

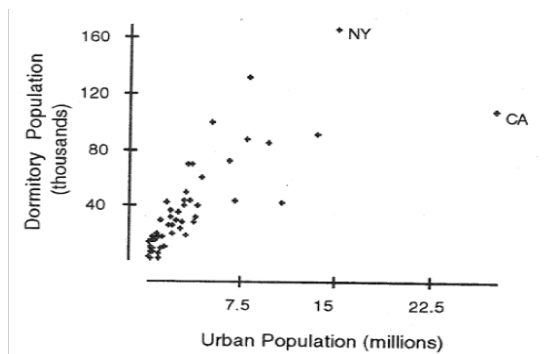
In the Long Run, We Will All Be Bayesians

George W Cobb

QME Seminar

GCobb@MtHolyoke.edu

1. Look, Ma: No Wars!
2. Gresham's Law of Learning?
3. The Tyranny of the Computable
4. Subversive Silicon: Three transitions
5. The Devil is in the Denominator
6. Whose Probability Is It, Anyway?
7. Laplace's Data Duplication Principle
8. A Russian Roulette Algorithm
9. Bayes Without His Theorem



BALLOON RULE FOR ESTIMATING A CORRELATION	
1. Balloon:	Draw a symmetric, oval balloon that best summarizes the point cloud.
2. Box of tangents:	Form a box by drawing the vertical and horizontal tangents
3. Outer:	Measure the vertical distance between the points of horizontal tangency
4. Inner:	Measure the vertical distance between the points of vertical tangency
5. Ratio:	correlation = Inner / Outer

11.1. An alignment problem from molecular biology.

Display 1 is a copy of a well-known data set from molecular genetics. The 18 rows correspond to DNA segments taken from 18 genes of the bacterium *E. coli*. Each segment is represented by a string of 105 letters chosen from a 4-letter alphabet $\{a, t, g, c\}$. These letters stand for the amino acids alanine, thymine, guanine, and cytosine, which pair with complementary amino acids to form the "rungs" (called base pairs) on the ladder-like structure of a DNA molecule. The genetic information in DNA is coded in the sequence of these base pairs.

Each of the segments shown in Display 1 contains, somewhere within its 105 base pairs, a sub-sequence of length 20 that functions as a binding site for a particular protein whose presence is required in order to "switch on" certain genes. Each sub-sequence can start at any position from 1 through 86. Although the 20 sub-sequences aren't exactly identical, nevertheless, when properly aligned, they will be the same in most positions, because they all have to bind with the same protein. The statistical challenge is (1) to find the starting positions of the binding sites, and (2) to find the sequence that best represents the "consensus" sub-sequence.

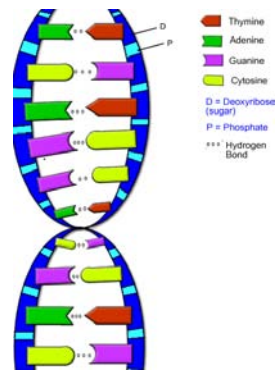
```

colel      taatgtttgtgctggtttttgtggcatcggggcgagaatagcgcgctgggtgtgaaagactgtttttttgatcgttttcacaaaaatggaagtccacagtccttgacag
ecoarabop  gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacggcgtcacacttggctatgccatagcatttttatccataag
ecobglrl  acaaatcccaataacttaattattgggatttggttatatataactttataaattcctaaaattacacaaaagttaataactgtgagcatgggtcatatttttatcaat
ecocrp    cacaaaagcgaaagctatgctaaaaacagtcaggatgctacagtaatacattgatgtactgcatgtatgcaaaggacgtcacattaccgtgcagtacagttgatagc
ecocya    acggtgctacacttgtatgtagcgcacttttctttacgggtcaatcagcatgggtgttaaattgatcacgtttttagaccattttttcgtcgtgaaactaaaaaac
ecodecop  agtgaattatttgaaccagatcgcattacagtgatgcaaacttgttaagtagatttcttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ecogale   gcgcataaaaaacggctaaattcttgtgtaaacgattccactaatttattccatgtcacacttttcgcatctttgttatgctatgggtattttcataaccataagcc
ecoilvbr  gctccggcgggggtttttgttatctgcaattcagtacaaaacgtgatcaaccctcaattttcccttggctgaaaaattttccattgtctcccctgtaaagctgt
ecolac    aacgcaattaatgtgagttagctcactcatttaggcaccccaggctttacactttatgtctccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ecomale   acattaccgccaattctgtaacagagatcacacaaaagcgacgggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtcggttta
ecomalk   ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgagggtcatgtaaggaatttctgtgatgttgcctgcaaaaaatcgtggcgattttatgtgcgca
ecomalt   gatcagcgtcgttttaggtgagttgttaataaagatttgggaattgtgacacagtgcaaatcagacacataaaaaacgctcatcgcttgcattagaaaggtttct
ecoompa   gctgacaaaaaagataaacataccttatacaagactttttttcatatgcctgacggagttcacacttgtaagttttcaactacgcttgtagactttacatcgcc
ecotnaa   tttttaaacattaaaaattcttacgtaattttataatctttaaaaaaagcatttaatatgtctcccgaacgattgtgattcgcattcaatttaacaatttcaga
eouxul   cccatgagagtgaattgttgtgatgtggttaacccaattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
pbr-p4   ctggcttaactatgcggcatcagagcagattgtactgagagtgaccatagcgggtgtgaaataccgcacagatgcgtaaggagaaaaataccgcatcagggcgtc
trn9cat  ctgtgacggaagatcacttcgcagaataaataaactcctgggtgtcctgttgataccgggaagcctgggccaacttttggcgaaaaatgagcgttgatcggcacg
(tdc)    gatttttatactttaacttggttgatatttaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttgtgagtggtcgcacatatcctggt

