MINIREVIEW

The Computational Detection of Functional Nucleotide Sequence Motifs in the Coding Regions of Organisms

HARLAN ROBINS,*,[1] MICHAEL KRASNITZ,† AND ARNOLD J. LEVINE†

*Computational Biology Group, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109; and †Institute for Advanced Study, Princeton, New Jersey 08540

A new algorithm has been constructed for finding under- and overrepresented oligonucleotide motifs in the protein coding regions of genomes that have been normalized for G/C content, codon usage, and amino acid order. This Robins-Krasnitz algorithm has been employed to compare the oligonucleotide frequencies between many different prokaryotic genomes. Evidence is presented demonstrating that at least some of these sequence motifs are functionally important and selected for or against during the evolution of these prokaryotes. The applications of this method include the optimization of protein expression for synthetic genes in foreign organisms, identification of novel oligonucleotide signals used by the organism and the examination of evolutionary relationships not dependent upon different gene sequence trees. Exp Biol Med 233:665–673, 2008

Key words: motifs; codon usage; relative entropy

The degeneracy of the genetic code allows for an enormous number of nucleotide sequences to encode the same protein. A protein of typical length (e.g. 300 amino acids) can be translated from well over 10^{50} different nucleotide sequences. This number is close to the estimate of the total number of protons that make up the planet Earth. As this number is extraordinarily large, the set of open reading frames (ORFs) within the genome of every organism has the capacity to carry vastly more information than just the amino acid order. Only a portion of this information, namely A + T/G + C content, the frequency of codon usage, and the amino acid content of all of the proteins of an organism has been employed to explain the unique DNA sequences of genomes.(1–3) But these variables appear insufficient to explain the stability of the *E. coli* DNA sequence.(4) Why does a specific organism like *E. coli*, isolated from many different parts of the world, have a very similar nucleotide sequence? Why doesn't that sequence drift and still produce all of the same proteins (due to selection) found in all the strains of *E. coli*? Is it possible that inherent in the genome of *E. coli* is additional information that preserves that sequence by selection?

Any genome of an organism can be divided into the portion of a gene which encodes the amino acid sequence of a protein and the non-coding portion that contains the regulatory motifs that impact the levels of a gene product, the timing of its expression, and the place (cell type) where it is produced. Clearly the sequence motifs or signals that have biological function are selected for and retained at high frequencies in a genome. Are there similar sets of sequence information in the protein coding regions of the genome which result from the degeneracy of the genetic code and preferential use of selected codons? In this discussion the search for functional sequence motifs or signals in the genome of an organism will be restricted to the protein coding regions of the genome. Of course, there is the confounding effect that some motifs are conserved for non-functional reasons. Only experiments can distinguish the truly functional motifs. The same approaches employed here

This work was supported in part by the Simons Foundation, the Ambrose Monell Foundation, and the Leon Levy Foundation.

[1] To whom correspondence should be addressed at Computational Biology Group, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, WA 98109. E-mail: hrobins@fhcrc.org

DOI: 10.3181/0704-MR-97
1535-3702/08/2336-0665$15.00

can also be usefully applied to the regulatory regions of genomes and this is work in progress.

The proportion of each codon used for a given amino acid varies widely between organisms as well as between genes in the same organism. For example, in *Saccharomyces cerevisiae* the highly expressed genes have different codon usage than poorly expressed genes (3). The typical explanation is that isoaccepting tRNAs have very different concentrations within a yeast cell, so the tRNAs that occur in lower abundance can cause rate-limiting steps in translation of an mRNA. The yeast genome selects against those rare codons that correspond to these low abundance tRNAs. The highly expressed genes have a stronger negative selection pressure on the rare codons than poorly expressed genes. Is this a sufficient explanation to shape the unique DNA sequence and use of codons of *S. cerevisiae*? The results reviewed here suggest other additional factors may play a role. Yet another example of this issue is the LINE-1 retrotransposon elements found in the genomes of many organisms. These LINE-1 elements are very poorly expressed in the somatic cells of their hosts. This is in part due to the heavy methylation of cytosine residues which decreases transcription rates but a second factor, inherent in the nucleotide sequences of LINE-1 elements, slows the rate of transcription of these genomes. Changing the nucleotides in the third position of codons in a LINE-1 element (no amino acid changes) increases the rate of transcription of these elements many fold (5, 6). This can impact the frequency of transposition and the forces of selection for these sequence motifs. Because the poor expression is a transcriptional problem, not translation, this increase due to codon change is not related to codon usage in translation. Could it be that inherent in these LINE-1 elements are sequence motifs that modulate transcription rates or RNA processing rates? Yet a third example of a possible role for nucleotide sequence motifs in protein coding regions of genes, which are not related to codon usage, is found in the Human Immunodeficiency Virus. The DNA copy of this virus integrated into the host cell genome is transcribed poorly and RNA processing of HIV mRNA is inefficient. Indeed two viral proteins are required to enhance transcription (Tat) and RNA processing (Rev). In addition to these aids in gene expression, however, synonymous changes in the third positions of this HIV DNA sequence greatly increases the gene expression of this virus even in the absence of Tat and Rev (7). Thus, there is growing evidence that the RNA sequence of an organism contains sequence motifs, other than codons, which impart biological information. Despite the direct biological relevance of codon usage, there has been very little effort to search for additional information within coding regions (8–11). Here we review the Robins-Krasnitz algorithm, a new method to search this vast amount of additional information contained in the coding region of each organism for biologically functional signals (12). In particular, the algorithm systematically determines the set of small motifs which are under- and overrepresented in a genome.

