. Bars represent one standard deviation of the means of the GA runs.

of particular interest. These tests indicate whether we should have statistical confidence that the direct search-based approach provides a better predictor of expressivity than the traditional tally based methods. The search-based

Fig. 5. Principal components analysis performed on strength criterion, ribosomal criterion, rank fitness, and Spearman rank correlation coefficient dimensions for confounded organisms. Reference sets were based upon experimentally determined high-expression genes (EDHE). There is clear separation between the GA (CAI_{opt}) and CAI groups.

Fig. 6. Principal components analysis performed on strength criterion, ribosomal criterion, rank fitness, and Spearman rank correlation coefficient dimensions for confounded organisms. Reference sets were based upon genes identified through the use of the high-expression database (HEDB). There is clear separation between the GA (CAI_{opt}) and CAI groups.

approach performs significantly better than traditional methods of determining CAI (Table 6) on confounded organisms. p -values of two-tailed, paired t -tests between search-based r_s values and traditional CAI r_s values are significant ($\alpha = 0.05$) for confounded organisms for both EDHE and HEDB gene reference sets. This is strong

TABLE 5
Performance Measurements for
Traditional CAI (Confounded Organisms)

Organism	CAI EDHE		CAI HEDB	
	r_s	RF	r_s	RF
<i>Campylobacter jejuni</i> NCTC 11168	-0.023	0.589	-0.004	0.745
<i>Halobacterium</i> sp. NRC-1	0.165	0.567	0.086	0.511
<i>Nostoc</i> sp. PCC 7120	0.046	0.608	0.150	0.721
<i>Pseudomonas aeruginosa</i> PAO1	0.347	0.730	0.303	0.658
<i>Streptomyces coelicolor</i> A3(2)	0.132	0.713	0.148	0.832

CAI EDHE: CAI determined using the traditional Sharp and Li approach using experimentally determined highly expressed genes as a reference set. CAI HEDB: CAI determined using the traditional Sharp and Li approach using HEDB genes as a reference set. r_s : Spearman rank correlation coefficient with experimentally determined expression data. RF: Rank fitness value (10). All Spearman rank correlation coefficients are significant ($p < 0.05$) with the exception of those for *C. jejuni*.

evidence that the direct search-based method is the preferred approach on confounded organisms.

3.4 Cross Validation

The direct-search methodology described here does not build a model that is then used on future genomes. Solutions for isolating translational efficiency identified by this approach are unique to each organism's genome. As such, the concept of a training and test set for use in cross validation is not applicable to this problem domain. However, the underlying purpose of cross validation is to reinforce confidence in the approach by showing evidence of consistent performance in multiple trials, and to show the absence of overfitting.

We achieve these same goals by validating our results against experimentally determined expression data (Tables 2 and 3) and then comparing the predictive ability of the search-based approach to traditional methods (Tables 4 and 5). Due to the stochastic nature of the search process, we validate consistency by executing multiple (10) trials per organism and then showing uniformity in the results across the runs (low standard deviations in Tables 2 and 3, and low standard deviations in Figs. 3 and 4). We do this for 10 different organisms (five of which are confounded by at least one competing bias, two by multiple biases, Table 1) and use two different approaches for identifying highly expressed genes (high-expression database and experimentally determined highly expressed genes). In addition to measuring the predictive capability of the search-based approach, we examine other attributes of the solutions to determine if they are consistent with that of translational efficiency bias (i.e., strength of bias and high bias adherence scores in ribosomal protein coding genes, Figs. 3 and 4).

To validate the similarity of solutions across multiple runs, analysis was performed on the resultant weight vectors. The variance in the weights across the 10 runs was calculated, and the outcomes averaged for the 10 different genomes, the two different maximized reference sets (HEDB and EDHE), and across all codons. The average variance in the weights was a remarkably low 0.021 (the range is from 0 to 1 for the weights). This is an indication that the solutions across runs are very similar. In almost all cases, the codons

identified as major (those with a weight of one) are the same across all 10 runs (variance of zero).

It is worth mentioning that trials were successfully run where randomly chosen subsets of genes (representing 90 percent of each genome) were used to identify optimal weights (fitness was correlated with actual expression values) (data not shown). These trials allowed for 10-fold cross validation; however, this approach is not applicable to the problem of predicting expression utilizing sequence information alone since expression data were required for the training set. The genetic algorithm described here requires no (and has no access to) prior knowledge of expression levels.

