

Contents

1	Probability theory	1
	Biological motivation: motif finding	1
	<i>E. coli</i> ribosome binding sites	1
	More biological motivation: score matrices for pairwise alignment	2
	Making customized score matrices	2
	Basics of probability	5
	Joint probability and independence	5
	Some notational issues, before we proceed	6
	Conditional probabilities	7
	Marginal probabilities	7
	Bayes' theorem	7
	fragmentary outline...	8
	Probabilistic models	9
	An ungapped matrix model of the <i>E. coli</i> Shine/Dalgarno motif	9
	Parameter estimation	13
	Maximum likelihood (ML) parameter estimation	13
	Laplace's law of succession	13
	Pseudocounts	14
	The rest of a fragmentary outline on parameter estimation	14
	ML estimation revisited more formally?; proof of ML	15
	MAP estimation and Dirichlet priors	15
	Expectation maximization	16
	Statistical inference	17
	Bayesian inference	17
	more fragmentary notes on Bayesian statistical inference	17
	Maximum likelihood inference	18
	Frequentist inference	18
	Probabilistic models of ungapped pairwise alignments	18
	Information theory	18
	Additional reading	19

Chapter 1

Probability theory

Biological motivation: motif finding

E. coli ribosome binding sites

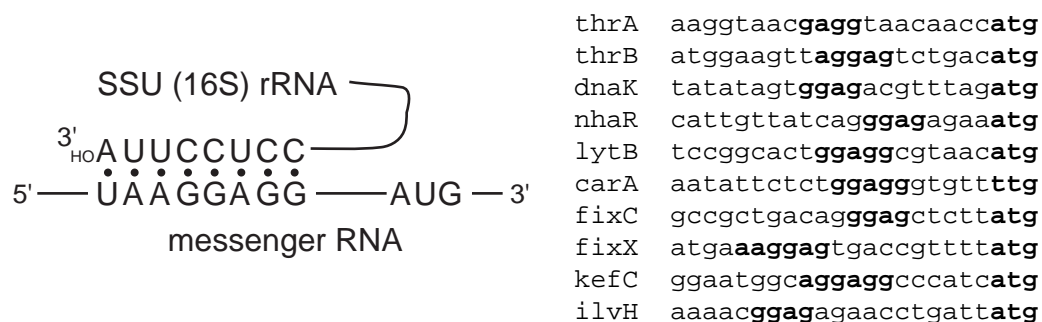


Figure 1.1: *E. coli* ribosome binding sites. Left: During translational initiation, a *E. coli* ribosome binding site (RBS) is recognized by base pairing of the 3' end of SSU (16S) ribosomal RNA to a short sequence just upstream of the correct initiation codon. The TAAGGAGG consensus sequence for the *E. coli* RBS is called the *Shine/Dalgarno* (SD) sequence [10]. Right: 20 nucleotides immediately upstream of the annotated initiation codon for ten *E. coli* genes with strong SD sequences. The initiation codon and SD sequences are highlighted in bold. The SD sequences only match part of the consensus, and they occur with variable spacing with respect to the initiator. (Note that one of the initiation codons is a TTG. *E. coli* does not always use ATG.)

More biological motivation: score matrices for pairwise alignment

When we discussed alignment algorithms, we used a *score matrix* $\sigma(a, b)$ that told us the score for aligning residue a to residue b . We needed these scores $\sigma(a, b)$ to obey two conditions, so that Smith/Waterman and other local alignment algorithms would work sensibly. First, at least one of the scores must be positive (favorable). Second, the overall expected score by chance, $\sum_{ab} f_a f_b \sigma(a, b)$, must be negative (unfavorable) - where f_a and f_b are the frequencies that residues a and b occur in sequences.

Then, in discussing BLAST and friends, we saw that there is a simple equation for calculating $\sigma(a, b)$ as a log-odds score:

$$\sigma(a, b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b} \quad (1.1)$$

We're going to use the $\sigma(a, b)$ equation as an excuse for introducing a whole chunk of relevant probability theory that you will need repeatedly in computational biology. But like many things theoretical, you don't have to understand it fully to use it effectively. First let's make sure you know how to use it to calculate your own customized substitution matrices - then we'll start delving into the theory behind them.

Making customized score matrices

To make our own score matrix, we just need to specify our own numbers for p_{ab} , f_a , and f_b and plug them into the log odds equation.