This search algorithm has overcome two major obstacles. First, the algorithm attempts to find signals which are obscured by the already known biological properties: coding for proteins and codon usage. These make up a set of complicated constraints that must be controlled. Second, motifs come in large families that are highly codependent. For example, CpG is significantly underrepresented in the human genome because a common mutation is C→U through the removal of an amine group. This negative evolutionary pressure significantly reduces the entire superfamily of motifs that include CG as a substring, e.g. ACG, CCG, CGA, etc. Additionally, the mutation increases the number of TAs and its superfamily. Disentangling the primary object of selection is a difficult challenge. Taking advantage of powerful information theory tools, the Robins-Krasnitz algorithm overcomes both of these obstacles. One drawback of the algorithm in the present form is that it is limited to exact motifs. Extending the formulation to include degenerate motifs is a difficult problem.

These obstacles, particularly the first one, are a specific case of a general problem in bioinformatics as well as other areas of computational biology (13). Finding a pattern or signal requires a null-hypothesis for what the genome would look like without signal. An incorrect null-hypothesis (or background model) leads to the discovery of false signals. Our background model needs to account for as much biological knowledge as possible without damping the signal too dramatically. For the case of finding transcription factor binding sites in upstream regions of genes, we have very little biological knowledge to use. Genomic algorithms have had much difficulty separating out false signals. In the case of coding regions, we have substantially more biological knowledge to incorporate into our background model. For coding regions, we must be careful of the opposite issue, over constraining the background and losing the signal. Of the two constraints we consider, amino acid order is not controversial. The interesting decision is how to include codon usage. We are limited for practical reasons to two choices; either we can fix codon usage in the whole genome or we can fix it individually in each gene. We have strong biological evidence that codon usage within a gene has a large effect on rates of gene expression in bacteria. Additionally, the variation of codon usage between genes in the same genome varies widely, with the variation strongly correlated with expression. If overall codon usage in the genome is used as a constraint, we will be ignoring substantial biological information. On a practical level, fixing codon usage in each gene also fixes mononucleotide content. Lifting this constraint creates false signals that correlate with GC content.

This review is divided into five sections. Sections one and two present the algorithm, explaining how both obstacles described above are overcome. The third section provides evidence for biological function of the motifs. The

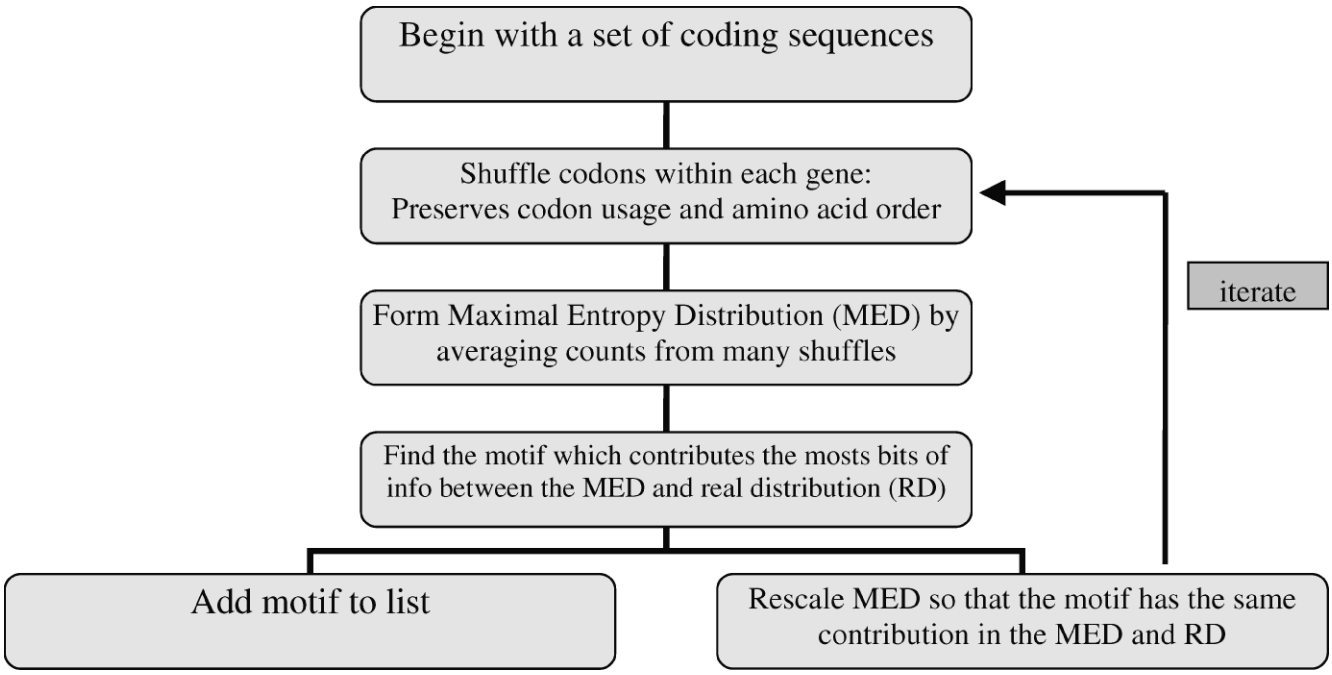

Figure 1. Outline of Robins-Krasnitz algorithm. Figure 1 shows a flow chart of the steps in the Robins-Krasnitz algorithm. The process is iterated with a new motif added to the output at each step.

fourth section presents a set of possible applications for this information. The last section explores the reasons why such coding region sequence motifs might arise in different organisms and differ in different organisms.