```

Display 1. Sequences of base pairs in DNA segments from 18 genes of E. coli.

Data from Stormo, G. D. and G. W. Hartzell (1989). "Identifying protein binding sites from unaligned DNA fragments." *Proceedings of the National Academy of Sciences*, v. 86, pp. 1183-1187.



For this particular problem, the correct alignment has been determined experimentally. The solution is shown in Display 2. Notice how, even with the sub-sequences lined up in parallel, it is not obvious that the sub-sequences match. In principle, each sub-sequence is a copy of the same underlying set of 20 letters, but in real life apparently random deletions, insertions, and substitutions alter the "true" sequence. The many mismatches illustrate part of what makes statistical work in molecular genetics so hard.

cole1	64	tttgatcgttttcacaaaa
ecoarabop	58	tttgcacggcgtcacactt
ecobglrl	79	tgtgagcatggtcatat
ecocrp	66	tgcaaaggacgtcacatta
ecocya	53	tgttaaattgatcacgttt
ecodecop	10	tttgaaccagatcgatta
ecogale	45	tttattccatgtcacactt
ecoilvbpr	42	cgtgatcaaccctcaatt
ecolac	12	tgtgagttagctcactcat
ecomale	17	tgtaacagagatcacacaa
ecomalk	64	cgtgatgttgcttgcaaaa
ecomalt	44	tgtgacacagtgcaaattc
ecoompa	51	cctgacggagttcacactt
ecotnaa	74	tgtgattcgattcacattt
eouxul	20	tgtgatgtggtaacccaa
pbr-p4	56	tgtgaaataccgcacagat
trn9cat (tdc)	81	tgtgagtggctgcacatat

Display 2. Starting positions and base pair sequences of binding sites for the 18 E. coli genes

Drill Exercise:

11.1 Here are four lines of letters:

```
qnerbmtkmfketltmgxpizachd
otspinvchteoisjmdjubpdkq
cyztnvpdspivachsvopddzaww
ozikozedbspinechurvbsobkv
```

Most of the letters in each line are random "junk" letters, but within each line there is one approximate copy of the same one actual word. (a) What is the word? What are the starting positions? (b) What is the dimension of θ ? of α ? What is the number of possible values for (θ, α) ?

11.2. Overview of Bayesian Analysis using Gibbs Sampling

In attacking the problem, our strategy will be to make several simplifying assumptions in order to create versions of the problem that we can actually solve. There will be two main assumptions, leading to two complementary versions of the problem:

If we assume that the alignment vector α is known, then "all" we need to find is the true coding sequence θ .

If we assume that true coding sequence θ is known, then all we need to find is the alignment vector α of starting positions.

Fortunately, these two assumptions lead to solvable problems. Finding either α or θ , given the value of the other, is quite possible. The difficulty of the challenge comes from needing to find *both* α and θ simultaneously. Once we've solved the two simpler one-at-a-time problems, however, we get an unexpected bonus: we can use the two solutions to create an alternating, back-and-forth solution to the harder problem:

(A) Assume that the alignment vector α is known. Find the posterior distribution (given α and the data \mathbf{Y}) for the true coding sequence θ , and use it to generate a random "true" coding sequence θ .

(B) Now assume that the true coding sequence θ is known. Find the posterior distribution (given θ and the data \mathbf{Y}) for the alignment vector α , and use it to generate a random alignment vector α . Then go back to (a), and use this random value for α .

By cycling back and forth between (A) and (B), you create a sequence of (θ, α) pairs:

$$\theta_0 \rightarrow \alpha_0 \rightarrow \theta_1 \rightarrow \alpha_1 \rightarrow \theta_2 \rightarrow \alpha_2 \rightarrow \dots$$

Notice that if you think of these in pairs (θ, α) , *the sequence obeys the Markov property*: the probabilities for $(\theta, \alpha)_{n+1}$ depend only on $(\theta, \alpha)_n$. This means that we can think of the process defined by (a) and (b) above as governed by a (very large!) transition matrix. The matrix may be humongous, but it is a transition matrix nonetheless. It has a stationary distribution, and if the chain is regular, it will converge to that stationary distribution as its limiting distribution. That distribution is in fact the joint posterior distribution of (θ, α) that we are looking for!

* * * * *

