Bioinformatics Advance Access published November 29, 2014

Genome analysis

Exploiting Hidden Information Interleaved in the Redundancy of the Genetic Code without Prior Knowledge

Hadas Zur[1,2] and Tamir Tuller[1,3,*]

[1]Department of Biomedical Engineering, the Engineering Faculty, Tel Aviv University.

[2]Blavatnik School of Computer Science, Faculty of Exact Sciences, Tel Aviv University.

[3]Sagol School of Neuroscience, Tel Aviv University, Tel-Aviv 69978, Israel.

Associate Editor: Dr. John Hancock

ABSTRACT

Motivation: Dozens of studies in recent years have demonstrated that codon usage encodes various aspects related to all stages of gene expression regulation. When relevant high quality large scale gene expression data is available it is possible to statistically infer and model these signals, enabling analysing and engineering gene expression. However, when these data are not available it is impossible to infer and validate such models.

Results: In the current study we suggest *Chimera* - an unsupervised computationally efficient approach for exploiting hidden high dimensional information related to the way gene expression is encoded in the ORF, based solely on the genome of the analysed organism.

One version of the approach, named *Chimera Average Repetitive Substring* (*ChimeraARS*), estimates the adaptability of an ORF to the intracellular gene expression machinery of a genome (host), by computing its tendency to include long substrings that appear in its coding sequences; the second version, named *ChimeraMap*, engineers the codons of a protein such that it will include long substrings of codons that appear in the host coding sequences, improving its adaptation to a new host's gene expression machinery.

We demonstrate the applicability of the new approach for analyzing and engineering heterologous genes and for analyzing endogenous genes. Specifically, focusing on *E. coli*, we show that it can exploit information that cannot be detected by conventional approaches (e.g. the CAI - Codon Adaptation Index), which only consider single codon distributions; for example, we report correlations of up to 0.67 for the *ChimeraARS* measure with heterologous gene expression, when the CAI yielded no correlation.

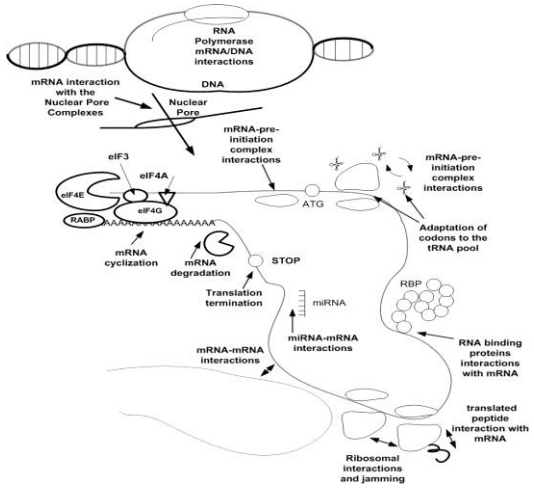
Contact: tamirtul@post.tau.ac.il (TT)

Availability: For non-commercial purposes, the code of the *Chimera* approach can be downloaded from

http://www.cs.tau.ac.il/~tamirtul/Chimera/download.htm

1 INTRODUCTION

The inherent redundancy of the genetic code, where 61 codons encode only 20 amino acids, is a widely studied phenomenon (Chamary, Parmley et al. 2006; Plotkin and Kudla 2010; Sauna and Kimchi-Sarfaty 2013). In recent years it has been shown that various gene expression regulatory aspects are interleaved in this redundancy. Specifically, during its lifetime the mRNA sequence interacts with various intracellular molecules and complexes such as the spliceosome (Cartegni, Chew et al. 2002), pre-initiation complex (Kozak 1986; Zur and Tuller 2013), ribosomes (Ramakrishnan 2002) and ribosomal RNA composing it (Li, Oh et al. 2012), tRNAs (Alberts, Johnson et al. 2002), miRNAs (Forman and Coller 2010), other mRNAs (Zur and Tuller 2012), proteins (Hogan, Riordan et al. 2008) (including transcription factors (Stergachis, Haugen et al. 2013)), and the mRNA sequence itself via its folding (Gu, Zhou et al. 2010; Tuller, Veksler-Lublinsky et al. 2011) (see Figure 1). The affinity of these interactions is affected by the nucleotide composition in various parts of the transcript (see for example (Kozak 1986; Alberts, Johnson et al. 2002; Chamary, Parmley et al. 2006; Hogan, Riordan et al. 2008; Kudla, Murray et al. 2009; Cannarozzi, Schraudolph et al. 2010; Forman and Coller 2010; Gu, Zhou et al. 2010; Plotkin and Kudla 2010; Schnall-Levin, Zhao et al. 2010; Tuller, Carmi et al. 2010; Li, Oh et al. 2012; Zur and Tuller 2012; Stergachis, Haugen et al. 2013; Zur and Tuller 2013)), and can usually be described by Markovian models and/or position specific scoring matrices (PSSMs) (Pevsner 2009). However, there are debates regarding the nature and the efficiency of some of these interactions (see for example (Plotkin and Kudla 2010)).



*To whom correspondence should be addressed.

Fig. 1. *Illustration of the various macro-molecules that interact with the ORF and the regulatory signals interleaved in the genetic code (see explanation in the main text).*

The studies mentioned above support the conjecture that it is possible to accurately predict gene expression aspects based solely on the coding sequence, and that by manipulating synonymous aspects of the coding sequence itself it is possible to affect all gene expression stages. There are three major drawbacks to the current measures for inferring gene expression based on the coding sequence (Open Reading Frame; ORF):

First, most of the conventional methods for estimating the adaptiveness of a transcript to the gene expression regulatory machinery are based on the independent distribution of single codons in the coding sequence (*i.e.* single codon usage bias). While there are dozens of measures based on single codon distributions (Sharp and Li 1987; dos Reis, Savva et al. 2004), these methods often fail to exhibit meaningful relations with the (usually heterologous) expression levels of genes (see, for example, (Kudla, Murray et al. 2009; Goodman, Church et al. 2013)).