1) Creating a Maximal Entropy Distribution

The Robins-Krasnitz algorithm finds short nucleotide motifs in coding regions that are independent of amino acid order and codon usage. Figure 1 contains a flow chart outlining the algorithm. Codon usage, for this paper, is defined to mean the total fraction of each codon used in a given gene. Finding a set of over- and underrepresented motifs requires a clear definition of the background or expected distribution of motifs. In other words, we must determine a standard of comparison to give the terms over- and underrepresented meaning. Because we are searching for signals that are orthogonal to the known information, amino acid order and codon usage, we treat them as constraints. The goal is to find the most random distribution of motifs from a given set of sequences subject to the constraints that amino acid order and codon usage are held constant. A probability distribution which is completely unstructured except for a set of constraints is called a maximal entropy distribution (MED). For the problem at hand, the constraints are too complicated to analytically find the MED. However, we can extract the distribution using a Monte Carlo procedure as long as we randomize at each step such that the constraints are obeyed exactly and no other conditions are imposed.

The Monte Carlo procedure is implemented using what we will refer to as a shuffling program. For each coding sequence in the set of genes of interest, we consider each type of amino acid separately. At each step in the Monte Carlo, all the codons in a gene (and then all of the genes in a genome) that are translated into the same amino acid are randomly permuted. For example, if a protein has four different Leucines coded by four different codons, each iteration of the Monte Carlo will produce a new nucleotide sequence, where the four codons for Leucine are randomly permuted amongst themselves in the gene. See Figure 2 for an illustration of this process. The same randomization is done for all 20 amino acids at each step in the Monte Carlo. This procedure clearly obeys all of the necessary requirements to generate a MED. The constraints are met by construction at each iteration. Amino acid order is never changed as we are only permuting the codons that code for the same amino acid. Additionally, the codon usage for each amino acid in each gene is held constant because the permutation does not change the number of any codon in

M	L_1	L_2	H_1	L_3	H_2	L_4	H_3	ST
ATG	CTA	CTG	CAT	TTA	CAT	CTG	CTT	TAG

Figure 2. Example of shuffling procedure. The procedure to get the maximal entropy distribution (MED) involves a set of randomized iterations. The triplets of nucleotides coding for each amino acid are permuted randomly among themselves. This is an illustrative example of mock short protein with eight amino acids. The shuffling procedure randomly permutes L_1, L_2, L_3, and L_4 and separately permutes H_1, H_2, and H_3. Each iteration produces a new sequence. For this example, there are 12 different combinations for the Leucines and three combinations for the Histidines giving 36 unique sequences. They are weighted in the shuffling procedure so that the MED is attained in the limit of a large number of iterations.

Box 1A. Extracting the Maximal Entropy Distribution from a Set of Sequences.

$N_{MED}(w)$ is the average count for each word w from the Monte Carlo runs.

$W_i^{\max}$ is the set of words of maximum length where $i = 0, \ldots, 4^{L_{\max}-1}$.

All counts $N_{MED}(w)$ can be determined from $N_{MED}(W_i^{\max})$, but a couple of additional quantities need to be defined.

$C(W_i^{\max},w)$ is the number of times the word w is contained in the word $W_i^{\max}$.

$L(w)$ is the length of word w.

The equation to extract the counts of all the shorter motifs from the counts of the longest motifs is

$$N_{MED}(w) = \sum_{i=0}^{(4^{L(W^{\max})}-1)} \frac{N_{MED}(W_i^{\max}) * C(W_i^{\max}, w)}{1 + L(W^{\max}) - L(w)}.$$

$P_{MED}(w) = N_{MED}(w)/L$ is the probability of motif w, where L is the total length of sequence.

$P_R(w) = N_R(w)/L$ is the probability of motif w in the real sequence (this is found by counting).

Box 1B. An Example.

Consider the case where the maximum length is 7, which is the maximum size motif that we can consider for most bacterial genomes.

Let $w = AAC$ and $W_{257}^7 = AACAAAC$, and where the label 257 is the base four conversion of the motif with A = 0, C = 1, G = 2, T = 3, i.e. 0010001 = 4^4 + 4^0.

$L(w) = 3$, $C(W_7^{257},w) = 2$, and $L(W^{\max}) = 7$.

To get $N_{MED}(w)$, this process is repeated for all 4^7 $W_i^{\max}$ terms. The values are then added in the formula above.

any gene. The shuffling procedure only moves the codons around within the same gene. Any additional variation beyond the shuffling procedure would violate one of the constraints, therefore, extracting the distribution as the average of a number of iterations of this Monte Carlo procedure approaches the MED. The distribution converges quickly, so computational time is not an issue even for the full human genome. Convergence usually occurs by 20 iterations, but we use 50 to be conservative.

