BIOL591: Introduction to Bioinformatics
Detection of anomalous regions of a genome

Outline:

I. Pathogenicity islands and their detection
   A. The problem
   B. Comparison of G+C content
   C. Genome signature contrasts
   D. Codon usage contrasts
   E. Detection of pathogenicity islands
II. Identification of alien genes by codon usage contrasts
III. Markov models

I. Pathogenicity islands and their detection

I.A. The problem

It is now clear that many bacterial pathogens differ from their nonpathogenic relatives because of the recent addition of foreign DNA, typically introduced by bacteriophages (viruses). These foreign regions carrying genes important in pathogenesis are called pathogenicity islands (PAIs), illustrated in Fig. 1. There is thus clear practical reason to be able to identify which DNA in a bacterial genome is foreign and which is native. You might skim Hentschel and Hacker (2001) available from the unit web page to learn more about PAIs.

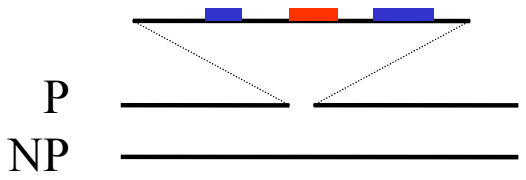

Fig.1 : Representation of a pathogenicity island. A large block of DNA is found in a pathogenic strain (**P**) but not a related nonpathogenic strain (**NP**). The insertion contains genes related by bacteriophage function (blue) and one gene (red) contributing to the pathogenesis of **P**.

Given the sequence of a bacterial genome, how can one identify PAIs? You might think that one straightforward approach would be to compare the sequences of a pathogenic bacterium and a nonpathogenic relative. We did something like that in Scenario 1. There are problems with this approach, however. First, one does not always have a good nonpathogenic relative for comparison. Second, such a comparison does not distinguish foreign DNA that was recently gained from native DNA that was recently lost.

I.B. Comparison of G+C content

What we would like is some method of looking at a region of the genome and recognizing it as foreign just from internal clues. We know that DNA from a certain organism has a typical value for the fraction of nucleotides that are G+C. For example, you'll recall from Scenario 2 that *Anabaena* DNA typically has a G+C fraction of about 42%. That fraction in *E. coli* DNA is closer to 50%. If you encountered a region of *E. coli* DNA that had a G+C fraction far away from 50% you might well be suspicious of its origin.

While analysis of G+C content has helped in identifying some PAIs, you can see that it isn't a very good method. There is significant variation in the G+C fraction even within native DNA, and so the deviation of the foreign DNA must be greater than that to be detected. Furthermore,

it's entirely possible for foreign DNA to originate in an organism with a similar G+C content as the organism you're working with. Then the analysis of G+C fraction will be of no use.

Sam Karlin and his colleagues have developed more sensitive tools to detect PAIs in bacterial genomes. Download the article Karlin (2001) from the link on the unit web page, and let's go through it.

Box 1 of Karlin (2001) lists four methods that might be used to identify PAIs (the fifth element of the list, *Putative alien gene clusters*, is just a variation on *Codon usage contrasts*). The first method, *Compositional contrasts*, is simply comparing the G+C fraction within a chunk of DNA to the overall G+C fraction. The other three methods are discussed in Boxes 2 through 4, and we'll consider the first two below. The fourth method, amino acid bias, is not widely used (and strikes me as peculiar). I won't consider it further here.

I.C. Genome signature contrasts (Box 2)

The name of this measure as well as the description may give you the idea that it is very complicated and difficult to understand. Not at all. It is just G+C fraction extended to two nucleotides rather than one. If one considers all the DNA in an organism, one need only determine the fraction of a single nucleotide to determine the fraction of them all, because the fractions of G and C will be equal, as will the fractions of A and T. How many ***di***nucleotide