It is clear that such indexes cannot fully capture all the gene expression information encoded in the ORF as some of it is not directly related to codon decoding. For example, the binding site of micro-RNAs is around 22 nt, information not fully described by the independent distribution of single codons. Thus, a simple approach that can capture various aspects of gene expression regulation (see Figure 1) is needed; we term this type of statistical information 'high dimensional information' as it is related to substrings of nucleotides longer than codons. This type of statistic has the potential to encapsulate all known and unknown intracellular interactions.

Second, currently most of the models related to the biophysical nature of the various interactions of macro-molecules with the ORF are based on the analyses of gene expression measurements (*e.g.* mRNA levels, protein levels, ribosomal densities, etc) (dos Reis, Savva et al. 2004; Vogel, Abreu Rde et al. 2010; Lee, Topper et al. 2011; Reuveni, Meilijson et al. 2011; Schwanhausser, Busse et al. 2011; Zur and Tuller 2013); specifically, this information can be used for inferring the parameters of the simulative and/or predictive gene expression models. However, today there are around 26,000 genomes of different organisms (http://www.ncbi.nlm.nih.gov/genome), but large scale gene expression information (*e.g.* protein abundance measurements (PA)) is available for only a few dozen (Wang, Weiss et al. 2012). Thus, we aim to develop an unsupervised measure that is based solely on the genome of the analyzed organism, without the necessity of additional gene expression measurements.

Third, there is a growing surge of new studies reporting novel rules related to the way aspects of gene expression are encoded in the transcript. One relatively simple rule suggests that the distribution of codon-pairs affects gene expression and the fitness of organisms and viruses (see, for example, (Irwin, Heck et al. 1995; Coleman, Papamichail et al. 2008; Tats, Tenson et al. 2008)); however, many of these rules are more complicated and complex (Plotkin and Kudla 2010; Li, Oh et al. 2012; Stergachis, Haugen et al. 2013; Zur and Tuller 2013) (see also Supplementary Note 1.1). Thus, it is apparent that numerous additional rules are yet to be deciphered. Moreover, many of the established rules are organism specific and/or condition/tissue specific, and may not hold in different organisms/conditions than the ones used for their inference. Thus, we aim at developing a measure that can exploit "hidden" (*i.e.* unexplored) gene expression information encoded in the ORF, which may be related to yet unknown gene expression rules.

In the following sections we describe an approach which encapsulates all the three points mentioned above. Additionally, we devise a novel approach for engineering genes for heterologous gene expression (see, for example, (Vervoort, Ravestein et al. 2000; Gustafsson, Govindarajan et al. 2004; Plotkin and Kudla 2010; Goodman, Church et al. 2013)) based on the aforementioned concepts. In the Methods (and Supplementary Methods) we will describe our new computationally efficient approach for exploiting hidden high dimensional information interleaved in the redundancy of the genetic code without prior knowledge. Since our Method is based on engineering new genes, or analyzing genes based on patterns that appear in different endogenous genes, we named it *Chimera*, which is a mythological creature composed of three different animals.

In the Results section, we will show that indeed such high dimensional information appears in the coding regions of the analysed organisms, and that we can at least partially infer it with our approach. As a model organism we analyse *E. coli* which is the only organism with large scale measurements of heterologous gene expression.

2 METHODS

Inspired by universal approaches for data compression without any prior knowledge of its statistical characteristics (Ziv and Lempel 1977; Ulitsky, Burstein et al. 2006); we suggest the *Chimera* approach. Generally, the approach is based on the idea that various aspects of gene expression (mentioned above) are encoded in the ORF; thus, these "codes" (information) are frequently repeated in the coding sequences of the organism; in addition we expect to see more of these "codes" in genes (both heterologous and endogenous) that are highly expressed and/or more tightly regulated. Furthermore, based on this idea we can optimize the expression levels of a heterologous gene by engineering its codons (substrings of codons and not only single ones) such that they will be similar to the ones that appear in the endogenous genes of the host. This approach can be extrapolated to many variants, two of which we will consider in the current study: 1) A new measure for the adaptation of the coding sequence to the intracellular gene expression regulatory machinery named *Chimera Average Repetitive Substring* or *ChimeraARS*. 2) A new algorithm for engineering heterologous genes without prior knowledge and based only on the genome of the host named *ChimeraMap*. In this subsection we briefly describe *ChimeraARS* and *ChimeraMap*.

The *ChimeraARS* is depicted in Figure 2**A**. A given gene which codes a protein, *P*, can be described as a sequence of codons, *S*; thus, the new measure is based on the tendency of substrings in *S* to appear in other genes, *i.e.* in a reference set *G*; it is important to mention that various definitions of *G* are possible, including considering only highly expressed genes; for simplicity and demonstrating the unsupervised advantage of our approach, we assume that *G* includes the entire genome. The measure is based on the assumption that evolution shapes the organismal coding sequences to improve their interaction with the intra-cellular gene expression machinery. Thus, if longer substrings of *S* tend to appear in the organism's ORFs, it suggests that *P* is more optimized to the intra-cellular gene expression machinery, and thus it is probably more highly expressed. In addition, as we explain in the following subsection, our measure also has important statistical and information theoretic properties. Computing the

ChimeraARS score, *ChimeraARS(G,S)*, of a coding sequence (*S*) given a reference genome (*G*) includes the following steps (Figure 2**A**; further details in the following subsections):

1) For each position *i* in the coding sequence *S* find the *longest* substring S_i^j that starts in that position, and also appears in at least one of the coding sequences of the genome *G*.

2) Let $|S|$ denote the length of a sequence *S*; the *ChimeraARS* of *S* is the mean length of all the substrings $S_i^j : \Sigma S_i^j / |S|$.