Let's deal with f_a and f_b first. f_a and f_b are simply the observed independent frequencies of residues a and b - that is, how often we expect to find these residues just by chance. We will typically get these numbers from some notion of the background frequency of residues in our target database.

For a DNA substitution matrix, a reasonable starting assumption is that A,C,G,T occur equiprobably, e.g. $f_x = 0.25$ for all four residues x . This is often not true in any single particular genome because genomic G+C composition varies substantially between organisms. But overall, if we use no auxiliary information about where the two DNA sequences being compared came from, our best guess at DNA base composition is close to equiprobability. Table 1.1 shows nucleotide frequencies in the genome sequences of three important organisms, *E. coli*, yeast, and human. *E. coli* does have a roughly equiprobable base composition, whereas yeast and human are somewhat AT-rich. (And there are also GC-rich genomes, of course.)

For an amino acid substitution matrix, we normally wouldn't assume that amino acids occurred equiprobably. Amino acid composition is biased in similar ways in all organisms - leucine (L), for example, is much more commonly used than tryptophan (W). We might obtain f_x from amino acid occurrence frequencies in some representative sequence database. Table 1.2 shows the residue frequencies observed in SWISS-PROT release 38.

What about p_{ab} ? This is the probability that we see an alignment of residue a and residue b in pairwise alignments of homologous sequences. We estimate these prob-

MORE BIOLOGICAL MOTIVATION: SCORE MATRICES FOR PAIRWISE ALIGNMENT3

Symbol	Nucleotide base	Residue frequencies f_x in:		
		bacterium <i>E. coli</i>	yeast <i>S. cerevisiae</i>	human <i>H. sapiens</i>
A	adenine	.246	.309	.295
C	cytosine	.254	.191	.205
G	guanine	.254	.191	.205
T	thymine	.254	.309	.295

Table 1.1: Nucleotide frequencies in three genome sequences.

abilities by counting aligned residue pairs in a database of trusted alignments. Some interesting issues arise. First, where do we get our trusted alignments from? (Since we're trying to estimate scoring parameters for alignment algorithms, the problem is a bit circular). Second, clearly p_{ab} is a number that is dependent upon a factor we haven't mentioned yet - the evolutionary distance between the pair of sequences. For two sequences that diverged recently, p_{ab} will be close to 1 for identical residues and close to 0 for nonidentical residues. For two sequences that diverged billions of years ago, p_{ab} will asymptotically approach $f_a f_b$, the probability of seeing these two residues independent of their evolutionary relationship. In between these two extremes is the realm in which we're interested, where subtle evolutionary relationships may be detected because p_{ab} is significantly different from $f_a f_b$.

There are three main strategies for estimating p_{ab} :

Algebraic Especially for DNA sequences, we can specify p_{ab} by making a few simple assumptions; for instance, if we assume that we're looking for 80% identical sequences, we might set p_{ab} to .2 for the 4 identical residue pairs (.8 / 4) and 0.0166 for the 12 nonidentical pairs (.2 / 12). (A slightly more sophisticated model would take into account that *transitions* ($A \leftrightarrow G, C \leftrightarrow T$) are more probable than *transversions* ($A \leftrightarrow C, A \leftrightarrow T, G \leftrightarrow T, G \leftrightarrow C$)).

Extrapolation Some of the first amino acid score matrices were derived by Dayhoff [2]. She used closely related protein sequences that could be confidently aligned, collected statistics on amino acid substitutions, and built a frequency table called PAM1. PAM stands for "point accepted mutation", reflecting the fact that an observed substitution is the result of mutation followed by selection. PAM1 contained residue substitution frequencies, normalized for pairs of sequences at 1% overall divergence. She could then estimate substitution matrices for longer evolutionary times just by matrix multiplication (PAM2 is PAM1²; PAM120 is PAM1¹²⁰; etc.) One subtle drawback to the PAM matrices is that amino acid substitutions at very close distances are significantly affected by the genetic code. Amino acid substitutions that require more than one base change in the codon are underrepresented. Over the longer time periods that are of more interest for remote homology detection, as base changes accumulate and equilibrate these code effects disappear, and p_{ab} is dom-





