4 DISCUSSION

The direct search-based algorithm presented here directly seeks a set of codon preferences (weights) that yield high adherence scores in the genes known (EDHE) and presumed (HEDB) to be highly expressed (relative to the rest of the genes in the target genome). In this way, we identify a set of weights that may be more representative of the underlying tRNA abundance values, and therefore, the relative selective pressure associated with choosing each codon. These weights lead to more accurate predictions of expressivity. This approach is tested on two sets of microbial genomes, one known to be dominant for some confounding bias or biases (Table 1), and one dominant for translational efficiency bias.

Given the results of traditional CAI on the confounded organisms, one would conclude that these organisms exhibit, at best, weak translational efficiency bias (low ribosomal and strength criteria; Figs. 3 and 4). In the case of *C. jejuni*, one might even conclude that there is no translational efficiency bias at all (particularly in light of the negative, though nonsignificant, correlation with respect to expression levels; Table 5). For these organisms, however, the genetic algorithm finds that there is actually a strong translational efficiency bias present (ribosomal criteria > 1 and strength criteria > 8 ; Figs. 3 and 4). A possible cause is that the translational efficiency bias is obscured from detection when using traditional methods by other biases and by nonuniform adherence in the most highly expressed genes. The utility of these findings is supported by the expression level predictive capability of the direct search-based approach. All five of the confounded organisms are easily cultivated and exhibit exponential growth phases making the existence of translational efficiency bias unsurprising.

The search-based method performs significantly better than traditional methods of determining CAI (Table 6) on confounded organisms. p -values of two-tailed, paired t -tests between GA r_s values and traditional CAI r_s values are significant ($\alpha = 0.05$) for confounded organisms for both EDHE and HEDB gene reference sets. For unconfounded organisms, the search-based approach performs similarly to the traditional CAI method (p -values of two-tailed, paired t -tests between GA r_s values and traditional CAI r_s values are 0.924 and 0.981 for EDHE and HEDB reference sets, respectively).

In addition to good performance in predicting expressivity, the direct search-based approach provides solutions that exhibit traits generally known to be shared by genomes dominant for translational efficiency bias [17]. These traits

TABLE 6

p-Values for t-Tests between Criteria of GA and Traditional CAI

	EDHE			HEDB		
	r_s	S	RC	r_s	S	RC
Un-Confounded Organisms	0.924	0.004	0.008	0.981	0.000	0.000
Confounded Organisms	0.035	0.000	0.001	0.050	0.000	0.001

EDHE: Experimentally determined expression data used to identify reference set. HEDB: High-expression database used to identify reference set. r_s : Spearman rank correlation coefficient between CAI data and expression data. S: Strength criterion (5). RC: Ribosomal criterion (4).

include the strength of bias exhibited by the reference set, and the degree to which ribosomal protein coding genes adhere to the translational efficiency bias. Both tend to be high in organisms dominant for translational efficiency bias and are measured by strength and ribosomal criteria, respectively [ribosomal criterion > 1 (4) and strength > 8 (5)]. The strength and ribosomal criteria values for the biases identified by the search-based approach are significantly greater than those of the traditional methods for all organisms, confounded and unconfounded (*p*-values of two-tailed, paired t-tests between GA strength and ribosomal criterion values and traditional CAI strength and ribosomal criterion values are all less than 0.05, Table 6).

All of these factors taken together [better predictive capabilities (more positive r_s values), stronger identified bias, and a gene ranking with higher ribosomal criteria] lend support for the conclusion that the direct search-based approach provides a bias identification methodology that is more indicative of the underlying translational efficiency bias than previous methods. This is true for confounded organisms, with performance equivalent to traditional methods for unconfounded organisms, whether experimentally determined high-expression genes are used as a reference set, or genes identified through the use of the high-expression database are employed. The use of the search-based method may be especially advantageous in organisms such as *C. jejuni* that exhibit little or no translational efficiency bias using traditional methods. Use of the direct search-based approach reveals the existence of a strong translational efficiency bias in the genome of *C. jejuni* (ribosomal criterion $\mu = 2.972$, $\sigma = 0.206$ and $\mu = 2.668$, $\sigma = 0.333$ for EDHE and HEDB reference sets, respectively, Figs. 3 and 4) with a significantly positive Spearman rank correlation between the GA results and experimentally determined expression data (Table 3).