The shuffling procedure described above gives a set of randomized sequences. How do we extract a probability distribution of motifs from a set of sequences? As long as the number of occurrences of each motif found in the total set of sequences is reasonably large, we can form a probability distribution, estimating the probability of a given motif by its fraction in the set of all motifs. If a certain motif of length four occurs 2000 times in the set of sequences and the total number of motifs of length four is 100,000, then we estimate the probability of that motif to be 1/50. An estimate of the variation for the count of a specific motif is the square root of the count. For counts as small as 30, the fraction is still a reasonable estimate for the probability. For the MED, we take the average of a set of randomizations, the Monte Carlo iterations.

For practical purposes, we actually only need to keep track of the set of the longest motifs we wish to consider. All shorter motifs can be retrieved from this set. We determine the longest length motif that is statistically robust for the set of sequences we are analyzing. For the case of a bacterium with 1 Mb of coding sequence, we can find motifs up to length seven. Since $10^6/4^7 \sim 60$, all the motifs of length seven would be expected to have reasonably large counts. Then, we move a sliding window of this length across the sets of sequences, moving one base at a time. This process determines the counts for the largest length motifs. However, the shorter length motifs are over counted, because they appear in more than one window each. The formula to deconstruct the counts for all the shorter length motifs is found in Box 1A. An example is found in Box 1B. The estimated probabilities for each motif are simply the counts divided by the total length of sequence. We now have two distributions to compare: one is the motif probabilities from the real set of sequences and the other is the MED distribution which is found by averaging the motif probabilities from a number of Monte Carlo iterations.

2) Using Relative Entropy to Choose Over- and Underrepresented Motifs

The shuffling procedure defines two distributions, the real distribution found from the actual sequence and the MED which we use as the surrogate for the background. The first obstacle is overcome, because the MED exactly conserves the constraints. Now we need a method for choosing under- and overrepresented motifs. The standard we use is from information theory. The motif that contributes the most bits of information to the difference between the real distribution and the MED is the first motif we choose. Using information theory has the nice feature of putting all results in the same units, number of bits. This allows us to compare motifs of different lengths and motifs that are either over- or underrepresented. The formula we employ to compute the motif contributing the most bits of information between the two distributions is the Kullback-Leibler distance or the Relative Entropy. The formula is found in Box 2. We compute the Relative Entropy contribution for each motif and pick out the one with the largest value.

The other choice of distance is a symmetrized version of the Kullback-Leibler (KL) distance, called the Jensen-Shannon (JS) divergence. The practical difference is small

for our application as the KL distance is symmetric to lowest order in our application. The JS divergence also smoothes the result, which removes some of the signal. JS converges faster than KL upon rescaling, so we use KL.

Now that we have found the most under- or over-represented motif in the sequence, we have to tackle the second obstacle. We need to pick out the motif which is the next most under- or overrepresented. However, we cannot simply take the motif which has the next largest Relative Entropy. This is because the motifs are overlapping, so under- or overrepresentation of a given motif affects the distribution of all the other motifs. The example of CpG from the introduction illustrates this point. If we take the human genome, CG will have the largest Relative Entropy. However, all eight trimers which contain CG as a subset fall within the top 50 highest Relative Entropy of all motifs. This is simply an artifact of the selection against CG. We must first remove the contribution of CG from the MED before recalculating the Relative Entropy to find the next motif. If we call the motif w, we rescale all motifs that contain w, by the same amount such that the rescaled MED has the same distribution for w as the real distribution. This forces the Relative Entropy of w to zero and, at the same time, removes the contribution of w from all other motifs.

The mathematical details of the rescaling are as follows with many of the definitions in Boxes 1 and 2. After a motif w is found to be over- or underrepresented, its contribution to the difference between the real distribution and the MED needs to be eliminated. The maximal entropy distribution is rescaled in a minimal way, such that the contribution of w becomes identical in both the real distribution and the MED. For the rescaling to be minimal, the ratios of frequencies of words $W_i^{\max}$ of length *max* that contain w the same number of times should not change. That is, all words $W_i^{\max}$ with the same $C(W_i^{\max},w)$ must be rescaled by an equal factor. Consider a coarse graining of the detailed probability distributions. The MED is defined as the set of words $W_i^{\max}$ of length *max*, with the probabilities $P_{MED}(W_i^{\max})$. The set of $W_i^{\max}$ is partitioned into disjoint subsets where each element of a given subset contains the motif w an equal number of times. These sets are

$$K_J(w) = \{W_i^{\max} | C(W_i^{\max}, w) = J\}$$

$$\text{with } J = \{0, \ldots, \max - 1\} \text{ and}$$

$$K_0 \cup \ldots \cup K_{\max-1} = \{W_i^{\max}\}.$$

These disjoint subsets, $K_J(w)$, need to be rescaled such that the probability of being in a given subset is equal in the real distribution and the MED. Define

$$Q_R(K_J) = \sum_{W_i^{\max} \in K_J} P_R(W_i^{\max}),$$

Box 2. The Relative Entropy or Kullback-Leibler Distance.

