Large-scale Asymmetries in the DNA of E. coli

Kenneth Jones
Department of Computer Sciences
University of Wisconsin, Madison, WI 53706, USA
kjones@cs.wisc.edu

Abstract

Compositional asymmetries in circular bacterial chromosomes exist on a large-scale. Using a sliding window to take the average concentration of nucleotides across a length equal to one-half of the chromosome (~2.3 million characters), it is possible to to split the E. Coli genome into two distinct region: one with a bias in C concentration and the other with a bias in G concentration. This genome wide asymmetry is known as GC skew and is due to the two-stranded structure of the chromosome.

What drives this difference in composition? How is it actualized in a given sequence? A generic method is presented using regular expressions and mutual information to find embedded structure in these two regions. This technique should be able to identify the still debated explanation of third codon position bias responsible for GC skew, but might also identify other possible explanations should they exist.

Key words: GC skew, Compositional asymmetries, Mutual information

Introduction

Large-scale asymmetries in composition can be found in circular bacterial genomes. The specific biological mechanism by which this phenomenon arises is not entirely certain, but it is hypothesized that a difference in mutation rate due to the process of cell division is responsible (Nikolaou and Almirantis 2005). Regardless of the cause, understanding the nature by which such a skew is actualized in the sequence of nucleotides might be revealing. Any bias must not interfere with encoding of protein sequences, so it naturally follows the most likely mechanism is a third position codon bias due to the degenerate of nature of the genetic code giving this position less significance on the resulting amino acid.

Some studies have used statistical methods to show there is indeed a third position bias, but many of the studies offer contradictory results. By using GC skew to split the genome into a "C" region and "G" region of comparable length, this study uses a generic method employing regular expressions combined with mutual information to search the two "documents" for discriminatory features.

Base composition

In most probabilistic models, the pattern of interest comes from a particular distribution(s), while the non-pattern is assumed to come from a fixed background distribution representing the entire genome. The following table shows the background for the E. coli genome.

Table 1

Base compositions

	A	%	T	%	G	%	C	%
Whole:	.246		.246		.254		.254	
C region:	.248		.244		.245		.261	
G region:	.244	-1.6	.246	.3	.263	7.5	.247	-5.3

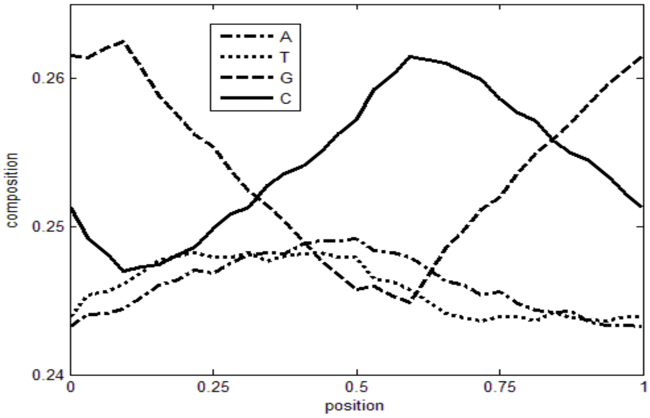
How well do these numbers actually represent what is happening in the genome? Simply treating one static set of numbers as the genome background probability might be an oversimplification given the dynamics present between the C and G halves.

Moving average

By using a smoothing technique known as a sliding window (aka moving average) to sample the nucleotide composition in a buffer of fixed length, a DNA sequence can be transformed into continuous measurements of the relative levels of concentration of nucleotides as shown in the below.

Figure 1

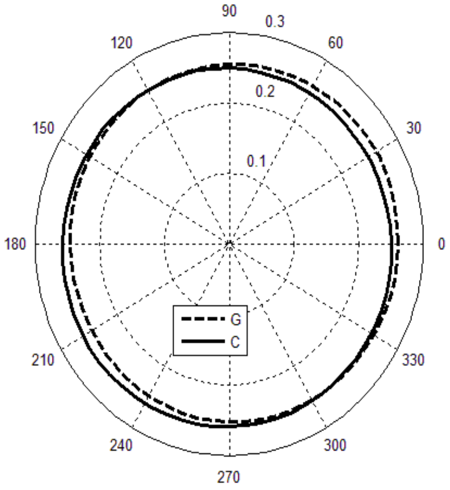
Sliding window using the ½ genome length



Another visualization of the same data highlights the difference in G vs C content along the genome.