Symbol	Amino acid	Observed counts	Frequency, f_x	
A	alanine	2205330	0.0758	
C	cysteine	483328	0.0166	
D	aspartic acid	1534944	0.0528	
E	glutamic acid	1851790	0.0637	
F	phenylalanine	1193226	0.0410	
G	glycine	1990933	0.0684	
H	histidine	654243	0.0225	
I	isoleucine	1690020	0.0581	
K	lysine	1729564	0.0595	
L	leucine	2745354	0.0944	
M	methionine	691771	0.0238	
N	asparagine	1292790	0.0444	
P	proline	1431286	0.0492	
Q	glutamine	1155060	0.0397	
R	arginine	1501462	0.0516	
S	serine	2074755	0.0713	
T	threonine	1650899	0.0568	
V	valine	1914781	0.0658	
W	tryptophan	360745	0.0124	
Y	tyrosine	928030	0.0319	
total:		29085965		

Table 1.2: Amino acid frequencies in SWISS-PROT 38.

inated by the propensity for residues to functionally substitute for each other in protein structures.

Empirical Currently, the best score matrices are probably the BLOSUM matrices [8]. BLOSUM62 is the default matrix for protein searches with BLAST. The starting point for estimation of the BLOSUM matrices was a large collection of automatically generated and reasonably trustworthy ungapped multiple sequence alignments, the BLOCKS database [7]. Observed residue substitutions were collected from all the pairwise relationships in the BLOCKS multiple alignments, using a weighting scheme that downweights the contribution of closely related sequences. (BLOSUM62 downweights the counts collected from pairwise alignments over 62% identity; BLOSUM40 downweights counts from alignments over 40% identity. Therefore, the lower the BLOSUM number, the more distant the sequence relationships; the opposite of PAM matrix numbers.) BLOSUM is believed to outperform PAM in homology searches because its p_{ab} 's are estimated directly from distantly related

proteins.

Even this brief discussion should raise a number of questions in your mind. Surely we can outperform the simple assumptions of these models, which are our current state of the art in bioinformatics! And indeed we can - but we will want to be fluent in the tools of probability theory.

Basics of probability

We all have at least an intuitive idea of what a probability is. For instance, for some event with N possible outcomes, the probability of obtaining outcome i is p_i . These probabilities p_i have the properties:

$$\sum_{i=1}^N p_i = 1.0; p_i \geq 0$$

Common examples of *discrete* probabilities include the probability of flipping heads or tails with a coin, rolling 1..6 on a die, observing A/C/G/T at some position in a DNA sequence, or observing one of the twenty amino acids at some position in a protein sequence.

“Uniform” probabilities (or “fair”, in the case of coins and dice) mean equiprobable outcomes, $p_i = \frac{1}{N}$. The probability of rolling a 6 with a fair die is $\frac{1}{6}$. The probability of some base in a DNA sequence being an A is $\frac{1}{4}$ (until we start worrying about nonuniform base composition in a genome sequence).

In many cases we will be able to estimate and interpret probabilities as frequencies; that is, from counting the number c_i of outcomes i in T repeated trials. As T gets large, the frequency of seeing i will approach p_i :

$$p_i = \lim_{T \rightarrow \infty} \frac{c_i}{T}$$

Joint probability and independence

The *joint probability* of a co-occurrence of event A and event B is written as $P(A, B)$ or just $P(AB)$. If the events are both rolls of a fair die, for example, $P(6, 6)$ is $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$. Multiplying the probabilities of individual events together to obtain the joint probability assumes that the events are *independent*. Independence is a reasonable assumption for simple repeated trials like die rolls and coin flips, but more often, we will not get off so easily.

We may explicitly assume independence for the sake of simplifying a calculation, even if it is only crudely true. For example, a common model of “random” (nonhomologous) sequences is the *i.i.d.* (*independent, identically distributed*) model. Given a protein sequence \mathbf{x} of length L , with residues numbered $x_1..x_i..x_L$, what is the probability $P(\mathbf{x})$ of that sequence? We need to specify a discrete probability distribution over the 20^L possible amino acid sequences of length L , which would require us to know

Bayesians and frequentists. What about events that only happen once and have a single outcome – like whether or not it will rain this afternoon? It doesn't make sense to talk about the frequency of it raining this afternoon; it's either going to rain or it's not. But we still talk about the probability of rain, even though we don't think of that probability in a frequentist sense. Thus, clearly we also sometimes interpret a probability as a degree of belief – or, more concretely, as betting odds. If my general mechanism of estimating probabilities is correct, over the long haul, the number of times I am right will approach the number of times I expected to be right. If my way of estimating odds is better than yours, I will make money from you in the long run.