As we demonstrate in the ensuing subsections, the *ChimeraARS* exploits information that does not appear in single codon distributions. Thus, among others, it can be used for estimating the adaptation of the composite codon content of a gene to the cellular gene expression machinery; since highly expressed genes are expected to be more adapted, it can specifically be used for predicting the protein levels of a gene from its codon distribution, while considering the high dimensional distribution of codons.

The objective of the *ChimeraMap* algorithm is described in Figure 2B (further details in the following subsections). Given a target protein, *P*, whose coding sequence is *S*, and a host genome (*G*), *ChimeraMap* finds a new coding sequence (*S**) that codes the protein *P* but is composed of a *minimal* number of (non-overlapping) 'codon blocks' that appear in the host genome (Figure 2B). If several blocks of the same length exist, we select the most frequent one, thus further improving the adaptability of *S** to *G*.

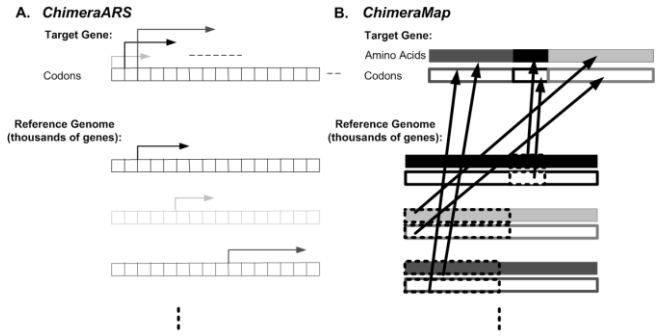


Fig. 2. An illustration of the *ChimeraARS* measure: To compute the *ChimeraARS* measure for a certain ORF we find for each codon (or nt) position in the ORF the longest substring that starts in this position, but also appears in one of the ORFs of the genome. The score is based on the average over the lengths of all these substrings. B. An illustration of the *ChimeraMap* measure. The objective function of this algorithm is covering the coding sequence of the target gene with a minimal number of most frequent codon blocks that appear in the host genome, such that the protein encoded in the resulting coding sequence is identical to the protein encoded in the target gene. Under this construction, the boundaries between blocks are the only regions with codon sequences that may not appear in the host genome; the *ChimeraMap* minimizes these regions by minimizing the number of blocks.

We believe that the *ChimeraMap* can be used for optimizing the coding sequences of heterologous genes for expressing them in a new host. It is easy to see that under this construction, the boundaries between blocks are the only regions with codon sequences that may not appear in the host genome; the *ChimeraMap* minimizes these regions by minimizing the number of 'codon blocks'. Since the 'codon blocks' already appear in the host genome they are expected to be compatible with the host gene expression machinery; the boundaries between blocks, on the other hand, correspond to substrings that do not appear in the host genome, thus they may not be compatible with the host gene expression machinery and *ChimeraMap* minimizes them.

2.1 Properties of the *ChimeraMap* and the *ChimeraARS* approaches

All the details regarding the algorithms for the computation of the *ChimeraARS* score and the *ChimeraMap* output appear in the following subsections. Here we only briefly outline the algorithmic approach and mention the general properties of these algorithms.

ChimeraARS is based on a pre-processing step of generating a suffix tree (or array) of all the coding sequences of the reference genome (*G*); based on this suffix tree all the longest substrings of the target *P* can be computed in an efficient manner, resulting in a total running time complexity of $O(|G|+|P|)$. The *ChimeraMap* is also based on the same pre-processing step in addition to a dynamic programming algorithm that finds the optimal solution with total running time complexity of $O(|G|+|P|)$.

ChimeraARS (and thus the *ChimeraMap* objective) is inspired by information theoretic approaches for universal compression of Markovian sequences, and estimating the number of bits required for describing one sequence (*S*) given a second one (*G*) (Ziv and Lempel 1977; Wyner 1993; Farach, Noordewier et al. 1995; Wyner and Wyner 1995; Ulitsky, Burstein et al. 2006). More specifically and formally, let x^n denote a codon sequence of length *n*. Specifically, if the codon distribution in *G* and *S* are generated by Markovian processes with probability distributions M_S and M_G ($M_S(x^n)/M_G(x^n)$ is the probability of emitting x^n based on the Markovian model M_S /M_G respectively), the *ChimeraARS(G,S)* estimates the following measure (see proofs and/or explanations in (Wyner 1993; Ulitsky, Burstein et al. 2006):

$$\log(G) \Big/ -\mathrm{E}_{M_S}\log(M_G) \qquad (1)$$

$$-\mathrm{E}_{M_S}\log(M_G) = \lim_{n\to\infty} \sum_{x^n} M_S(x^n)\cdot\log(1/M_G(x^n)) \qquad (2)$$

If the distribution of *S* and *G* are similar, *S* can be better compressed by *G*. If $M_s = M_G$ the *ChimeraARS(G,S)* (equation (1)) converges to $\log(|G|)/H(M_S)$ where $H(M_S)$ is the entropy of M_S and it is known that $H(M_S)$ is smaller than $-\mathrm{E}_{M_S}\log(M_G)$ (equation (2)) for $M_s \neq M_G$.

Finally, by definition, genes designed according to the *ChimeraMap* algorithm should have higher *ChimeraARS* scores: *ChimeraMap* engineers the coding sequence of the target gene such that it will include long substrings that appear in a reference genome, while the *ChimeraARS* measure detects the tendency of a coding sequence to include long substrings that appear in a reference genome, and thus its adaptability to the genome's gene expression machinery.