Figure 2

Polar plot of sliding window using ½ genome length



The focus of this particular study is on only one window size, but polar plots help identify trends which are also present at shorter wavelengths.

Markov Transitions

Beyond just compositions, what about the order? Changes from one base to the next are known as first-order Markov probabilities. The following table shows these transitions for the entire genome alongside the C and G regions.

Table 2

First-order Markov probabilities

	A	%	T	%	G	%	C	%
Whole:								
A	.296		.271		.208		.225	
T	.186		.298		.282		.234	
G	.227		.217		.230		.326	
C	.276		.200		.294		.230	
C region:								
A	.295		.272		.202		.230	
T	.187		.302		.270		.241	
G	.227		.216		.220		.336	
C	.281		.197		.284		.238	
G region:								
A	.297	.5	.270	-.7	.215	**6**	.219	**-5.1**
T	.184	-1.6	.293	-2.9	.296	**9.8**	.227	**-6.1**
G	.227	-.2	.218	.8	.238	**8.3**	.317	**-5.8**
C	.270	-3.8	.203	3.2	.305	**7.2**	.222	**-6.8**

Long sequences of repeated G nucleotides might account for the bias present in the G region. If that were the case, transitions from G->G would increase markedly. Notice how transitions to the G nucleotide increase uniformly when moving from C to G regions, along with a corresponding uniform decrease in transitions to the C nucleotide. No one transition stands out, hence the first-order properties do not help explain the GC skew.

Second-order Markov transitions

Since one transition does not yield much information, what do the transitions between pairs of nucleotides tell us? The following table shows these second-order Markov transitions for the entire genome alongside the C and G regions where each row shows transitions from one pair to the next. For example, the cell (AA, xT) shows the transition from AA -> TT while the cell (GG, xT) shows the transition from GG -> GT. Only four transitions are possible out of each pair shown in individual rows.

Table 3

C vs G genome second-order Markov probabilities

	xA	xT	xG	xC
C->G:				
AA	-.91%	4.69%	**10.33%**	**-3.85%**
AT	3.13%	-1.26%	**9.50%**	**-4.35%**
AG	-2.67%	.57%	**10.40%**	**-4.54%**
AC	-1.96%	1.97%	**3.28%**	**-2.87%**
TA	-1.52%	2.50%	**2.26%**	**-2.10%**
TT	-.05%	-1.85%	**10.30%**	**-6.32%**
TG	1.17%	-1.16%	**8.38%**	**-7.21%**
TC	-4.65%	6.14%	**4.01%**	**-3.75%**
GA	0.90%	-1.19%	**6.56%**	**-4.29%**
GT	-2.21%	-4.36%	**13.86%**	**-6.59%**
GG	-3.73%	-3.45%	**14.00%**	**-1.55%**
GC	2.47%	1.79%	**9.78%**	**-9.95%**
CA	-1/13%	-0.86%	**8.27%**	**-9.74%**
CT	-2.37%	-5.79%	**7.62%**	**-7.07%**
CG	3.05%	5.32%	**5.48%**	**-8.74%**
CC	-4.00%	2.25%	**10.28%**	**-12.16%**

Just like in the first-order case, a uniform increase in transitions to G occurs when moving from the C -> G region, but the probability mass is spread out fairly evenly between all xG states. No one transition adequately accounts for the observed bias.

Mutual Information

Since Markov chains fail to explain what is at the root of the GC skew, we now turn to more generic patterns. By treating the DNA as nothing but a text document and regular expressions as "words" in the DNA vocabulary, a very clean measure known as mutual information might help identify features of the sequence that can help explain the seen asymmetries.