And indeed, a consistent (Bayesian) theory of probability can be derived from a few first principles, the desiderata for an optimal statistical inference (e.g. betting) system. The frequentist view of probability falls out of the derivation as a special case. A particularly lucid exposition of this derivation and the Bayesian canon is in [9].

The “frequentist” (probability = frequency) and “Bayesian” (probability = degree of belief) views can be argued to great depths of subtlety and fundamentalism, but both views have merit. I will take a pragmatic approach and use whichever view seems appropriate for the biological problem at hand – though I confess a fondness for the clarity of the Bayesian view.

20^L numbers. By assuming *independence*, we assume that $P(\mathbf{x}) = \prod_{i=1}^L P_i(x_i)$. Now we need to know “only” $20L$ numbers, a distribution of 20 probabilities $P_i(x_i)$ for each position i in the sequence. By additionally assuming that residue probabilities are *identically distributed* (independent of position in the sequence), we further reduce the number of parameters to just 20 residue probabilities $P(x)$ in the final i.i.d. model – the same numbers we saw as f_a and f_b in log-odds scores.

Some notational issues, before we proceed

Before we go further, let's clarify some notation. When I say $P(AB)$, the upper case letters A and B refer to *random variables*. When I say $P(ab)$ or p_{ab} , the lower case letters a and b refer to *specific outcomes*, e.g. values that A or B may adopt. A random variable A corresponds to a set of possible events a_1, a_2, \dots, a_n . For example, A might refer to a roll of a die, which has 6 possible outcomes. Or, if I say $P(AB)$ is the probability of a pair of aligned homologous amino acid residues, the random variable AB can assume 400 values over all possible amino acid pairs. ab , in contrast, is a *specific event*, where a and b correspond to particular choices of amino acids. p_a is the probability that A assumes the specific value a : $P(A = a) = p_a$, and $\sum_A p_a = 1$.

Thus I talk about a *probability distribution* $P(A)$ for a random variable A , when I can specify the individual probabilities p_a for all the kinds of events that A represents.

The key thing to keep in mind is that p_a is a *single number* and a is a *specific outcome*, whereas A is a *set of outcomes* and $P(A)$ is a *set of numbers* (e.g. a probability distribution).

So far we're only talking about *discrete random variables* that can assume one of an enumerable set of outcomes (like the nucleotides ACGT, or an integer). Soon we'll also

see *continuous random variables*, where the outcomes are real numbers. When we do, we'll also see the word *distribution* show up in a very different context as a *probability distribution function*, as opposed to a *probability density function*, where distribution will have an almost entirely different meaning. We'll worry about that when we get there.

Conditional probabilities

The *conditional probability* $P(a|b)$ is the probability of event a given that we have observed an event b . A *conditional probability distribution* $P(A|B)$ is the probability distribution for random variable A , given that we have observed the outcome of random variable B .

For example, the probability $P(u|q)$ that the letter u follows the letter q in written English is close to 1.0, and quite different than $P(u)$ by itself.

In calculating score matrices, we are implicitly using a conditional probability $P(a, b|t)$, the probability of seeing residues a and b aligned, given an evolutionary divergence "time" of t .

There is an algebraic relationship between conditional and joint probabilities:

$$P(AB) = P(A|B)P(B)$$

and

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

Marginal probabilities

Given a joint probability distribution, we can always obtain the *marginal* probability distribution for any subset of random variables by summing over the others:

$$\begin{aligned} P(A) &= \sum_B P(AB) \\ &= \sum_B P(A|B)P(B) \end{aligned}$$

We can use *marginalization* to eliminate so-called *nuisance* random variables, ones for which we're not currently interested in their distribution.