2.2 The *Chimera* Algorithms

1. The algorithm of the first version, *ChimeraARS*: The preprocessing step of the algorithm is based on building a suffix tree (or suffix array) (Manber and Myers 1993; Gusfield 1997) for the coding sequences of the host genome. This can be done in O(|*G*|) where |*G*| is the length of the proteome/coding sequences of the host (Manber and Myers 1993; Farach 1997). We will discuss the complexity of a suffix tree implementation, though due to the genomes' size in practice sometimes a suffix array implementation is advisable if space considerations are more critical than time.

Then, the length of the longest substring starting at each position in the target gene that appears in the host genome can be found in an efficient manner in O(|*P*|)) (Gusfield 1997) (matching statistics algorithm, pp 132-134, this is achieved by building the suffix tree for *G* maintaining the suffix links (shortcuts between internal nodes related to substrings and their suffixes), which are then utilized together with the skip/count trick to shorten traversal time). Thus, the total time complexity of the algorithm is O(|*G*| + |*P*|)).

2. The algorithm of the second version, *ChimeraMap*: This is a dynamic programming (DP) algorithm that builds an optimized representation of a given protein *P*, maintaining the encoded protein, to that of a specified reference (host) genome *G*, by minimizing the number of (most frequent)

substrings from the reference genome required to cover it. This problem naively solved is essentially exponential. *ChimeraMap* reduces this to polynomial time based on the observation that the optimal solution can be greedily extended in each DP step. Similarly to the *ChimeraARS*, the preprocessing step of the algorithm is based on building a suffix tree (Gusfield 1997) for the coding sequences of the reference genome in time O(|*G*|).

The premise of the algorithm is that all that is necessary is to consider only the previous step in the DP optimal solution space. Thus utilizing dynamic programming, at each step *i* (in which the length *i* of the substring of *P* we are looking at grows by 1 from the previous step, with *i* being from 1 to |*P*|) we look at the previous step's (*i*-1) optimal solution and determine based on it what the optimal solution for step *i* is. The manner in which *ChimeraMap* tries to elongate the previous optimal solution to length *i* is as follows. Each optimal solution is represented by a list of pairs of numbers, symbolizing the start and end positions of the substrings (blocks) covering *P* up to that point (Figure 2B). *ChimeraMap* looks at the last such block ([*start end*]) and tries to find a match in the suffix tree of *G* for *P*(*start*:i), and if that fails for *P*(*end*+1:i). If we maintain a pointer to the end of solution *i*-1 in the suffix tree, then we can check in O(1) if there is a match for *P*(*start*:i), by simply continuing the down traversal on the current edge to the next character. If a match is found we are done, otherwise a pointer is kept to the root of the tree and we can begin the search for *P*(*end*+1:i) in constant time. Thus, we get a total time complexity of O(|*G*| + |*P*|). One interesting observation regarding the properties of the *ChimeraMap* is that there exists a maximal length for a 'covering' substring from a reference genome. Let *lcsMAX* denote the length of the longest substring that is common to *P* and *G*. *lcsMAX* can be calculated using the common suffix tree of *P* that is found in *G* and all its prefixes. If the sequences *G* and *P* are generated by a Markov process O(*lcsMAX*) is expected to be of the order of O(log|*G*|) (see equation (1)); in the analyzed organism |*G*| equals 3,958,573, the log of which is 15.2, while the mean *lcsMAX* is 15.3 with an *STD* of 38.

2.3 The Optimality of the *ChimeraMap*

Here we prove the dynamic programming algorithm for the second version of the *Chimera* approach, the *ChimeraMap*, indeed finds an optimal solution with induction, *n*=|*P*|.

The base case: *i*=1:

The algorithm initiates with the first character of *P*. The first solution is therefore the block [1,1], and it is trivially optimal.

Inductive step: assume optimality for *i*=*n*-1:

We assume we have optimal solutions for all the algorithm steps up to *n*-1.

Prove optimality for *i*=*n*:

The algorithm at each step looks back at the previous solution, and elongates it according to the following rule: look at the last mapped block ([*start end*]) of the solution and try to find a match in the suffix tree for *P*(*start*:*i*), and if that fails for *P*(*end*+1:*i*). If a match is found, merge to the current previous solution in the appropriate manner. For each solution, we either elongate its last block, thus not increasing the number of substrings covering *P*, and since that solution was optimal, so is our solution for *i*. If we cannot elongate an existing block, we will open a new one: [*end*+1 *i*], thus increasing the previous optimal solution by one. If an existing block could have been elongated, this solution is chosen as the optimal for step *i*. If not, the optimal solution grows by no more than one. Now assume that the optimal solution of step *i*=*n* includes *l* blocks. We assume that the algorithm found all the optimal solutions for *i* < *n* and show that it will find the optimal solution for *i*=*n*.

Since the optimal solution of *i*=*n* includes *l* blocks, the optimal solution for *i*=*n*-1 includes either *l*-1 or *l* blocks. If it is of length *l*-1, then no extension of the last block exists, and the algorithm adds a new block to the *i*th solution. If it is of length *l* blocks, this means that an extension exists. Let us assume by negation that the last block of the solution *i*=*n*-1 is of length *w* and represents the string α, and cannot be extended. Thus, since according to our assumption the optimal solution for the *i*th step includes *l* blocks there must be a block representing the string β of length > *w* that can be extended. But α is a suffix of β and therefore can also be extended, contradicting our negation assumption.

2.4 Additional information

Due to lack of space additional information related to the *Chimera* approach, datasets analysed, statistical analysis, and the CAI (Sharp and Li 1987) appear in the Supplementary Methods.

3 RESULTS

The following analysis further demonstrates the relation described above between the *ChimeraMap* algorithm and the *ChimeraARS* scores. We uniformly selected 100 *E. coli* genes according to their PA levels, and created the following variations of them: 1. Performing 100 randomizations of these genes while maintaining the encoded protein, the amino acid bias, and the codon usage bias per gene (Supplementary Methods). 2. We optimized them according to the CAI rationale, replacing every synonymous codon with its most abundant version, which we termed MFSC (Most Frequent Synonymous Codon), a variation representing the encapsulation of single codon distribution. 3. We engineered them according to the *ChimeraMap*. The results of this analysis can be seen in Figure 3, where clearly and significantly the *ChimeraMap* engineered genes obtain the highest *ChimeraARS* scores compared to both the randomized and MFSC versions.