The Kullback-Leibler distance between the real and MED probability distributions is

$$D_{KL} = \sum_i P_R(W_i^{\max}) \log \frac{P_R(W_i^{\max})}{P_{MED}(W_i^{\max})}.$$

A figure of merit that measures the extent to which *any* word w, of length 2 to *max*, contributes to D_{KL} is

$$S(w) = P_R(w) \log \frac{P_R(w)}{P_{MED}(w)} + (1 - P_R(w)) \log \left(\frac{1 - P_R(w)}{1 - P_{MED}(w)} \right).$$

This can also be thought of as a Kullback-Leibler distance between two probability distributions, namely the coarse-grained real and background distributions where we only know if a given word is or is not w.

$$Q_{MED}(K_J) = \sum_{W_i^{\max} \in K_J} P_{MED}(W_i^{\max}).$$

These are well-defined probability distributions because they are grouped elements from the old probability distribution (and their probabilities are added). A rescaling that factors out the contribution of w while conserving probability is given by

$$N_{MED}(W_i^{\max}) \vec{w} \frac{Q_R(K_J)}{Q_{MED}(K_J)} N_{MED}(W_i^{\max})$$

for all i. Note that with this rescaled distribution the figure of merit for w is now $S^{rescaled}(w) = 0$, so the contribution of w to D_{KL} has been removed. Now, the whole procedure can be repeated to find the next word w, etc.

It is not hard to see that in this iterative algorithm, the MED $P_{MED}(W_i^{\max})$ converges to the real distribution $P_R(W_i^{\max})$. This is because D_{KL} is monotonically decreasing (proof is found in Robins *et al*. 2005). It is well known that D_{KL} is nonnegative, and is 0 if and only if the two distributions are identical. The algorithm will not halt before convergence is achieved since $S(w) = 0$ for all w also implies that the real and background distributions are identical. Finally, D_{KL} cannot converge to a positive value, since one could then find a word that would reduce it below that value.

The procedure is iterated, so that we remove the contribution of one motif at a time from contributing to the Relative Entropy through rescaling of the MED. Then, we choose the next motif. The rescaling automatically over-

Table 1. The Top 100 Motifs for *E. coli* in Order of Significance. The + and – Denote Over- and Underrepresentation, Respectively

1	GGCC	–	21	TTGG	–	41	CCTGGA	–	61	TGGCCT	+	81	TCGCA	–
2	TAG	–	22	TCAAG	–	42	CTCC	–	62	GGAG	–	82	TATCG	+
3	GCTGG	+	23	CGCTG	+	43	GCGGA	–	63	GAGCTC	–	83	TGCGA	–
4	TTGGA	–	24	CGGTA	+	44	TCAGG	+	64	CTCGAC	+	84	CTGGGGC	+
5	CCC	–	25	CACGTG	–	45	AGCGCT	–	65	CAGCAA	+	85	TGAAG	+
6	GGCGCC	–	26	CCGCGG	–	46	TGGTG	+	66	AGAC	–	86	GTCTGG	+
7	GTCC	–	27	GGCTC	–	47	TCGTA	–	67	TCCAA	–	87	GTCGAT	+
8	GCCGGC	–	28	CATG	–	48	CGCC	+	68	TATGAT	–	88	GCATGC	–
9	GGCG	+	29	CAGCTG	–	49	CGGCA	+	69	TATA	–	89	GAAT	–
10	GAGG	–	30	TCGGA	–	50	TTCCCG	+	70	GGAC	–	90	GTTGA	+
11	CCAAG	–	31	CGGCCG	–	51	CACCA	+	71	TGGCCA	–	91	CAGTAA	+
12	GGGT	–	32	AATT	–	52	GGTACC	–	72	TTCCTCG	+	92	GAGCC	–
13	TTGCC	+	33	TTGCT	+	53	TTCG	–	73	CTGTC	–	93	ACGCT	+
14	CTAG	–	34	CGAG	–	54	TGGG	–	74	TGTG	–	94	CAGCGA	+
15	GCCAG	+	35	GGGGG	–	55	GAGACC	–	75	TCTGA	–	95	TCCGA	–
16	CTGCAG	–	36	TCGTG	–	56	GATCG	–	76	CAGAT	–	96	CTGGAAG	+
17	AAGAG	+	37	CTGAG	–	57	CCCTG	–	77	CGTG	–	97	TCTTA	–
18	ACTGG	+	38	GACC	–	58	CTGGCTG	+	78	CAAG	–	98	CTCAGA	–
19	GAAC	–	39	CTTG	–	59	CTGGCA	+	79	CTGCTGG	–	99	TTGACA	–
20	GTGT	–	40	GGTCTC	–	60	AAAAAAA	–	80	TTTTTT	–	100	CCATGG	–

comes the second obstacle. The impact of selection on a motif w, affects all other motifs, especially its supersets and subsets. Our scaling procedure systematically handles all of these motifs at the same time.

Iterations are performed until the largest Relative Entropy for any motif is small enough that we would expect it by chance. This cutoff is the point where it becomes likely that chance fluctuations would create the most significant remaining word, appropriately corrected for multiple hypotheses (the set of all words of length w). The cutoff occurs when the selected word w satisfies

$$erfc\left(\frac{|N_R(w) - N_{MED}(w)|}{\Delta(w)\sqrt{2}}\right) * 4^{L(w)} > 1/2,$$

where $\Delta(w)$ is the standard deviation of the background count for w. For all our applications we stopped after 100 iterations, which is substantially below the cutoff.

The algorithm gives a set of motifs of varying lengths which are under- or overrepresented in the coding region. Each motif varies in a statistically significant manner between the real and Maximal Entropy Distributions. Having this list, the question then arises; do these sequence motifs found under- or overrepresented in the coding regions of genes have any biological function?