genome (available DNA)	CG	GC	TA	AT	CC GG	TT AA	TG CA	AG CT	AC GT	GA TC	G+C
Escherichia coli (4.6Mb)*	1.16	1.28	0.75	1.10	0.91	1.21	1.12	0.82	0.88	0.92	51%
Haemophilus influenzae (1.8Mb)*	1.09	1.43	0.75	0.95	1.01	1.25	1.12	0.82	0.85	0.87	38%
Neisseria gonorrhoeae (877kb)	1.32	1.26	0.63	1.05	0.99	1.50	0.99	0.67	0.83	0.89	53%
Neisseria meningitidis (2.2Mb)	1.31	1.27	0.64	1.05	0.96	1.44	1.01	0.70	0.84	0.91	52%
Rhodobacter capsulatus (1.4Mb)	1.19	1.19	0.33	1.61	0.88	1.30	1.03	0.84	0.71	1.16	67%
Rickettsia prowazekii (1.1Mb)*	0.77	1.53	0.98	0.98	1.03	1.05	1.02	1.06	0.86	0.91	29%
Helicobacter pylori (1.7Mb)*	0.93	1.56	0.73	0.86	1.17	1.37	0.97	0.97	0.67	0.87	39%
Campylobacter jejuni (1.6Mb)*	0.62	1.75	0.77	0.83	1.11	1.25	1.03	1.09	0.71	0.92	31%
Bacillus subtilis (4.2Mb)*	1.04	1.27	0.65	1.02	0.97	1.24	1.08	0.91	0.75	1.06	44%
Streptococcus pyogenes (985kb)	0.71	1.19	0.76	0.89	1.04	1.17	1.12	1.04	0.86	0.99	39%
Clostridium acetobutylicum (4.0Mb)	0.45	1.28	0.93	0.95	1.22	1.08	1.02	1.12	0.81	0.97	31%
Streptomyces coelicolor (2.4Mb)	1.14	0.97	0.51	0.93	0.88	0.82	1.00	0.95	1.14	1.25	72%
Mycobacterium leprae (1.7Mb)	1.13	1.07	0.75	1.10	0.88	1.04	1.14	0.86	1.05	1.02	58%
Mycobacterium tuberculosis (4.4Mb)*	1.18	1.07	0.58	1.24	0.86	1.05	1.11	0.80	1.05	1.08	65%
Mycoplasma genitalium (580kb)*	0.39	1.19	0.75	0.77	1.13	1.23	1.16	1.06	0.96	0.89	32%
Mycoplasma pneumoniae (816kb)*	0.82	1.14	0.77	0.71	1.12	1.30	1.08	0.96	1.02	0.81	40%
Synechocystis sp. (3.6Mb)*	0.75	1.02	0.75	1.00	1.36	1.32	1.05	0.85	0.79	0.86	48%
Deinococcus radiodurans (3.0Mb)	1.07	1.16	0.49	0.89	0.87	1.24	1.12	1.00	0.93	1.01	67%
Treponema pallidum (1.1Mb)*	1.08	1.22	0.74	0.93	0.86	1.18	1.13	0.94	0.96	0.95	53%
Borrelia burgdorferi (911kb)*	0.48	1.47	0.77	0.88	1.29	1.22	1.02	1.07	0.69	1.01	29%
Chlamydia trachomatis (1.0Mb)*	0.79	1.12	0.77	0.89	1.01	1.16	0.96	1.18	0.76	1.15	41%
Aquifex aeolicus (1.6Mb)*	0.87	0.75	0.82	0.66	1.24	1.29	0.74	1.18	0.89	1.12	43%
Methanococcus jannaschii (1.7Mb)*	0.32	1.12	0.83	0.94	1.38	1.14	1.03	1.11	0.72	1.05	31%
Methanobacterium thermoautotrophicum (1.8Mb)*	0.51	0.76	0.74	1.13	1.25	0.95	1.17	1.07	0.85	1.14	50%
*Archaeoglobus fulgidus (2.2Mb)**	0.78	1.02	0.61	0.86	1.04	1.21	1.01	1.17	0.77	1.19	49%
Pyrococcus horikoshii (1.7Mb)*	0.61	0.89	0.90	0.92	1.30	1.11	0.85	1.22	0.73	1.13	42%
Pyrobaculum aerophilum (2.2Mb)*	0.97	1.15	1.07	0.93	1.10	1.18	0.86	1.06	0.83	0.90	51%
human (5.8Mb)	0.25	1.00	0.74	0.88	1.25	1.12	1.20	1.17	0.83	0.99	43%
mouse (1.1Mb)	0.22	0.95	0.72	0.80	1.19	1.08	1.24	1.26	0.88	1.01	46%
Drosophila melanogaster (4.3Mb)	0.94	1.29	0.75	0.97	1.08	1.23	1.12	0.87	0.84	0.90	41%
Caenorhabditis elegans (74Mb)	0.97	1.04	0.62	0.86	1.05	1.28	1.09	0.90	0.86	1.09	36%
yeast (12Mb)*	0.80	1.02	0.77	0.94	1.06	1.14	1.10	0.99	0.89	1.06	38%
Arabidopsis thaliana (2.0Mb)	0.72	0.93	0.74	0.90	1.03	1.13	1.11	1.04	0.91	1.11	36%
Plasmodium falciparum (947kb)	0.74	0.93	0.99	1.07	1.54	1.00	1.10	0.83	0.92	0.97	20%