Bayes' theorem

To get $P(A|B)$ from $P(B|A)$, we use *Bayes' theorem*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

which is equivalent to

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)}$$

Bayes' theorem is trivial, falling straight out of the algebraic relationships between conditional, joint, and marginal probabilities:

$$\begin{aligned} P(AB) &= P(A|B)P(B) \\ P(AB) &= P(B|A)P(A) \\ P(A|B)P(B) &= P(B|A)P(A) \\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ P(B) &= \sum_A P(B|A)P(A) \\ P(A|B) &= \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)} \end{aligned}$$

Bayes' theorem becomes more interesting when it is used as the heart of *Bayesian statistical inference*, where one of our random variables represents possible data sets we could observe, and the other represents possible hypotheses we want to consider to explain the data. As scientists we are interested in knowing $P(H|D)$, the probability of our hypotheses given our observed data, because we want to choose the most probable hypothesis. If we express our hypotheses as probabilistic models, we will generally be able to calculate $P(D|H)$, the probability of data given a hypothesis. We can use $P(D|H)$ to calculate $P(H|D)$ by applying Bayes' theorem. To do this, we need to specify $P(H)$ – the *a priori* probability of our hypotheses, *before the data arrive*. As scientists, we are quite used to doing this, at least intuitively. Occam's Razor tells us to favor simple hypotheses over complex ones. But can we actually assign objective numbers to $P(H)$? $P(H)$ is a difficult number to know, and you might even wonder if there isn't some circular logic lurking here – can we claim to be objectively calculating a data-dependent probability $P(H|D)$ for our hypotheses, if we had to explicitly assume a subjective prior distribution $P(H)$ over our hypotheses before we saw any data?

This is getting us into an area of deep controversy among statisticians. We'll defer further discussion of Bayesian inference 'til later in the chapter, when we take more time to contrast it against other ways of testing models, such as maximum likelihood inference and classical statistical inference (null hypothesis rejection testing). It's important to remember that Bayes' theorem itself is noncontroversial – indeed, it's true by definition.

fragmentary outline...

- contingency table ("truth table") illustration of how to think about joint, conditional, and marginal probabilities.

- Example: The occasionally dishonest casino. Two types of dice; 99% are fair, 1% are loaded such that $p_6 = 0.5$. Conditional $P(6 \text{ — loaded})$; joint $P(6, \text{loaded})$; marginal $P(6)$.
- Example: Alan Templeton's critique of "disease genes". Imagine a gene with two alleles A and a , and another gene with two alleles B and b . The a and b alleles interact to cause a genetic disease in ab individuals. The loci are unlinked, and the allele frequencies are $a=.01$, $A=.99$, $b=.9$, $B=.1$. The prevalence of the disease in the population is $P(ab)$, 0.001. If you have the a allele, the probability you have the disease is $P(ab \text{ — } a)$, 90%. If you have the b allele, the probability you have the disease is $P(ab \text{ — } b)$, 1%. a is a better predictor of disease status... but does that mean that a is a "disease" allele?
- Example: AUG initiator codon consensus. $P(1,2,3 \text{ — initiator}) = P(3 \text{ — } 1, 2, \text{ initiator}) P(1,2 \text{ — initiator})$, etc.
- Markov dependency: examples of probabilistic models of genomic DNA.

Probabilistic models

An ungapped matrix model of the *E. coli* Shine/Dalgarno motif

Let's return to our example of *E. coli* translational initiation sites. The SD sequence often does not exactly match the consensus, and it is located at different distances from the initiation codon. If someone gives us a sequence upstream of some *E. coli* coding region of interest, can we automatically predict the location and sequence of its SD?

To frame this as a probabilistic modeling problem, we first need to state $P(S|M)$: the probability distribution over sequences S , given a model M for what SD's look like. There are an infinite number of possible sequences, and we'd hate to have to specify an infinite number of parameters for our model M , so we'll make two simplifying assumptions. First, we'll assume that the SD can be modeled as an *ungapped, fixed length* motif of width W ; thus $P(S)$ is $P(x_1, x_2, \dots, x_W)$, the joint probability of the W contiguous bases in the motif. Second, we make an independence assumption: the probability of seeing a particular base at some position in the motif is independent of the specific bases at the other positions. Thus for each position i in the model, we need 4 probability parameters $p_i(a), p_i(c), p_i(g), \text{ and } p_i(t)$ specifying the probability of each base at that position (a total of $4W$ parameters). Then:

$$P(S|M) = P(x_1 \dots x_W|M) = \prod_{i=1}^W p_i(x_i)$$

This kind of model is the probabilistic version of a so-called *weight matrix* or *position-specific sequence motif* (PSSM), a commonly used model of ungapped sequence consensus in computational biology. The difference right now is that our numbers are products of probabilities rather than sums of scores. (Later, we'll start working with probability parameters as log probabilities or log-odds scores, at which point we'll

The prosecutor's fallacy.