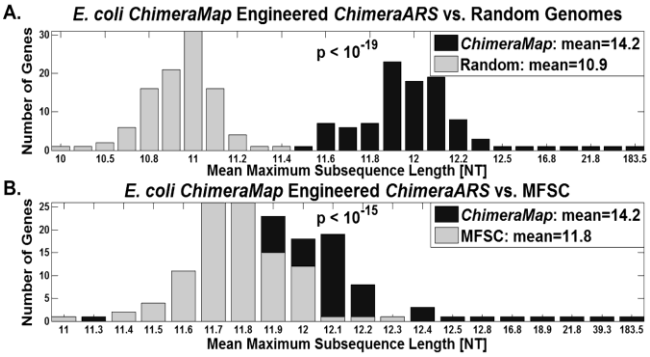


Fig. 3. A. The *ChimeraMap* engineered genes' *ChimeraARS* scores, as compared to those of the averaged 100 randomizations, of the *E. coli* subset of genes (14.2 vs. 10.9; performing a Wilcoxon signed rank test we received a p-value < 10^{-19}). B. The *ChimeraMap* engineered genes' *ChimeraARS* scores, as compared to those of the MFSC version (14.2 vs. 11.8; performing a Wilcoxon signed rank test we received a p-value < 10^{-15}).

3.1 High dimensional information is encoded in the codon usage bias and can be exploited by the *Chimera* approach

In the current section we show that high dimensional information appears in the coding sequences of organisms via several tests. This was achieved by comparing the *ChimeraARS* scores of endogenous *E. coli* genes to the ones obtained for randomized genomes that maintain the protein content and frequencies of single codons. If indeed patterns of substrings of codons (longer than

one) tend to repeat in the endogenous genome more than expected by chance, the *ChimeraARS* scores will tend to be higher in the real genome in comparison to the randomized genome.

Specifically, in order to compute the *ChimeraARS* measure for endogenous genes, for each gene instead of using the entire genome, we considered all the genes excluding the current as the reference genome. First, we computed the *ChimeraARS* measure for the real and randomized *E. coli* genome; the randomized genome encoded the same protein and single codon frequencies in each gene as in the original *E. coli* genome; however, it did not include the same higher dimensional distributions (further details in the Supplementary Methods). For each gene, we calculated its *ChimeraARS* score, which is the mean over the maximum substring length of each of its codon/nucleotide positions, that can be found in all the other genome genes. The distributions of the *ChimeraARS* scores in the real *vs.* the randomized genome appear in Figure 4A, after removing paralogs in order to demonstrate that the relation can't be attributed to sequence similarity among paralogs (see Supplementary Figure S1 for an analysis including all the genes). As can be seen, the *ChimerARS* scores are significantly higher in the real genome (16.7 *vs.* 11.1; $p = 10^{-454}$). This result supports the conjecture that long substrings of codons/nucleotides tend to appear in the coding sequences of the analysed organism more than expected by chance; thus, this result supports the hypothesis that at least some of the repetitive codon substrings affect the fitness of *E. coli*. The analyses performed in the next subsections support the conjecture that this high dimensional information is probably related at least partially to gene expression regulation, as *ChimeraARS* scores correlate with the expression levels of endogenous and heterologous genes.

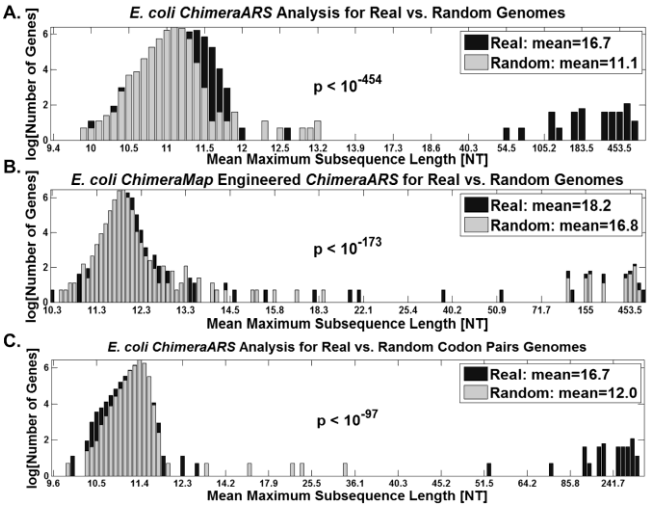


Fig. 4. A. *ChimeraARS* scores for the real and randomized *E. coli* genome. The mean *ChimeraARS* score for the real genome was significantly higher than the random (16.7 vs. 11.1). Performing a Wilcoxon signed rank test we received a p-value < 10^{-454}. B. *ChimeraARS* scores for the *E. coli* real and random genome, as engineered by the *ChimeraMap* algorithm. The mean *ChimeraARS* score for the *ChimeraMap* engineered real genome was significantly higher than that of the engineered random genome (18.2 *vs.* 16.8; performing a Wilcoxon signed rank test we received a p-value < 10^{-173}); the increase in the *ChimeraARS* scores relatively to A. is due to the fact that by definition, *ChimeraMap* genes that are engineered based on this algorithm tend to include longer repetitive substrings that appear in the host genomic coding sequences, and thus result in higher *ChimeraARS scores*; this phenomenon can be seen in the random genome as well, further demonstrating *ChimeraMap*'s ability to engineer genes. In both analyses paralogs were removed in-order to show that the reported signal can't be attributed to sequence similarity among paralogs. C. *ChimeraARS* scores for the real and randomized *E. coli* genome which maintains the codon pairs distribution, in addition to the single codon distribution and encoded protein (additional details appear in the Supplementary Methods). The mean *ChimeraARS* score for the real genome was significantly higher than the random (16.7 vs. 12). Performing a Wilcoxon signed rank test we received a p-value < 10^{-97}.