Feature selection is done by selecting words that have the highest mutual information with the class variable C, in this case positive for the C region and negative for the G region. W_t is a random variable over all word occurrences. This method calculates the values of terms by sums over word occurrences instead of over documents, so P(c) is the number of word occurrences appearing in documents with class label c divided by the total number of word occurrences; $P(f_t)$ is the number of word occurrences of word w_t divided by the total number of word occurrences; and $P(c,f_t)$ is the number of word occurrences of word w_t that also appear in document with class label c divided by the total number of word occurrences. Average mutual information is the difference between the entropy of class variable, H(C), and the entropy of the class variable conditioned on the absence or presence of the word, $H(C|W_t)$ (Cover and Thomas 1991):

$$\begin{aligned} I(C;W_t) &= H(C)-H(C|W_t) \\ &= -\sum_{c\in C} P(c)\log(P(c)) \\ &\quad + \sum_{f_t\in\{0,1\}} P(f_t)\sum_{c\in C} P(c|f_t)\log(P(c|f_t)) \\ &= \sum_{c\in C}\sum_{f_t\in\{0,1\}} P(c,f_t)\log(\frac{P(c,f_t)}{P(c)P(f_t)}) \end{aligned}$$

This particular formulation works well since the same "words" in DNA appear numerous times in both classes of documents. Also, given that we are searching for the most general explanation of these regions and not bias due to one particular instance of the chromosome, several strains of the E. Coli genome were used as input, each split according to its own C and G regions.

Exact Patterns

Before trying to find generic patterns, the first step is to confirm mutual information can indeed find the most elementary explanation. The first experiment performed tests all permutations of nucleotides for lengths between one to six, referred to here as "exact" patterns. Increasing the length of word beyond this not only substantially lengthens the computation time due to exponential growth of permutations, but it also substantially decreases the expected number of counts, which is in opposition to our goal of finding the most general explanation.

What should we expect to see? If a particular repeat sequence such as "GGGCGG" were the culprit, it should appear near the top of the list. If no such singular exact pattern exists, we should expect to see things like "G" and "GG" having the highest rank as indeed can be seen in the results shown in Table 4. Note that all results shown henceforth were pruned if they failed to pass the Chi-squared test, which accounts for differences in the number of times a particular pattern is expected to appear in a given length of DNA sequence. Rank is derived first by sorting on mutual information, then by the Chi-squared value, and then by length of sequence.

Table 4

Mutual information results for all permutations of nucleotide sequences between length 1-6

Rank		Rank	
1	*G*	11	GGG
2	GG	12	GGCG
3	C	13	CCC
4	CA	14	GTGG
5	CC	15	GGC
6	GTG	16	GT
7	TG	17	GGGG
8	TGG	18	GGGC
9	CAC	19	GA
10	CCA	20	GCCC

By including all possible permutations, we have not let our assumptions bias the outcome, but rather have let mutual information prove itself as an effective formalism. With this positive confirmation using trivial patterns, we can now turn to more interesting trials.

Structure Patterns

Given the strand bias should not interfere with coding of proteins, the only explanation fitting with the genetic code would be a change in the third position of codons since such a mutation typically does not alter the amino acid sequence due to the degenerate nature of the genetic code where ATA, ATT, ATG, ATC can all encode the same thing. Using our generic pattern method, we can test "structure" patterns such as "GOOGOOG", where O can be any nucleotide (using "O" to keep clear alignment in tables to follow). By building many permutations of such periodic structure patterns to search each document, we should finally be able to find different explanations for the G rich region.

Method A of Pattern Construction

For the purposes of reasonable computation time and leveraging the permutation algorithms developed in the testing phase, a simple approach was taken to insert 1-n wild cards between each nucleotide of the previously generated nucleotide sequences. For example, the length three exact permutation "GGG" led to "GOGOG", "GOOGOOG", GOOOGOOOG", etc. up to n wild cards. This does exclude many other possibilities such as "GGOG", GOGG", "GOGOOG", etc..