The difference between a likelihood and a posterior probability can be illustrated by the prosecutor's fallacy.

A person stands accused of a murder. DNA evidence recovered from crime scene has been subjected to genetic fingerprinting, and is found to match the DNA fingerprint of the accused person. There is an extremely low probability that a randomly chosen person would have this fingerprint. The prosecutor stands before the jury and says, "The odds of obtaining such a good DNA fingerprint match by chance are one in a million. This probability is so low that the accused is clearly guilty beyond reasonable doubt."

We can see the fallacy here by expressing this in terms of probability theory. We have two random variables: V (the correct verdict) which can assume two values g and i representing guilt or innocence; and F (the DNA fingerprint) which represents one of many possible discrete fingerprints f . What we want to know is the probability $P(i|f)$, the probability that the accused is innocent given that we've seen the fingerprint data f . What the prosecutor is telling us is $P(f|i)$, the probability that we obtain fingerprint f from an innocent person. (We'll assume that $P(f|g)$ is 1: we assume that the fingerprint we obtained at the scene is from the murderer, not the result of a mistake at the lab or a deliberate mishandling of the evidence.) Bayes' rule tells us that:

$$\begin{aligned} P(i|f) &= \frac{P(f|i)P(i)}{P(f)} \\ &= \frac{P(f|i)P(i)}{P(f|g)P(g) + P(f|i)P(i)} \\ &= \frac{P(f|i)P(i)}{P(f|g)P(g) + P(f|i)(1 - P(g))} \end{aligned}$$

So to calculate the posterior probability $P(i|f)$, we need to know the prior probability $P(g)$ that a randomly chosen individual would be guilty (note that $P(i) = 1 - P(g)$). If we assume that the killer is somewhere in Los Angeles, with a population of about 10 million, the probability $P(g)$ that a randomly chosen Angeleno is the killer is $1/10^7$. Substituting the relevant numbers and solving gives us:

$$P(g|f) = \frac{10^{-6} * (1 - 10^{-7})}{1 * 10^{-7} + 10^{-6}(1 - 10^{-7})} = 0.91$$

So the probability the accused person is innocent is 91%, if this is the only evidence the prosecutor is presenting! Truly a far cry from the "one in a million" chance that the prosecutor is telling the jury.

This is intuitive if you think about it this way. If there's a one in a million chance of a spurious DNA fingerprint match, this means that a total of 10 innocent people out of the 10 million population of Los Angeles have the fingerprint, in addition to the 1 guilty person. If you show me an individual that has this particular fingerprint, all I know is that they're one of these 11 people. The odds that any given one of them is the guilty one are 1 in 11.

Another example of the prosecutor's fallacy arose in a debate in the pages of Nature about the probability that the pope is a human or an alien [3].

also be summing terms instead of taking products, and the parallel to weight matrices and PSSMs will be even more clear.)

Figure 1.2 shows an example of a parameterized matrix model of the *E. coli* RBS, using a width $W = 8$.

residue	position							
	1	2	3	4	5	6	7	8
A	0.342	0.350	0.407	0.644	0.075	0.234	0.614	0.274
C	0.159	0.116	0.310	0.111	0.082	0.037	0.038	0.041
G	0.165	0.171	0.234	0.079	0.805	0.721	0.097	0.542
T	0.334	0.363	0.049	0.166	0.038	0.008	0.252	0.142
consensus:	A/T	A/T	A	A	G	G	A	G

Figure 1.2: Ungapped probabilistic model of *E. coli* ribosome binding sites. (The exact numbers here should not be taken as biologically “real”. They’re just an example that comes from running an EM algorithm on a dataset of 823 *E. coli* translational initiation sites of length $L = 20$, using an ungapped matrix model of width $W = 8$. The EM algorithm is introduced later in the chapter.

If I showed you the 8-mer sequence ATAAGGAG (the best match), you would calculate that its probability $P(S|M)$ is 0.00629. The sequence CCTGTTCC (the worst match) has probability 3.38×10^{-11} under this model.