* indicates complete genome

<0.50 | 0.50–0.70 | 0.70–0.78 | 0.78–1.23 | 1.23–1.30 | 1.30–1.50 | >1.50

Fig. 2: Dinucleotide frequencies from the sequenced genomes of several organisms. The relative frequency (r*) of each dinucleotide is shown and those frequencies deviating from random expectation (i.e. one) y more than 22% are highlighted as indicated in the bar below the table. *(Taken from Alan Campbell, Jan Mrazek, and Sam Karlin (1999). Proc Natl Acad Sci USA 96:9184-9189)*

frequencies must be determined? There are sixteen possible dinucleotides (see Fig. 2), but most of them can be grouped as complementary pairs. Thus, dinucleotide CC and GG are paired as are AC and GT. The four palindromic dinucleotides, AT, CG, GC, and TA, pair with themselves.

SQ1. What fraction of dinucleotides of a random DNA sequence would you expect to be AT?

SQ2. Suppose you homogenized the genome of *Anabaena*. What fraction of dinucleotides of the resulting sequence would you expect to be AT?

Clearly, the expected frequency for a dinucleotide depends on its nucleotide content and the frequency of those individual nucleotides. Karlin defines the ***dinucleotide relative abundance*** (ρ^*_{XY}) as the observed frequency of a dinucleotide divided by the expected frequency of that dinucleotide, calculated as:

(1) $\rho^*_{XY} = f^*_{XY} / f^*_X f^*_Y$

where f^*_{XY} is the observed frequency of the dinucleotide XY in the genome, f^*_X is the observed frequency of X, the first nucleotide of the dinucleotide, f^*_Y is the observed frequency of Y, and the second nucleotide (the * indicates that the observations are over ***both*** strands of DNA).

SQ3. Calculate ρ^*_{AA} in the following 50-nt sequence:
TGATGACAGTCGATTTTTCGGTAGGATAACTGCCATGCCTCTCAAAGTAC
(the answer I got was 0.911) (not 0.888, not 0.893, and not 0.906)

What Karlin calls the genome signature profile is the set of ρ^*_{XY}, all ten of them. Fig 2 shows the genome signature profile for a variety of genomes. You'll note that there are some dinucleotides are highly biased, i.e. their relative abundance differs markedly from the random expectation of 1 (same abundance as a random sequence would have). The gross deficiency of the CG dinucleotide in mammalian genomes is well known and is the result of methylation of CG by mammals, important in gene regulation and imprinting.

The genome signature profile is used to assess foreignness of a block of DNA by comparing the profile of the block to the profile of the entire genome, one dinucleotide at a time. Karlin expresses the procedure to calculate the difference, $\delta^*(f, g)$, between the signature profile of sequence f and sequence g as:

(2) $\delta^*(f, g) = (1/16) \Sigma \, | \rho^*_{XY}(f) - \rho^*_{XY}(g) |$

The sum of the differences is divided by the number of dinucleotides types (i.e. multiplied by 1/16) to express the sum as an average. For example, if you wanted to compare the difference between the relative dinucleotide frequencies of humans and mice, you would begin:

(3) $\delta^*(human, mouse) = (1/16) \Sigma \, | \rho^*_{XY}(f) - \rho^*_{XY}(g) |$

$= (1/16) \{ \, | \rho^*_{CG}(human) - \rho^*_{CG}(mouse) |$

$+ \quad | \rho^*_{GC}(human) - \rho^*_{GC}(mouse) |$

$\ldots$

$+ \quad 2 * | \rho^*_{CC}(human) - \rho^*_{CC}(mouse) |$

$\ldots$

$+ \quad 2 * | \rho^*_{GA}(human) - \rho^*_{GA}(mouse) | \}$

which equals 0.049. Note that it is equivalent (and easier) to add $2 * \rho^*_{CC}$ rather than add $\rho^*_{CC} + \rho^*_{GG}$, because ρ^*_{CC} is equal to ρ^*_{GG} (because ρ^* looks at both strands, and for every GG on one strand, there is a CC on the other).