All coding sequences, GA-derived weights, and GA-derived CAI values for all coding sequences (for all organisms for all runs) can be viewed or downloaded at this paper's supplementary data Web site (http://web-dev.cs.umd.edu/cub_db/supplemental.shtml), and which can also be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.24>.

ACKNOWLEDGMENTS

This work was supported by a Research Initiation Grant from the Wright State University Research Council to Michael L. Raymer (grant no. 282026).

REFERENCES

- [1] M. Nei, "Molecular Population Genetics and Evolution," *Frontiers of Biology*, vol. 40, pp. 26-28, 1975.
- [2] 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.
- [3] 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.
- [4] 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&pubmed%did=6760125>, 1982.
- [5] T. Ikemura, "Correlation between the Abundance of *Escherichia Coli* Transfer rnas and the Occurrence of the Respective Codons in Its Protein Genes: A Proposal for a Synonymous Codon Choice that Is Optimal for the *E. Coli* Translational System," *J. Molecular Biology*, vol. 151, pp. 389-409, 1981.
- [6] P.M. Sharp and W.H. Li, "An Evolutionary Perspective on Synonymous Codon Usage in Unicellular Organisms," *J. Molecular Evolution*, vol. 24, nos. 1/2, pp. 28-38, 1986.
- [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] 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, 2003.
- [9] H. Akashi and T. Gajbordi, "Metabolic Efficiency and Amino Acid Composition in the Proteomes of *Escherichia Coli* and *Bacillus Subtilis*," *Proc. Nat'l Academy of Sciences USA*, vol. 99, no. 6, pp. 3695-3700, Mar. 2002.
- [10] J. Heizer, M. Esley, I. Raiford, W. Douglas, M.L. Raymer, T.E. Doom, R.V. Miller, and D.E. Krane, "Amino Acid Cost and Codon-Usage Biases in 6 Prokaryotic Genomes: A Whole-Genome Analysis," *Molecular Biology and Evolution*, vol. 23, no. 9, pp. 1670-1680, <http://mbe.oxfordjournals.org/cgi/content/abstract/23/9/1670>, 2006.
- [11] H. Seligmann, "Cost-Minimization of Amino Acid Usage," *J. Molecular Evolution*, vol. 56, no. 2, pp. 151-161, 2003.
- [12] J.R. Lobry and C. Gautier, "Hydrophobicity, Expressivity and Aromaticity Are the Major Trends of No-Acid Usage in 999 *Escherichia Coli* Chromosome-Encoded Genes," *Nucleic Acids Research*, vol. 22, pp. 3174-3180, 1994.
- [13] A. Grote, K. Hiller, M. Scheer, R. Münch, B. Nörtemann, D.C. Hempel, and D. Jahn, "JCat: A Novel Tool to Adapt Codon Usage of a Target Gene to Its Potential Expression Host," *Nucleic Acids Research*, vol. 33, pp. W526-W531, <http://dx.doi.org/10.1093/nar/gki376>, July 2005.
- [14] G. Hannig and S.C. Makrides, "Strategies for Optimizing Heterologous Protein Expression in *Escherichia Coli*," *Trends in Biotechnology*, vol. 16, no. 2, pp. 54-60, Feb. 1998.
- [15] M. Carrondo and M. Fussenegger, "Advances in Heterologous Protein Production," *J. Biotechnology*, vol. 120, no. 1, p. 55-64, <http://dx.doi.org/10.1016/j.jbiotec.2005.06.001>, Oct. 2005.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] M. Stenico, A.T. Lloyd, and P.M. Sharp, "Codon Usage in *Caenorhabditis Elegans*: Delineation of Translational Selection and Mutational Biases," *Nucleic Acids Research*, vol. 22, no. 13, pp. 2437-2446, July 1994.