3) Evidence for Biological Function

A computational sequence analysis like the one preformed above can provide evidence for function in two ways. Our functionality tests are applied to bacterial genomes because of the wide range of sequence data available on the NCBI website. Below we show that these over- and underrepresented sequence motifs have been evolving with the bacteria over billions of years of evolution. We need to be a bit careful here, because motifs can appear conserved for reasons other than direct biological function. In particular, motifs can be also be conserved by historical accident as well as by unknown biochemical constraints. Since we are considering billions of years of evolution, the rate of accidental conservation should be very low. However, we are not able to separate biochemical constraints from functional elements. Next, we can, in some circumstances, directly associate a motif with a known function. Within the set of underrepresented bacterial motifs, most of the known restriction sites are found. For example, Table 1 shows the results of the top 100 most under- and overrepresented motifs from the coding regions of the *E. coli* genome. We suggest that the algorithm presented is a valid method to determine an exhaustive list of restriction sites that are found in different bacterial organisms. This also shows that different species and genera of bacteria have different sequence motifs over- or underrepresented. There is no universal sequence motif avoided or preferred by all bacterial organisms sequenced.

Before providing evidence for evolutionary conservation, we must first provide a necessary check that the motifs are real properties of bacterial genomes. We choose a particular genome, in our case *E. coli*, and randomly divide the full set of genes in half. We run our algorithm on both sets independently and focus on the top 100 motifs from each half. This self consistency check requires both lists to be the same up to minor fluctuations. This is exactly what we get. There are minor variations in order, but every one of the top 80 motifs from one half is within the top 100 on the other list. We repeated this process for many random divisions and always pass this consistency check. We examine two different Null hypotheses relating to get statistics for the previous statement. We reject the Null

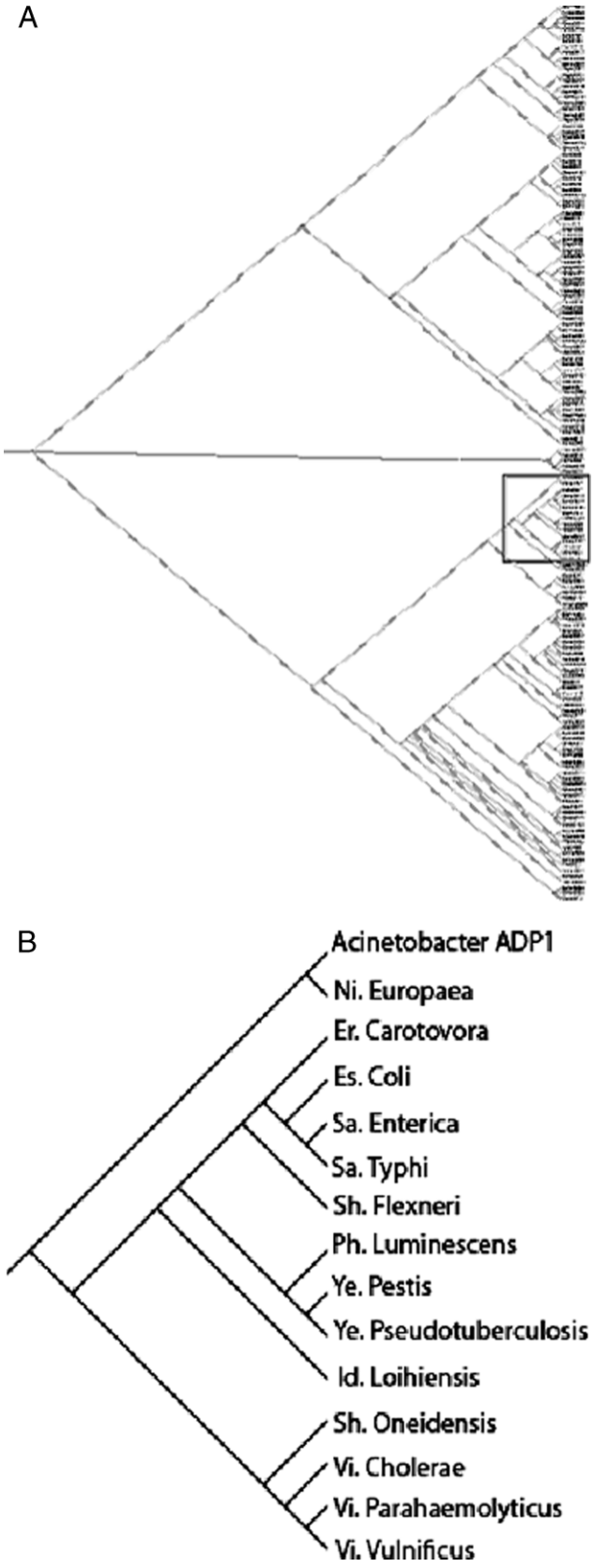

Figure 3. A bacterial phylogenetic tree using a metric based solely on motifs differences. Figure 3A is the phylogenetic tree for the 164 bacterial species found in the NCBI database. The black rectangle encloses the *Enterobacteria* clade. Figure 3B contains a blow up of this section of the tree. The only *Enterobacteria* missing from this group is *Buchnera Aphidicola*.

hypothesis that two different sets of motifs from separate runs are drawn from independent distributions with a *P*-value $\ll 10^{-7}$. We are unable to reject the Null hypothesis that the two sets of motifs from different runs are drawn from the same distribution, *P*-value = 4. We are unable to find evidence for different motifs in separate divisions of the genome.