SQ4. Calculate d* (*Escherichia coli, Haemophilus influenzae*) and d* (*Escherichia coli, Pyrobaculum aerophilum*). (Note that the GC% of *E. coli* and the archaebacterium *P. aerophilum* are both 51%).

SQ5. If a fragment of DNA from *H. influenzae* and *P. aerophilum* were inserted into the genome of *E. coli*. Which one would be flagged as the more foreign by the method of genome signature profiles?

I.D. Codon usage contrasts (Box 3)

The 61 sense codons (64 minus 3 stop codons) are spread over 20 amino acids, and most amino acids are encoded by more than one codon. We saw in Scenario 1 that different organisms have different preferences when it comes to encoding the amino acids. Fig. 3 may serve as a reminder of the comparison between the full codon usage tables of two bacteria (see notes for _).

B. burgdorferi				*M. tuberculosis*			
CAU	His	0.73	8.6	**CAU**	His	0.29	6.6
CAC	His	0.27	3.2	**CAC**	His	0.71	16.0
CAA	Gln	0.84	22.8	**CAA**	Gln	0.26	8.2
CAG	Gln	0.16	4.2	**CAG**	Gln	0.74	22.9

Fig. 3: Comparison of codon usage between two bacteria. Usage for four codons is shown for *Borrelia burgdorferi* and *Mycobacterium tuberculosis*. Each line gives the triplet codon, amino acid abbreviation, relative abundance of codon within family of synonymous codons, and frequency of codon (per 1000).

Given the codon usage tables for *B. burgdorferi* and *M. tuberculosis*, you could probably do a good job of distinguishing genes taken from one of the two organisms from genes taken from the other. Karlin quantitates the different biases (**B**) in codon usage between two organisms in much the same way as he quantitated differences in patterns of dinucleotides frequencies. The bias in codon usage in one set of genes (**F**) relative to another set of genes (**G**) is given as:

(4) $\mathbf{B}(\mathbf{F}|\mathbf{G}) = \Sigma\, \mathbf{p}_{\mathbf{a}}(\mathbf{F})\ \{\ \Sigma\ |\boldsymbol{f}(x,y,z) - \boldsymbol{g}(x,y,z)|\ \}$

Where $\boldsymbol{f}(x,y,z)$ and $\mathbf{g}(x,y,z)$ are the relative abundance of codon *xyz* within the genes of set **F** and set **G**, respectively, and $\mathbf{p}_{\mathbf{a}}(\mathbf{F})$ is the frequency of amino acid **a** within the genes of set **F**. For example, to calculate **B**(*B. burgdorferi* | *M. tuberculosis*), you would go through each of the 20 amino acids and calculate:

(5) $\mathbf{p}_{\mathbf{a}}(\mathbf{F})\ \{\ \Sigma\ |\boldsymbol{f}(x,y,z) - \boldsymbol{g}(x,y,z)|\ \}$

The factor $\mathbf{p}_{\mathbf{a}}(\mathbf{F})$ is analogous to 1/16 in equation (2) and serves to cast the sum of the differences as a weighted average, taking account the different frequencies of amino acids in real protein. For the amino acid histidine, the calculation would look like this:

(6) $(11.8/1000)\ \{\ |\ 0.73 - 0.29\ |\ +\ |\ 0.27 - 0.71\ |\ \}\ =\ .010384$

SQ6. Calculate the part of the calculation of B(*B. burgdorferi* | *M. tuberculosis*) shown in expression (5) above using values for the amino acid glutamine (gln).

F and **G** could be any set of genes, but within the context of the article, the first set is always the genes within a small region of the genome of an organism and the second set is always the set of all genes of the same organism.

I.E. Detection of pathogenicity islands

Figure 2 in Karlin (2001) shows how these measures of foreignness can be applied to detect pathogenicity islands within a genome. For each measure of foreignness, a window either 20 kb or 50 kb was moved along the genome, and the contents of genes within the window was

compared to the sum of all genes in genome. For example, panel (**a**) of the figure shows an analysis of the genome of *Vibrio cholerae*, the causative agent of cholera. The known pathogenicity island (labeled **C** in the figure) was identified by all the methods used, but each method also identified other regions as deviating from the norm.

II. Identification of alien genes by codon usage contrasts

It is important to note the scale of the x-axis in Figure 1 of Karlin (2001). The distance between the tick marks represents 0.1 Mbase or 100,000 nucleotides. Compare this with the size of a typical gene: 1000 nucleotides. Clearly, a stray alien gene would be too small to be seen in the graphs nor detected within a window 20- to 50-times larger. The methods described act on large chunks of DNA, many tens of genes at once. How can we identify whether an ***individual*** gene is foreign or not?