We performed an additional validation where we wanted to verify that the engineered genes which the *ChimeraMap* algorithm produces maintain the noted higher length distribution, and that it is higher than the one obtained for randomized genomes also in this case. To this end, taking each gene as a target, we built its *ChimeraMap* version, which encodes the same protein, but is composed of the maximal most frequent substrings in all the other genes of the genome (excluding the current gene). We performed this for the real and randomized *E. coli* genome respectively. As can be seen in Figure 4B, indeed the *ChimeraMap* engineered genes have higher *ChimeraARS* scores in the real genome (18.2 vs. 16.8; $p = 10^{-173}$). This result further substantiates the conjecture that long substrings of codons do tend to appear in the coding sequences of the analysed organism more than expected by chance, and that the *ChimeraMap* algorithm can exploit this information.

3.2 Measures based on the *Chimera* approach correlate with various aspects of gene expression and include information that does not appear in conventional codon usage bias measures

In the previous section we showed that long substrings of codons/nucleotides tend to repeat in the coding sequences of *E. coli* more than expected by chance. In the current section we will show that the repeated substrings, and thus the *ChimeraARS* score, are related to the expression levels of endogenous genes. To this end, we compared the correlation obtained between the CAI (Sharp and Li 1987) (a measure based on the *independent* distribution of *single* codons; see details in the Supplementary Methods), and measurements related to various gene expression aspects/stages (mRNA levels, ribosomal density, and protein levels), to the one obtained based on a regressor of both the CAI and *ChimeraARS* (the analysis was based on cross validation and control for the number of features in the regressor/predictor; more details in the Supplementary Methods). As can be seen in Figure 5, the correlation with gene expression indeed increases when adding the *ChimeraARS* feature relatively to regression based on the CAI alone, also when controlling for the number of features by computing adjusted correlations (Supplementary Methods, see Figure S2 for the correlations achieved for each measure separately); the result supports the conjecture that the *ChimeraARS* infers information related to expression levels which cannot be detected by conventional approaches such as the CAI; thus, information related to gene expression regulation is encoded in high-dimensional distributions of codons and nucleotides in the coding sequence.

3.3 Analyses of heterologous gene expression by the *Chimera* approach demonstrate its advantages over conventional codon usage bias measures

Goodman *et al.*(Goodman, Church et al. 2013) recently designed a heterologous gene library utilizing the first 11 amino acids including the initiating methionine from 137 essential genes in *E. coli.* They generated 13 variants of each gene, where they changed the

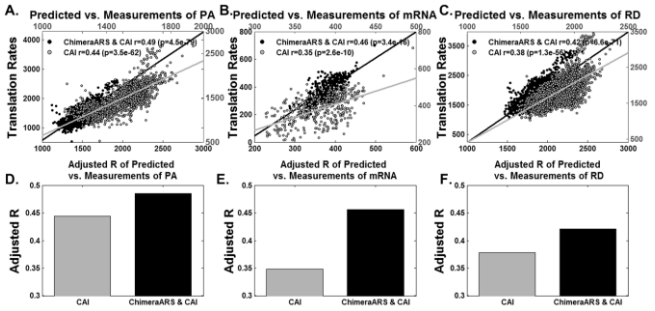


Fig. 5. Dot plots (A. – C.) and Adjusted Spearman correlations (D. – F., Supplementary Methods) of the prediction of (1) a regression model which is based only on the CAI (gray) and (2) a model which is based on the CAI and *ChimeraARS* (pale blue), vs. measured protein abundance (A,D), mRNA levels (B,E), and ribosomal density (C,F), respectively.

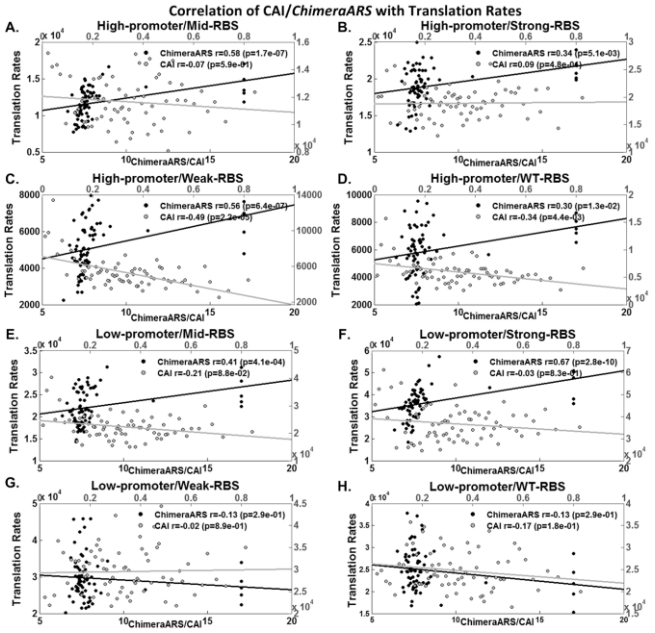


Fig. 6. The Spearman correlation between the *ChimeraARS* score and the CAI respectively with the Goodman *et al.* translation rates, according to their promoter (High/Low), and RBS (ribosomal binding site; Strong/Weak/Mid/WT) gene groups. The figures include linear regression lines, though the relations are monotone but clearly not linear.

synonymous codons used to encode the peptide, always keeping the start codon as ATG. Using two promoters, and four RBSs (Ribosome Binding Site) they generated 14,234 heterologous gene sequences, and measured their translation rates. In their paper Goodman *et al.* show that there is no correlation between the CAI (Sharp and Li 1987) and the translation rate. Here we show that the *ChimeraARS actually correlates with* the translation rates of the Goodman *et al.* (Goodman, Church et al. 2013) experiment.