Additionally, we have analyzed the frame bias for the start of the motifs. We asked the question, are the motifs preferentially beginning in a particular reading frame. If the answer was yes, this could either be an artifact of the method or be a biologically functional. However, there is no frame preference overall. The distribution among individual motifs agrees with a random distribution (data not shown).

We use the *E. coli* example to explore the relative effect of the constraints. If we remove the constraint on amino acid order in each gene, we maintain the constraints on GC bias and codon usage. Lifting this constraint makes the least change between the distributions. The result is that 40% of the top 100 motifs are completely different, while 60% are the same. Removing further constraints drops the overlap of motifs below 25%.

How do the motifs in one bacterium relate to the motifs in another bacterium? Using the list of motifs the Robins-Krasnitz algorithm produces for each of the 164 bacterial species found in the NCBI database, we defined a Hamming distance as a metric. We constructed a phylogenetic tree of these bacteria based entirely on the motifs and this metric (Figs. 3A, 3B). The tree formed from over- or under-represented sequences was then compared with an evolutionary tree constructed from the ribosomal gene family sequences, which is considered at this time the gold standard. The tree created by the Robins-Krasnitz algorithm captures most of the standard bacterial taxonomy with a whole or partial genome approach (12). This distribution of lists of over- and underrepresented sequences for each organism should permit the assembly and classification of other unknown organisms from their sequence data. Thus, this method can be used to take advantage of shotgun sequence data, because it does not require alignment of orthologous genes. As long as a sufficient amount of sequence from a bacterium, e.g. over 50,000 bases, are sequenced, this method can accurately place the unknown organism into a region of the bacterial phylogenetic tree.

The last piece of evidence for biological function is found in the relationship between viruses and their hosts. As viruses are intercellular parasites, their genomes are subject to similar selection pressures to the host genomes. There has been a large amount of work comparing bacterial genomes to their infecting phage genomes. Codon usage and dinucleotide content have both been utilized with some

Table 2. Summary of Scores for Phage/Host Predictions. Top Score Is the Percent of Phages that Scored Highest with Their Known True Hosts. Top Three Is the Percent of Phages such that the True Host Scored in the Top Three out of All 108 Possible Bacterial Genera.

Phage type	All	dsDNA	Lytic	Temperate
% top score	50	58	45	70
% top three	71	82	59	93

success to illuminate features that are shared preferentially between phage genomes and their hosts (14). The motifs from the Robins-Krasnitz algorithm capture substantially more information about the relationship between phages and hosts. These over- and underrepresented motifs provide a prediction accuracy of phages to their known hosts of almost double the other methods. Table 2 displays the results of the Robins-Krasnitz algorithm applied to 131 phages with known bacterial hosts. The algorithm is used to match the phages to the set of 108 bacterial genera using a scoring algorithm described in Robins *et al.* 2005.

4) Applications

The set of applications of the Robins-Krasnitz algorithm may be grouped into three different categories. The first is a direct utilization of the algorithm to enhance medical research. The second is as a bioinformatics tool to assist in the identification and categorization of defined motifs or signals in the genome. Third, the motifs themselves can be used as a stepping stone for basic research. We have shown indirectly that the motifs have biological function. Future experiments will be required to determine the possible biological roles of each of these motifs.

Two projects are presently underway for direct application of these motifs for medical research. The first project extrapolates the phage/bacterial analysis to humans and our viruses. Like bacteria, many of the motifs derived from the human genome match the motifs from our infecting viruses. However, viral genes can have regulatory differences when compared to the sequences in the human genome. The evolutionary pressures on the viral genomes may result in design differences when compared to those in the human genome. These differences may well be reflected in differences in the motifs detected by the Robins-Krasnitz algorithm. There are striking differences between the motifs found in the mRNA of HIV-1 and those found in the set of all human genes in their coding regions. As mentioned previously the genes of HIV are expressed very poorly in their human host. Focusing on a single motif which is overrepresented in the HIV-1 genome and underrepresented in the coding regions of the human genome, we have recoded the Gag gene of HIV-1 to remove this signal. This simple change in coding doubled the expression of Gag in a human kidney cell line. The same recoded sequence was then used as a DNA vaccine in a mouse model. The anti-gag immune response was six fold higher for our recoded sequence. This avenue of research is presently ongoing and would be a first step in improving an HIV DNA vaccine. It might also help explain the modest success of the HIV DNA vaccines.

The second practical application is to increase the gene expression of coding regions of genes from one host expressed in a second host. Human or viral genes are often expressed in foreign organisms, usually bacteria, yeast, or baculoviruses, for either medical purposes or basic research applications. There is a promising computational project underway to create a tool which combines codon optimization and the contribution of the motifs to recreate a sequence for a particular protein which has increased expression in a given organism.

Additionally, the motifs can be used as a bioinformatic tool to determine the origin and classification of DNA sequences of unknown origin or to assemble a genome from its parts when multiple organisms compose a sample for DNA sequencing. With ongoing shotgun sequencing projects to explore the bacterial diversity in the Sargasso Sea, the air above Manhattan, and the human digestive track, among others, classifying bacterial phylogeny based on novel sequence is a useful tool (15). For this purpose a genome wide property that accurately classifies a genome will be essential. The common tool presently used is dinucleotide content. This method has advantages over the Robins Krasnitz algorithm for very short sequences, but as is demonstrated by phage/host analysis, our algorithm becomes significantly better than the other methods when the sequences reach a few tens of thousands of bases.