Method B of Pattern Construction

For longer sequences beyond five nucleotides in length, method A suffers from the same exponential explosion of permutations mentioned with exact patterns, this time even worse due to each exact sequence spawning even more permutations when wild cards of different lengths are included. Also, the execution time to search for regular expressions increases greatly with length and the number of wild cards. Since the goal is to test long structure patterns, "method B" generated all exact permutations of length eleven, and then pruned for those containing greater than an 80% concentration of G. This greatly reduced the number of test patterns while still generating some variation in the mutations of long sequences.

Results

The result of using method A with sequences length 2-5 and number of wild cards 1-8 can be seen in table 5. The total size of such a vocabulary is 10,080.

Table 5

Mutual information results for method A

Rank		Rank	
1	GOG	11	COOC
2	GOOG	12	COCOC
3	GOOGOOG	13	GOOOGOOOG
4	COC	14	GOGOT
5	GOOOG	15	COOCOOC
6	GOGOG	16	GOOOOOOG
7	GOOOOOG	17	TOGOG
8	GOOOOG	18	AOGOG
9	GOOGOOGOOG	19	GOOOOOOOOG
10	GOOOOOGOOOOOG	20	GOOOOOOOOGOOOOOOOOG

The first observation is that G and C patterns rank the highest as is expected. Results 1, 2, 3, 7, 9, 10, 11, 15, 19, and 20 all support the third position mutation bias hypothesis by containing either 1, 2, or 5 wild cards, which leaves the first and second position unaltered. It is also very interesting how 10 and 20 have such high rank though they are much longer patterns, thus supporting the idea that periodic structure plays a key role in the observed bias. The strongest support for third position bias is clearly shown by top three ranking words.

The positive results from method A leads one to wonder how even longer periodic chains rank. As previously mentioned, method B was developed to effectively prune the extremely large vocabulary that can result from such brute force word generation. The second trial used method B with sequences length of 11, number of wild cards 1-5, and a G concentration of at least 80%. In an attempt to make the structure more readable in such long chains, patterns were split into two row entries, where necessary "_" was inserted at the front to help align repeated parts within one sequence.

Table 6

Mutual information results for method B

Rank	
1	GOOGOOGOOGOOTOOGOOGOOGOOGOOG
2	GOOOOOGOOOOOGOOOOOTOOOOOGOOOOOG _OOOOOGOOOOOGOOOOOGOOOOOGOOOOOG
3	GOOOOOGOOOOOGOOOOOGOOOOOGOOOOOG _OOOOOGOOOOOGOOOOOGOOOOOGOOOOOG
4	GOOOGOOOGOOOGOOOGOOOG _OOOTOOOGOOOGOOOGOOOG
5	GOOGOOGOOGOOGOOGOOCOOGOOGOOGOOG

Results 1, 2, 3, and 5 all exhibit a repeated third position structure as can be easily detected visually by the alignment of the Gs. Though space limited the results shown, eight out of the top ten has either two or five wild cards, thus conforming to the same general periodic structure. Though other methods exist, mutual information is a very straight forward formulation to search for any such structured pattern in a generic fashion, here finding good evidence that a third position codon bias is indeed related to the observed large-scale GC skew found in the E. coli chromosome.

Future Work

A more elegant approach would be to build up promising structures from smaller, high ranking seed patterns using genetic algorithms. If an alignment of the longest subsequence were done between pairs of winning seeds, then each round could recombine by keeping conserved subsequences, generalizing any differences, and adding a stochastic element. Many such rounds of competition would more effectively prune the large search space with less prior assumptions made than the methods used here.

References

Cover, T., and Thomas, J. 1991. *Elements of Information Theory*. John Wiley.

Lobry, J.R. 1995. Properties of a general model of DNA evolution under nostrand bias conditions. *J. Mol. Evol*. 40: 326-330.

Lobry, J.R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13: 660-665.

Nikolaou, C., Almirantis, Y. 2005. A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species. *Nucleic Acids Res*. 33: 6816-6822.

Rocha, E.P.C., Danchin, A. 2001. Ongoing evolution of strand composition in bacterial genomes. *Mol. Biol. Evol.* 18: 1789-1799.

Sueoka, N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol*. 40: 318-325.