Introduction to Bioinformatics

Genome Analysis – Sequence contrasts (Part III)

Codon Frequencies

Rationale

How do you relate with any confidence a small fragment (~500 nt, i.e. the size of a single read) to the genome from which it might have come? We've examined thus far two candidate methods: GC-content and dinucleotide biases. You've found that there just isn't enough information in one measure, GC-content, to do the job. Perhaps dinucleotide biases will be better, with 15 separate measures. We'll see. But even if it is, it smacks of magic. Why should dinucleotide biases vary from organism to organism? Why should they remain constant over the span of an entire genome? They must do both, otherwise the biases will not be useful for our task.

In these set of notes, we turn to a different candidate method, codon frequencies. In principle, it provides much more information than either GC-content or dinucleotide biases, and better, one can understand ***why*** it might show variability amongst different organism and ***why*** it ought to be constant within a given organism. Take another look at two codon frequency tables (in the previous set of notes and reproduced on the next page). Consider the six leucine codons and their frequencies in *Borrelia burgdorferi.* Suppose that each of the six codons has a corresponding tRNA that recognizes it (this is an oversimplification, but no harm). If *B. burgdorferi* cells have equal concentrations of the six tRNAs, then one might expect that translation of genes would proceed equally efficiently regardless of what codon is used for leucine. On the other hand, suppose that the six are not in equal concentrations. For example, perhaps the tRNA that recognizes UUA is in the highest abundance. Then those genes that prefer the UUA codon may be translated faster than those that use other leucine codons. In short, if there is an unequal distribution of tRNAs, then there is selection for an unequal distribution of codons. Now suppose that genes happen to have more UUA codons than other leucine codons. Then there is selection to maintain a higher concentration of tRNA that recognizes that codon. So there is a feedback loop: high usage of UUA favors high expression of the tRNA that recognizes it, and vice versa. The situation is much like handedness. If you tend to use your right hand, then you become more proficient in using it, and so you prefer to use your right hand.

It stands to reason, then, that the genes in an organism might all respond to the concentration of tRNAs within the cell, and if those concentrations vary from organism to organism, then we might be able to use codon frequencies to identify the organism from which a gene comes.

Problems in calculating codon frequencies

Using codon frequencies for this purpose pose problems we didn't have to face with either GC-content or dinucleotide bias. First, to determine dinucleotide frequencies, prerequisite for the calculation of bias, it's enough to count each dinucleotide in the DNA you're considering. You can't do that with codons. If you have a piece of DNA that has within it "…CGACTTAG…", does it contain a TTA codon? Yes, if (a) the sequence occurs within a protein-encoding gene and (b) the reading frame of the gene is such that the sequence is read "…C-GAC-TTA-G…". Otherwise, no. To determine codon frequencies, we need to first determine the extent of genes within the DNA under consideration.

***Borrelia burgdorferi*: 2294 CDS's (612759 codons)**

fields: [**triplet**] [amino acid] [fraction] [frequency: per thousand]