Finally, as shown in section 3, the motifs themselves appear to have biological function. We expect a subset of these motifs to be protein binding sequences. Further work needs to be done to determine which proteins are binding these particular sequences and what their biological function could be in a cell. Having these overrepresented sequences provides the tools to isolate and identify DNA sequence specific binding proteins and then to determine their functions. Some of the DNA sequence motifs could be providing physical properties such as flexibility for bending of DNA, packaging DNA in a cell or interactions with RNA polymerases. These DNA sequence motifs could function in the DNA or the mRNA produced from that DNA sequence. One way to distinguish between these alternatives is to look for the reverse complement of an over- or underrepresented sequence in the non-coding strand of DNA (note all of the motifs discussed here come from the coding strand). An analysis of the genes in the human genome with a modified Robins-Krasnitz algorithm has identified overrepresented sequences that function in both DNA and RNA molecules (transcription factor binding sites, splice junctions, splice enhancers, etc.). More than half of these sequence motifs have unknown functions.

If it is correct that these sequence motifs that are over- and underrepresented in a genome are derived from bases in the third position of the coding regions of a gene and the motifs that are formed are selected for or against, then some third position mutations may not be neutral to selection even though they don't change an amino acid. Synonymous single nucleotide polymorphisms that have a phenotype could well uncover sequence motifs (out of reading frame) with functions that have important consequences. For example a synonymous single nucleotide sequence in the multidrug resistance gene-1 alters the functionality of that gene product (16). Many of the mutations in the IARC p53 gene mutant database reside in the third position but no one has tested them for their functional consequences, if any. Clearly these ideas bring up some new directions for research in this area.

5) Why Do These DNA Sequence Motifs Arise in the Genome?

This review and the prior publication upon which it is based (12) demonstrates that:

1. The coding regions of genes in a genome contain sequence motifs (oligonucleotide sequences) that are significantly over- or underrepresented (compared to the MED) even when codon usage, G + C content and amino acid order is normalized.
2. The DNA sequence motifs differ in different organisms and are distributed uniformly around the genome (a real exception to this is the G/C rich regions in the genomes of mammals, birds and some reptiles).
3. Several tests suggest that these over- and underrepresented sequences are functional and they can be used to form an evolutionary tree much like the ribosomal gene family.
4. The best classifier of the host of a bacteriophage is based upon these sequence motifs.
5. Several of these sequence motifs can be assigned functions in the cell.

These observations provide good evidence that these sequence signals in a genome are under positive or negative evolutionary selection pressures and so genome sequences are shaped by many diverse sources. But it could be that these sequence motifs arise in a more complex way. The mutation rate and even the nature or the frequency of different types of mutations (transitions, transversions, direction of mutations; A to G or G to A) can differ in different organisms. Mutation rates can be set by the variation in the environment or by diversity of the environment. Thus, there is an evolutionary bias for some sequence motifs in a genome and this could differ between organisms, be distributed evenly around the genome, provide evolutionary diversity and an evolutionary tree of related organisms and would impact upon both the host and viral genomes. It seems likely that some of these sequence motifs could arise from a bias in the mutation type and frequency and might not have a functional consequence even though it was selected for by the host. It may be likely that both functional and non-functional sequence motifs reside in this distribution of over- and underrepresented sequences. What is clear is that such sequence motifs identify important processes in the evolution of the host and contain information about those processes that will need to be decoded in the future.

1. Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci U S A 101:12588–12591, 2004.
2. Vinogradov AE. Isochores and tissue-specificity. Nucleic Acids Res 31: 5212–5220, 2003.
3. Sharp PM, Li WH. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295, 1987.
4. Fuglsang A. Evolution of prokaryotic DNA: intragenic and extragenic divergences observed with orthologs from three related species. Mol Biol Evol 21:1152–1159, 2004.
5. Han JS, Boeke JD. A highly active synthetic mammalian retrotransposon. Nature 429:314–318, 2004.
6. Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429:268–274, 2004.
7. Schneider R, Campbell M, Nasioulas G, Felber BK, Pavlakis GN. Inactivation of the human immunodeficiency virus type 1 inhibitory elements allows Rev-independent expression of Gag and Gag/protease and particle formation. J Virol 71:4892–4903, 1997.
8. Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res 32:W187–190, 2004.
9. Fuglsang A. The relationship between palindrome avoidance and intragenic codon usage variations: a Monte Carlo study. Biochem Biophys Res Commun 316:755–762, 2004.
10. Karlin S, Cardon LR. Computational DNA sequence analysis. Annu Rev Microbiol 48:619–654, 1994.
11. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11: 377–394, 2004.
12. Robins H, Krasnitz M, Barak H, Levine AJ. A relative-entropy algorithm for genomic fingerprinting captures host-phage similarities. J Bacteriol 187:8370–8374, 2005.
13. Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, Stone L. Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". Science (New York, NY) 305:1107; author reply 1107, 2004.
14. Blaisdell BE, Campbell AM, Karlin S. Similarities and dissimilarities of phage genomes. Proc Natl Acad Sci U S A 93:5854–5859, 1996.
15. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. Environmental genome shotgun sequencing of the Sargasso Sea. Science (New York, NY) 304:66–74, 2004.
16. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science (New York, NY) 315:525–528, 2007.