Analyzing the heterologous *E. coli* data of Goodman *et al.* (Goodman, Church et al. 2013) using as a reference genome the first 11 codons of each endogenous *E. coli* gene (such that it will correspond to the first 11 codons that were modified in the heterologous gene library), we calculate the *ChimeraARS* score for each of the heterologous 11 codon long coding sequences. We calculated the correlation of the Goodman *et al.* translation rates with the *ChimeraARS* score, and compared these to the ones achieved for the CAI. As can be seen in Figure 6, while the CAI does not correlate with the translation rates of the heterologous gene library (correlation between -0.34 and 0.09; and mostly not significant; where the negative correlation is in the *wrong* direction), the *ChimeraARS* does (correlation mostly between 0.3 and 0.67; all $p < 0.02$,). This result demonstrates that the *ChimeraARS* can detect the expression of genes also in heterologous systems; in addition, it supports the conjecture that the relation between the *ChimeraARS* score and expression levels reported in the previous section is at least partially *causal/direct* – higher *ChimeraARS* scores contribute to higher expression levels (since these are heterologous and not endogenous genes). Finally, since by definition (as explained above) genes designed by the *ChimeraMap* algorithm have higher *ChimeraARS* scores, these results support the conjecture that the *ChimeraMap* algorithm should be able to optimize expression levels of genes at least in the analysed organism.

4 DISCUSSION

We describe a novel computational approach named *Chimera* for exploiting high dimensional information related to gene expression that is interleaved in the redundancy of the genetic code, and for engineering coding regions of heterologous genes without prior knowledge. Our approach is inspired by an information theoretic technique for data compression, and is very efficient in terms of computational running time.

One version of the approach, *ChimeraARS*, can be used for estimating the amount of information related to gene expression encoded in a coding sequence, and thereby its adaptability to the cellular gene expression machinery; as we demonstrate here, this estimation is expected to correlate with the expression levels and/or gene expression regulation levels of a gene. In addition we show that the *ChimeraARS* exploits high dimensional information that is not included in indexes, such as the CAI, that are based on the distribution of single codons.

Furthermore, we suggest a version of the *Chimera* approach, *ChimeraMap*, that can be used for engineering new genes for their efficient expression in a new host. The *ChimeraMap* optimizes the coding sequence encoding a protein such that it will include as few as possible substrings of the host genome (*i.e.* longer substrings). We show that the output of the *ChimeraMap* correlates with the *ChimeraARS* score; in addition, we show that the *ChimeraARS* score predicts the expression levels of heterologous genes in *E. coli* well*;* thus, the *ChimeraMap* is expected to be a useful approach for heterologous coding sequence optimization.

More generally, the analyses reported in this study suggest that codon bias, if defined accurately, is useful in detecting highly expressed genes in cases where conventional approaches do not work, such as heterologous gene expression (Goodman, Church et al. 2013). In addition, we show that the protein levels of endogenous and heterologous genes can be defined by their codon and amino acid content based only on genomic information.

Naturally, our approach can be generalized in various ways. For example, here for simplicity (and demonstrating the unsupervised

advantage of our approach) the reference set of genes used for the *Chimera* approach included the entire genome; we can readily think of other relevant reference sets such as highly expressed genes, tissue specific genes, or genes with a certain function or property. Another variation is related to the objective function of the *ChimeraMap* algorithm; there are many relevant objective functions, such as functions that penalize shorter codon blocks non-uniformly along the ORF (if we have prior knowledge that certain regions in the ORF contribute less to its regulation), function(s) that trade-off frequency *vs.* substring length in a different manner than the one reported here, or weighted functions according to the frequency of the substrings. Thus, we can extrapolate many variants of the approach, and in that manner calibrate it to suit specific problems. The last example is related to the alphabet used; here for simplicity we worked with nucleotides (Supplementary Methods); however it may make sense to work with codons, codon pairs, or divide the codons to sub-sets of codons assumed to be "identical".

Moreover, a version of the *Chimera* approach may also be used for engineering and estimating the information related to gene expression encoded in other parts of the gene such as UTRs and introns, known to also include signals related to gene expression regulation (Wang and Cooper 2007; Goodarzi, Najafabadi et al. 2012).

The *ChimeraARS* algorithm may also be modified to consider both substrings' lengths and their frequencies. In such a case, the major challenge is to model the trade off between length and frequency. We believe that the trade-off is organismal specific and should be inferred for each organism separately, possibly based on gene expression (adding additional layers of complexity to such a measure).

Finally, in this study we analysed *E. coli* since this is the only organism with large scale measurements of *both* heterologous and endogenous gene expression data. However, we believe that the results reported here will be even more significant for eukaryotes, and specifically multi-cellular organisms such as plants; there are different stages of gene expression, including many types of interactions with the mRNA molecules that occur only in these groups of organisms; for example, splicing, interaction with the nuclear pores, and regulation by miRNA (see, Figure 1), occur only in eukaryotes; all these examples include interactions between the intracellular machinery and the mRNA molecule, and are at least partially encoded in the ORF via the high dimensional distribution of codons. These signals are expected to be detected by the *Chimera* approach, but not by single codon measures of codon usage bias.

ACKNOWLEDGEMENTS

Funding: This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

REFERENCES

Alberts, B., A. Johnson, et al. (2002). Molecular Biology of the Cell. New York.

Cannarozzi, G., N. N. Schraudolph, et al. (2010). "A role for codon order in translation dynamics." Cell **141**(2): 355-367.

Cartegni, L., S. L. Chew, et al. (2002). "Listening to silence and understanding nonsense: exonic mutations that affect splicing." Nature Reviews Genetics **3**(4): 285-298.