- [20] G.E. Andersson and P.M. Sharp, "Codon Usage in the Mycobacterium Tuberculosis Complex," *Microbiology*, vol. 142, pp. 915-925, Apr. 1996.
- [21] 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.
- [22] D.W. Raiford, D.E. Krane, T.E. Doom, and M.L. Raymer, "Automated Isolation of Translational Efficiency Bias that Resists the Confounding Effect of GC(AT)-Content," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 7, no. 2, Apr.-June 2010, <http://doi.ieeecomputersociety.org/10.1109/TCBB.2008.65>.
- [23] 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.
- [24] J.H. Holland, *Adaptation in Natural and Artificial Systems*. Univ. of Michigan Press, 1975.
- [25] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.
- [26] D.E. Goldberg and K. Deb, "A Comparative Analysis of Selection Schemes Used in Genetic Algorithms," *Foundations of Genetic Algorithms (FOGA 1)*, G.J. Rawlins, ed., pp. 69-93, Morgan Kaufmann, 1991.
- [27] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, Nov. 1998.
- [28] Z. Michalewicz, G. Nazhiyath, and M. Michalewicz, "A Note on Usefulness of Geometrical Crossover for Numerical Optimization Problems," *Evolutionary Programming*, vol. 5, no. 1, pp. 305-312, 1996.
- [29] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, pp. 671-680, citeseer.ist.psu.edu/kirkpatrick83optimization.html, May 1983.
- [30] V. Cerny, "A Thermodynamical Approach to the Travelling Salesman Problem: An Efficient Simulation Algorithm," *J. Optimization Theory and Applications*, vol. 45, pp. 41-51, 1985.
- [31] T. Bäck, F. Hoffmeister, and H.-P. Schwefel, "A Survey of Evolution Strategies," *Proc. Fourth Int'l Conf. Genetic Algorithms (ICGA)*, R.K. Belew and L.B. Booker, eds., pp. 2-9, 1991.
- [32] K.A. De Jong, "An Analysis of the Behavior of a Class of Genetic Adaptive Systems," PhD dissertation, Univ. of Michigan, 1975.
- [33] J.E. Baker, "Reducing Bias and Inefficiency in the Selection Algorithm," *Proc. Second Int'l Conf. Genetic Algorithms and Their Application (ICGA2)*, pp. 14-21, 1987.
- [34] 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, pp. D760-D765, <http://dx.doi.org/10.1093/nar/gkl887>, Jan. 2007.
- [35] R. Wüschiers and H. Eckes, "HyDaBa Hydrogen Database: Nostoc," <http://www.hydaba.uni-koeln.de/index.php>, Apr. 2005.
- [36] E. Cantú-Paz and D. Goldberg, "Efficient Parallel Genetic Algorithms: Theory and Practice," *Computer Methods in Applied Mechanics and Eng.*, vol. 186, nos. 2-4, pp. 221-238, 2000.
- [37] 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.
- [38] 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.
- [39] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *J. Educational Psychology*, vol. 24, pp. 417-441, 1933.
- [40] I.T. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 1986.

soula, Montana, and is an associate member of the UM Biochemistry program's faculty.

areas of molecular evolution and population genetics. Since 1991, he has also 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.

Sciences PhD program faculty and the codirector of WSU's Bioinformatics Research Group. His research interests include design automation, computational biology, optimization theory, and engineering education. He currently serves as the 2007-2008 chair of the IEEE Technical Committee on Bioinformatics. He is a senior member of the IEEE.

faculty and the codirector of WSU's Bioinformatics Research Group. His research interests include protein structure and function, bioinformatics, evolutionary computation, and pattern recognition. He is the coauthor of one of the first bioinformatics textbooks developed for undergraduate students. He is member of the IEEE.

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. His research interests include bioinformatics, investigating global forces involved in genomic and proteomic evolution, codon usage bias, evolutionary computation, and pattern recognition. He is currently an assistant professor in the Department of Computer Science, University of Montana, Missoula, Montana, and is an associate member of the UM Biochemistry program's faculty.

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 the Cell and Molecular Biology Department at Pennsylvania State University in 1990. He pursued postdoctoral research at Washington University and Harvard University before accepting a faculty appointment at Wright State University, Dayton, Ohio, in 1993. His research interests are primarily in the

Travis E.W. Doom (M'98-SM'03) received the BS degree in computer science and mathematics from Bowling Green State University, 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. He is an associate member of WSU's Biomedical

Michael L. Raymer (M'03) 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 fa-

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.