Mrazek et al (2001), available from the unit web site, adapted the method of codon contrasts to look at the foreignness of individual genes (actually, most of the article is concerned with detection of highly expressed genes, but we'll focus on the part of concern to us). The authors used equation 4 to determine different biases in codon usages between a gene (***g***) and four different gene sets:

B(***g***|***C***) = Difference in bias between the gene and all genes
B(***g***|***RP***) = Difference in bias between the gene and genes encoding ribosomal proteins
B(***g***|***CH***) = Difference in bias between the gene and genes encoding chaperones
B(***g***|***TF***) = Difference in bias between the gene and genes encoding factors involved in transcription and translation

The RP, CH, and TF special sets were chosen because in many if not all bacteria these genes are highly expressed. A gene was judged to be putatively alien (**PA**) if it had distinctly different codon usage compared both to highly expressed genes (**B**(***g***|***RP***), **B**(***g***|***CH***), and **B**(***g***|***TF***) are all high) and to average genes (**B**(***g***|***C***) is also high). Their choice of how different the codon usage has to be seems pretty arbitrary to me.

The article, unfortunately, says very little about genes that are **PA**, and after a request to one of the authors for more information went unanswered, I decided to reproduce their results myself. The program I wrote does not exactly follow their procedure (since I did not know exactly what genes they used in the special sets), but the results I got for highly expressed genes were close to theirs. I was disappointed, however, that several genes were apparently **PA** that did not seem to me likely at all to be alien to the bacterium. The 100 most alien genes, according to the described method included three ribosomal proteins, and seven proteins involved in photosynthesis, all ten having orthologs in other cyanobacteria. I don't believe these genes are really alien.

For this reason, I was not completely happy with the method and sought an alternative.

III. Markov models

All three methods described above look at a small slice of the information contained in a DNA sequence. G+C fraction throws away any positional information. Dinucleotide frequencies obviously don't look beyond dinucleotides, and codon bias ignores any information except that related to the genetic code. An alternative method, Markov analysis, takes a broader view of information. Before describing it, let me show you Markov analysis in action.

Consider the following excerpts:

> *But you to other galled eyes, and this author of thousand moments; while most on to tell uses offence 'twill stale or there!*
>
> *That wickled all though thin infines, my somes life itsell; And wift, anday to to beasy;*
>
> *To be as into night, what fried indeed willion, Or lord; that I am not shall a said was it rant.*
>
> *Hold it truly; it escoted? Will my her in you know a knave.*

Who wrote these words? Francis Bacon? Christopher Marlowe? Edward de Vere? Perhaps even ***The Bard?*** Actually, they were written by a Dell PC running Hamlet.pl (after having digested all words uttered by Hamlet). At times the texts produced seem like monkeys at a typewriter, at other times Hamlet seems to shine through. The program knew nothing of English, let alone the human condition. All it did was analyze the tendencies of one letter to follow another and produce random text that followed those tendencies.[1]

Now try these:

> *Away to the little for your knees!*
> *Oh, gathe little to the new angels sight*
> *Round gather and gathey shepherds quakes.*
>
> *Hark! the feast who,*
> *lowinter Proclaid dream*
> *from her and every worth.*
> *A ther King Glory hear thy dearth.*
>
> *Mary wondering love's because of old:*
> *Peace. Sleep on the Saviour knees!*
> *O night nightly night, his the the sing heaven.*
> *Bright, O holy shephen on Mary world Jesus,*
> *holy her and child in and nature see the star,*
> *He ago in their watch are shepher King;*
> *Let evenly peat ther King willness loods,*
> *whild, angels sing to the hay.*

Definitely not Shakespeare, but the writer is the same: Hamlet.pl, but using a different input text, composed of a variety of Christmas carols.

Hamlet.pl begins by performing a Markov analysis of the input text, creating a table of tendencies. Most often, we use that table to ***identify*** unknown input. For example, if you had access to the two tables underlying the first set of texts and the second, but you did not have access to either bona fide Hamlet or Christmas carols and had never heard either, you would no doubt be able to identify whether a text given to you was closer to Hamlet's speech or the conventions of Christmas carols.

Next time, we'll unpack Hamlet.pl and see how it works and see if we can modify it to determine whether a gene is closer in its structure to a native gene of an organism or one foreign to it.

[1] The program is freely available and might be useful for those writing theses at 3 am.