UUU	Phe	0.88	48.3	**UCU**	Ser	0.32	24.1	**UAU**	Tyr	0.77	31.6	**UGU**	Cys	0.68	4.9
UUC	Phe	0.12	6.3	**UCC**	Ser	0.05	3.4	**UAC**	Tyr	0.23	9.2	**UGC**	Cys	0.32	2.3
UUA	Leu	0.41	41.5	**UCA**	Ser	0.24	17.6	**UAA**	*	0.65	2.4	**UGA**	*	0.16	0.6
UUG	Leu	0.16	16.3	**UCG**	Ser	0.03	2.3	**UAG**	*	0.19	0.7	**UGG**	Trp	1.00	4.4
CUU	Leu	0.28	29.0	**CCU**	Pro	0.42	10.0	**CAU**	His	0.73	8.6	**CGU**	Arg	0.07	2.1
CUC	Leu	0.02	2.3	**CCC**	Pro	0.15	3.7	**CAC**	His	0.27	3.2	**CGC**	Arg	0.04	1.1
CUA	Leu	0.10	10.6	**CCA**	Pro	0.37	8.9	**CAA**	Gln	0.84	22.8	**CGA**	Arg	0.06	1.8
CUG	Leu	0.03	2.7	**CCG**	Pro	0.06	1.3	**CAG**	Gln	0.16	4.2	**CGG**	Arg	0.02	0.5
AUU	Ile	0.54	53.1	**ACU**	Thr	0.39	17.4	**AAU**	Asn	0.80	60.0	**AGU**	Ser	0.22	16.5
AUC	Ile	0.07	7.2	**ACC**	Thr	0.12	5.6	**AAC**	Asn	0.20	15.1	**AGC**	Ser	0.14	10.4
AUA	Ile	0.39	38.0	**ACA**	Thr	0.44	19.9	**AAA**	Lys	0.80	87.8	**AGA**	Arg	0.65	20.1
AUG	Met	1.00	18.1	**ACG**	Thr	0.05	2.2	**AAG**	Lys	0.20	22.2	**AGG**	Arg	0.18	5.5
GUU	Val	0.55	27.9	**GCU**	Ala	0.44	21.2	**GAU**	Asp	0.79	42.0	**GGU**	Gly	0.28	13.7
GUC	Val	0.05	2.4	**GCC**	Ala	0.11	5.1	**GAC**	Asp	0.21	11.3	**GGC**	Gly	0.16	7.7
GUA	Val	0.30	15.1	**GCA**	Ala	0.39	18.9	**GAA**	Glu	0.75	53.9	**GGA**	Gly	0.41	20.0
GUG	Val	0.11	5.4	**GCG**	Ala	0.06	2.7	**GAG**	Glu	0.25	17.8	**GGG**	Gly	0.15	7.4

Coding GC 29.27% 1st letter GC 38.52% 2nd letter GC 28.30% 3rd letter GC 21.01%

***Mycobacterium tuberculosis CDC1551*: 4187 CDS's (1329826 codons)**

fields: [**triplet**] [amino acid] [fraction] [frequency: per thousand]

UUU	Phe	0.21	6.2	**UCU**	Ser	0.04	2.3	**UAU**	Tyr	0.30	6.1	**UGU**	Cys	0.26	2.4
UUC	Phe	0.79	22.9	**UCC**	Ser	0.21	11.6	**UAC**	Tyr	0.70	14.5	**UGC**	Cys	0.74	6.9
UUA	Leu	0.02	1.7	**UCA**	Ser	0.07	3.8	**UAA**	*	0.15	0.5	**UGA**	*	0.55	1.7
UUG	Leu	0.19	18.1	**UCG**	Ser	0.35	19.5	**UAG**	*	0.30	1.0	**UGG**	Trp	1.00	14.8
CUU	Leu	0.06	5.6	**CCU**	Pro	0.06	3.6	**CAU**	His	0.29	6.6	**CGU**	Arg	0.12	8.7
CUC	Leu	0.18	17.2	**CCC**	Pro	0.29	17.0	**CAC**	His	0.71	16.0	**CGC**	Arg	0.38	28.7
CUA	Leu	0.05	4.8	**CCA**	Pro	0.11	6.4	**CAA**	Gln	0.26	8.2	**CGA**	Arg	0.10	7.6
CUG	Leu	0.51	49.7	**CCG**	Pro	0.54	31.7	**CAG**	Gln	0.74	22.9	**CGG**	Arg	0.33	24.9
AUU	Ile	0.15	6.5	**ACU**	Thr	0.07	3.8	**AAU**	Asn	0.21	5.2	**AGU**	Ser	0.07	3.7
AUC	Ile	0.79	33.4	**ACC**	Thr	0.59	34.6	**AAC**	Asn	0.79	19.4	**AGC**	Ser	0.26	14.6
AUA	Ile	0.05	2.3	**ACA**	Thr	0.08	4.8	**AAA**	Lys	0.26	5.4	**AGA**	Arg	0.02	1.4
AUG	Met	1.00	18.6	**ACG**	Thr	0.27	15.7	**AAG**	Lys	0.74	15.1	**AGG**	Arg	0.05	3.4
GUU	Val	0.10	8.2	**GCU**	Ala	0.08	11.2	**GAU**	Asp	0.28	15.9	**GGU**	Gly	0.19	18.6
GUC	Val	0.38	32.4	**GCC**	Ala	0.45	59.0	**GAC**	Asp	0.72	41.9	**GGC**	Gly	0.51	49.3
GUA	Val	0.06	4.9	**GCA**	Ala	0.10	13.0	**GAA**	Glu	0.35	16.2	**GGA**	Gly	0.10	10.0
GUG	Val	0.47	40.1	**GCG**	Ala	0.37	48.4	**GAG**	Glu	0.65	30.4	**GGG**	Gly	0.20	18.9

