Introduction to Bioinformatics

Genome Analysis – Sequence contrasts

Dinucleotide and Codon Frequencies

The previous set of notes introduced the article

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

but focused only on one method described in it: the detection of genomic islands by differences in GC-frequency. Unfortunately, we found that what worked well with 50-Kb-sized fragments did not work well at all with much smaller fragments, those about the size you might expect to find from a metagenome project.

Here we'll discuss two other analytical methods considered by Karlin (2001): comparisons of dinucleotide biases (which he calls genomic signatures) and comparisons of codon frequencies.

Dinucleotide Biases

Dinucleotide? Why start there? What happened to mononucleotide? Well, we've already considered mononucleotide frequencies. Since in double-stranded DNA [A] = [T], [G] = [C], and [A]+[G] = [T]+[C], only one piece of information is all that is needed to describe the frequencies of nucleotides. You can express it as [A] or as [G] or as [G]+[C] (i.e., the GC-frequency), but the information content is the same.

If there's too much information for this single measurement to identify an organism from a short (~500-nt) fragment, maybe we'll have better luck with multiple measurements. Dinucleotides give you 15 independent frequencies for a DNA fragment. Maybe that ***will*** provide enough information to do the job.

SQ1. Where did the number 15 come from?

SQ2. Use a BioBIKE function to generate all possible dinucleotides.

SQ3. Find the counts of each dinucleotide in PMed4.chromosome

SQ4. Convert the counts to frequencies.

SQ5. Relate the frequencies to the dinucleotides. Which dinucleotides have the highest frequencies? Which have the lowest? Does this make sense? What is the GC fraction of PMed4?

Unfortunately, it should make all too much sense. The dinucleotide frequencies at first glance tell us not much more than we already knew from GC-frequencies. But there's a lot more information hidden, waiting to get out. To see it, we need to take into account the frequencies of the individual nucleotides. Of course AA has a high frequency in PMed4, since A has a high frequency. What we want to know is whether AA has a ***higher*** frequency than one would expect ***given*** the frequency of A. The expected frequency of AA is *frequency*[A] * *frequency*[A].

SQ6. What is the frequency of AA in PMed4? How close is it to the expected frequency? What is the *ratio* of the frequency to the expected frequency?

SQ7. Repeat the previous question, but considering GG dinucleotide.

Now we have a way of giving a fair measure of dinucleotide frequencies. Dividing the frequency by the expected frequency (based on mononucleotide frequencies) tells us how unusual the dinucleotide is ***discounting the organism's biases for A, C, G, and T***. Not surprisingly, someone has thought of all this before, Sam Karlin for one. Look at Box 2 of the article (remember the article?) and without swallowing your eyeballs, consider the equation on the left side of the box. For the most part, it should look familiar, especially if you ignore the detail about concatenating the inverted complement of the sequence (i.e. considering both strands of DNA). We'll call the ratio of observed to expected frequencies the ***bias***.

SQ8. Calculate the dinucleotide frequencies and dinucleotide biases for two organisms: a low GC organism like *Prochlorococcus* PMed4 and a high GC organism like *Prochlorococcus* P9313. How do they differ from one another?

Comparisons of Dinucleotide Biases

Back to the main question: Can this measure be useful in identifying pieces of DNA that are part of larger pieces of DNA. Of course Karlin would ask the opposite question, whether it can be useful in identifying pieces of DNA that are ***foreign*** to a larger piece of DNA. We addressed this question with GC-fraction by looking at the range of values in a genome as the fragment size got progressively smaller. We could do the same thing with dinucleotide biases, but… how do you compare ***sixteen*** numbers? Do we need sixteen separate graphs for each of the sixteen dinucleotides? I hope not! It would be far better if we could come up with a single number that combines the information from the sixteen dinucleotides. Given the dinucleotide biases for a fragment of DNA, we'd like to know how close they are to the biases of a different fragment of DNA (perhaps the genome it's part of). How to make that comparison?

Now return to Box 2, this time focusing on the equation on the right side of the box. Read the description of the equation (which lies both before and after the equation)

SQ9. Can you make sense of the equation? It's the right side of the equation that's important. The left side is just a name.

SQ10. Write a function that will take dinucleotide biases and from them derive a single number representing their similarity.

Why should organisms maintain constant dinucleotide biases over their genomes? I've never heard a convincing explanation, but evidently they do!

Codon usage contrasts

It's a fact of life, at least life on earth, that most amino acids can be encoded by more than one codon. How does an organism decide which one to use? On one hand, the choice makes no difference. A phenylalanine encoded by UUU is just as much a phenylalanine as one encoded by UUC. Nonetheless, organisms ***do*** make choices and their choices are definitely not random. The codon usage chart on the next page gives an example of the highly divergent choices of two bacteria: *Borrelia burgdorferi* and *Mycobacterium tuburculosis*. Note that the table is given in terms of RNA codons. You can transform them into DNA codons by replacing U with T. The first number for each codon is the ratio of the instances of that codon to the total number of codons for that amino acid. The second (less useful) number is the number of times that codon is used per 1000 codons.

***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.

SQ11. Choose an organism and list the two triplets that code for Lys. What is fraction of total lysine codons is taken up by each of the two codons? What is the sum of the two fractions? Why?

SQ12. Compare the codon usage for lysine between *Borrelia burgdorferi* and *Mycobacterium tuberculosis*? How do they differ?

SQ13. Consider the two tables as a whole. Which organism do you predict has a higher GC fraction?

With dinucleotide biases, we had the problem of how to combine 16 pieces of information. With codon usage, we have that same problem multiplied by a factor of 4! There are many possible ways to compare codon usage. Karlin's Box 3 shows one way.

SQ14. Consider Box 3 and the equation at the bottom left of the box. What sense can you make out of it?

We'll discuss how to compare codon usage soon.