Chamary, J. V., J. L. Parmley, et al. (2006). "Hearing silence: non-neutral evolution at synonymous sites in mammals." Nat Rev Genet **7**(2): 98-108.

Coleman, J. R., D. Papamichail, et al. (2008). "Virus attenuation by genome-scale changes in codon pair bias." Science **320**(5884): 1784.

dos Reis, M., R. Savva, et al. (2004). "Solving the riddle of codon usage preferences: a test for translational selection." Nucleic Acids Res **32**(17): 5036-5044.

Farach, M. (1997). Optimal suffix tree construction with large alphabets. Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on, IEEE.

Farach, M., M. Noordewier, et al. (1995). On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics.

Forman, J. J. and H. A. Coller (2010). "The code within the code." Cell Cycle **9**(8): 1533-1541.

Goodarzi, H., H. S. Najafabadi, et al. (2012). "Systematic discovery of structural elements governing stability of mammalian messenger RNAs." Nature **485**(7397): 264-268.

Goodman, D. B., G. M. Church, et al. (2013). "Causes and Effects of N-Terminal Codon Bias in Bacterial Genes." Science **342**(6157): 475-479.

Gu, W., T. Zhou, et al. (2010). "A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes." PLoS Comput Biol. 2010 **6**(2): 1-8.

Gusfield, D. (1997). Algorithms on strings, trees and sequences: computer science and computational biology, Cambridge University Press.

Gustafsson, C., S. Govindarajan, et al. (2004). "Codon bias and heterologous protein expression." Trends Biotechnol **22**(7): 346-353.

Hogan, D. J., D. P. Riordan, et al. (2008). "Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system." PLoS biology **6**(10): e255.

Irwin, B., J. D. Heck, et al. (1995). "Codon pair utilization biases influence translational elongation step times." Journal of Biological Chemistry **270**(39): 22801-22806.

Kozak, M. (1986). "Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes." Cell **44**(2): 283-292.

Kudla, G., A. W. Murray, et al. (2009). "Coding-sequence determinants of gene expression in Escherichia coli." Science **324**(5924): 255-258.

Lee, M. V., S. E. Topper, et al. (2011). "A dynamic model of proteome changes reveals new roles for transcript alteration in yeast." Mol Syst Biol **7**(514): 514.

Li, G. W., E. Oh, et al. (2012). "The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria." Nature(28).

Manber, U. and G. Myers (1993). "Suffix arrays: a new method for on-line string searches." siam Journal on Computing **22**(5): 935-948.

Pevsner, J. (2009). Bioinformatics and functional genomics, John Wiley & Sons.

Plotkin, J. B. and G. Kudla (2010). "Synonymous but not the same: the causes and consequences of codon bias." Nat Rev Genet **12**(1): 32-42.

Ramakrishnan, V. (2002). "Ribosome structure and the mechanism of translation." Cell **108**(4): 557-572.

Reuveni, S., I. Meilijson, et al. (2011). "Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Model. ." PLoS Comput. Biol. : 1-18.

Sauna, Z. E. and C. Kimchi-Sarfaty (2013). "Understanding the contribution of synonymous mutations to human disease." Nat Rev Genet **12**(10): 683-691.

Schnall-Levin, M., Y. Zhao, et al. (2010). "Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3′ UTRs." Proceedings of the National Academy of Sciences **107**(36): 15751-15756.

Schwanhausser, B., D. Busse, et al. (2011). "Global quantification of mammalian gene expression control." Nature **473**(7347): 337-342.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-1295.

Stergachis, A. B., E. Haugen, et al. (2013). "Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution." Science **342**(6164): 1367-1372.

Tats, A., T. Tenson, et al. (2008). "Preferred and avoided codon pairs in three domains of life." BMC genomics **9**(1): 463.

Tuller, T., A. Carmi, et al. (2010). "An evolutionarily conserved mechanism for controlling the efficiency of protein translation." Cell **141**(2): 344-354.

Tuller, T., I. Veksler-Lublinsky, et al. (2011). "Composite Effects of Gene Determinants on the Translation Speed and Density of Ribosomes " Genome Biol **12**(11): R110.

Ulitsky, I., D. Burstein, et al. (2006). "The average common substring approach to phylogenomic reconstruction." J Comput Biol **13**(2): 336-350.

Vervoort, E. B., A. v. Ravestein, et al. (2000). "Optimizing heterologous expression in Dictyostelium: importance of 5' codon adaptation." Nucl. Acids Res. **28**(10): 2069-2074.

Vogel, C., S. Abreu Rde, et al. (2010). "Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line." Mol Syst Biol **6**(400): 1-9.

Wang, G.-S. and T. A. Cooper (2007). "Splicing in disease: disruption of the splicing code and the decoding machinery." Nature Reviews Genetics **8**(10): 749-761.

Wang, M., M. Weiss, et al. (2012). "PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life." Molecular & cellular proteomics : MCP **11**(8): 492-500.

Wyner, A. and A. Wyner (1995). "An improved version of lempel-ziv algorithm." IEEE Tran. Inf. Theory.

Wyner, A. J. (1993). String matching theorems and applications to data compression and statistics, Stanford University.

Ziv, J. and A. Lempel (1977). "A universal algorithm for sequential data compression." Information Theory, IEEE Transactions on **23**(3): 337-343.

Zur, H. and T. Tuller (2012). "Strong association between mRNA folding strength and protein abundance in S. cerevisiae." EMBO Rep.

Zur, H. and T. Tuller (2013). "New Universal Rules of Eukaryotic Translation Initiation Fidelity." PLoS Comput Biol **9**(7): e1003136.

Zur, H. and T. Tuller (2013). "Transcript features enable accurate prediction and understanding of gene expression in S. cerevisiae." BMC Bioinfomatics. **14**((Suppl 15)): S1.