Coding GC 65.77% 1st letter GC 67.82% 2nd letter GC 50.22% 3rd letter GC 79.27%

Tables 1 and 2. Codon usage in *Borrelia burgdorferi* and *Mycobacterium tuberculosis* derived from sequence analysis of each genome. The number of coding sequences "CDS's" and codons from which these frequencies were derived are indicated above each table. [fraction]-proportion of occurrences of a particular amino acid encoded by a particular codon. For each amino acid, the fractions associated with each codon sum to 1. [frequency: per thousand]-the number of times each codon was used per genome ÷ 1000. * -stop codon.

Second, codons are tied to the amino acids they encode. If a protein requires a tryptophan residue in a certain position, you may be sure that the gene will contain TGG, because that's the only tryptophan codon that exists! A fairer assessment of codon frequency may therefore be achieved by confining the analysis to each of the twenty amino acids. We could ask of an organism: What is the relative frequency of alanine codons ***amongst all four possible alanine codons***? What is the relative frequency of asparagine codons ***amongst both possible asparagines codons***? And so forth. Calculating this way, there will be no information to be gained looking at methionine or tryptophan codons, because the relative frequency will always be 100%. These amino acids have only one codon, and so all methionines and tryptophans will be encoded by the same codons.

Now we can appreciate better the two values given in the codon tables on the previous page. The fraction is the relative frequency, the frequency considering only one amino acid. The number, when divided by 1000, is the absolute frequency, the frequency considering all twenty amino acids.

SQ1. From the absolute frequencies of codons in *Borrelia bergdorferi*, what are the amino acids (not codons) that most commonly appear in proteins? What are the least commonly appearing amino acids? How about *Mycobacterium tuburculosis*?

How to calculate relative codon frequencies?

Now let's return to our article:

> Samuel Karlin (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends in Microbiology 9:335-343.

focusing on Box 3, which describes how Karlin calculated relative codon frequencies and a measure of codon bias. As with dinucleotides and Box 2, there are two parts: (1) Calculate the relevant measure (dinucleotide bias in Box 2 and relative codon frequencies in Box 3), and (2) Compare the measure calculated from a DNA fragment with that calculated from a genome.

How to calculate the relative codon frequencies? The expression at the top of Box 3 may look mysterious, but taken piece by piece, it's not so bad. It says the sum (Σ; capital sigma) of something is equal to 1. The something $g(\ldots)$ is the average codon frequency for a triplet (x,y,z) in some collection of genes (called g). Well, codons ARE triplets. Perhaps the codon GAC can be represented (G, A, C). So x is the letter in the first position, y in the second, z in the third. We also learn that the sum is to be taken for all possible letters such that (x,y,z) encodes a, a specific amino acid. In plain English, if you add up all the average codon frequencies for each codon encoding a specific amino acid, it better add up to one. But that won't happen without some help.

SQ2. What is the sum of the absolute codon frequencies of the two phenylalanine codons of *Borrelia bergdorferi*?

But Box 3 also tells us that the frequencies should be normalized. Normalized means divided by some factor.

SQ3. By what factor must the absolute codon frequencies of the two phenylalanine codons of *Borrelia bergdorferi* be divided by in order for the sum to come to 1?

SQ4. What two normalized average frequencies do you get if you do that division?

Karlin's expression simply describes the relative frequencies of the type shown on the previous page.

How to compare sets of relative codon frequencies?

The next step is to find a way of comparing sets of relative codon frequencies. If you do so by eye, comparing the frequencies of *B. bergdorferi* with those of *M. tuburculosis*, it's clear that the two sets are very different. But how do we move from a fuzzy sense of difference to a quantitative measure of difference? One straightforward thing to do is to subtract each relative codon frequency of one organism from the corresponding frequency of the other and add up the differences.

SQ5. Do this for the cysteine codons: What is the sum of the differences between the codons for the two organisms?

If you did the subtraction consistently, then you see a problem. The sum is zero. It MUST be zero, because one codon has a higher frequency in *B. bergdorferi*, and the other has an equally higher frequency in *M. tuburculosis*. Nonetheless, we'd like to say that the two organisms are different with respect to their phenylalanine codon usage. The solution is simple: Add the differences, but only after taking the absolute value.

SQ6. What is the sum of the absolute differences between the codons for the two organisms?

We could repeat this operation for all 20 amino acids and add up all the absolute differences. But that brings us to a second computational problem. The sum of the absolute differences of the cysteine codons is about the same as the absolute differences of the glutamate codons. But there are about five or ten times more glutamate codons than cysteine codons. You'd think that glutamate codons should contribute five or ten times as much to the total comparison.

SQ7. How did I get "…five or ten times more glutamate codons than cysteine codons"?

Now look at the second expression in Box 3. I'm supposed to add something (Σ, Sigma, again), and the something is the absolute value of one thing subtracted from another… that should sound familiar.

SQ8. From the foregoing discussion, how do you translate the second expression into English?

Overview of calculation of codon bias in BioBIKE

If you really know what these quantities are, you should be able to calculate them yourself. But who has the time to do what amounts to 100's of computations? Answer: computers! As always with a complex computational task, the trick is to break it up into smaller, simple tasks. I'll break it up this way:

A. Calculate a table containing absolute codon frequencies
   1. Break the DNA up into codon triplets
   2. Count how many total triplets there are
   3. For each of the 64 possible triplets
      a. Count the triplet
      b. Divide the count by the total number of triplets
      c. Save that fraction in the table
   4. Package all this into a function that takes DNA and returns a table of absolute codon frequencies

B. Calculate a table containing relative codon frequencies
   1. Calculate a table containing absolute codon frequencies
   2. For each of the 20 possible amino acids
      a. Determine the sum of the absolute codon frequencies for the amino acid
         i. For each of the possible codons for that amino acid
            Sum the absolute codon frequency for that codon
      b. For each of the possible codons for that amino acid
         i. Get the absolute codon frequency for that codon
         ii. Divide that frequency by the sum of the absolute codon frequencies
         iii. Save that fraction in the table
   3. Package all of this into a function that takes DNA and returns a table of relative codon frequencies

C. Compare two relative codon frequency tables
   1. Calculate $Table_1$ containing relative codon frequencies for DNA_1
   2. Calculate $Table_2$ containing relative codon frequencies for DNA_2
   3. Translate DNA_1 to get a list of amino acids
   4. Calculate the frequencies of the amino acids
   5. For each of the 61 amino-acid encoding codons
      a. Determine the amino acid for that codon
      b. Look up the frequency of that amino acid
      c. Determine the relative codon frequency for that codon from $Table_1$
      d. Determine the relative codon frequency for that codon from $Table_2$
      e. Determine the absolute difference of the two frequencies
      f. Multiply that absolute difference by the amino acid frequency
      g. Sum the product
   6. Package all of this into a function that takes two segments of DNA and returns a single value, what Karlin terms the codon bias of DNA_1 relative to DNA_2.

SQ9. Try to translate this outline into something BioBIKE can understand, using as raw material two genes of your choosing. What problems do you encounter?