Now consider an upstream region of length L : for instance, the 23 nucleotides upstream of the *proA* gene, ACCCGTTAAGGAGCAGGCTGATG, inclusive of the initiator ATG. Where is the SD sequence?

Now we need a probabilistic model of the complete upstream sequence S . We assume that S contains *one and only one* SD motif. We also assume that the other residues (outside the SD) occur with position-independent background probabilities $f(a), f(c), f(g), f(t)$; and for further simplicity, let’s assume that the $L - W$ bases in the non-motif sequence occur equiprobably ($f(a) = f(c) = f(g) = f(t) = 0.25$). If we knew the SD motif started at position k (where $k = 1 \dots L - W + 1$), then:

$$\begin{aligned}
 P(S|k, M) &= \prod_{j=1}^{k-1} f(x_j) \prod_{i=1}^W p_i(x_{k+i-1}) \prod_{j=k+W}^L f(x_j) \\
 &= 0.25^{L-W} \prod_{i=1}^W p_i(x_{k+i-1})
 \end{aligned}$$

Our problem is that we don’t know the motif position k . That’s what we want to determine. Specifically, we want $P(k|S, M)$: the probability that the motif starts at position k , given the sequence and the model.

Applying Bayes’ theorem gives us:

$$P(k|S, M) = \frac{P(S|k, M)P(k|M)}{P(S|M)} = \frac{P(S|k, M)P(k|M)}{\sum_k P(S|k, M)P(k|M)}$$

So to get our *posterior* $P(k|S, M)$, we need not only the *likelihood* $P(S|k, M)$ but also a *prior* $P(k|M)$: the probability that the site is at position k *a priori*, that is, before the sequence data “arrive”. The simplest assumption we can make is that the priors $P(k|M)$ are a uniform distribution (all k are equiprobable *a priori*; then the $P(k|M)$ will cancel out and leave us with:

$$P(k|S, M) = \frac{P(S|k, M)}{\sum_k P(S|k, M)}$$

And because of our assumption of equiprobable background frequencies, every $P(S|k, M)$ term contains a constant factor of $0.25(L - W)$ which also cancels, leaving us with:

$$P(k|S, M) = \frac{\prod_{i=1}^W p_i(x_{k+i-1})}{\sum_{k'=1}^{L-W+1} \prod_{i=1}^W p_i(x_{k'+i-1})}$$

And that’s all we need to get an answer. The numbers calculated for the log likelihood $\log P(S|k, M)$ and the posterior $P(k|S, M)$ for each of the 16 positions in the proA sequence are shown in Figure 1.3 (23-8+1 = 16 positions at which the motif can occur; it can’t start in the last 7 positions because of the edge effect). According to our model, the proA SD is almost certainly at position 6 ($P = 0.972$), with the sequence TTAAGGAGG - as it happens, a very good consensus SD.

position k	motif at k	likelihood $\log P(S k, M)$	posterior $P(k S, M)$
1	accggtta	-35.106	0.000
2	ccggttaa	-38.367	0.000
3	ccgttaag	-33.853	0.000
4	cgttaagg	-36.200	0.000
5	gttaagga	-33.604	0.000
6	ttaaggag	-25.888	0.972
7	taaggagc	-33.577	0.000
8	aaggagca	-34.400	0.000
9	aggagcag	-30.140	0.014
10	ggagcagg	-34.696	0.000
11	gagcaggc	-35.743	0.000
12	agcaggct	-31.016	0.006
13	gcaggctg	-33.692	0.000
14	caggctga	-38.629	0.000
15	aggctgat	-33.308	0.001
16	ggctgatg	-30.993	0.006

Figure 1.3: Posterior probability calculations for the position of the SD in 23 nucleotides of the *E. coli* proA ribosome binding site, ACCCGTTAAGGAGCAGGCTGATG.

Some discussion of how the theory immediately allows us to see how to expand our power...

The decision to assume a fixed width W for the motif.

The decision to allow one and only one motif.

The decision to assume a uniform prior for k.

The decision to assume equiprobable base frequencies.

The decision to assume the motif contains no indels.

Computing with probabilities. ... why we work in log prob's, not prob's. ... the LogSum() trick for doing sums while still in log space. ... and the even fancier static-table version.

Parameter estimation

We are given some observed counts of events, such as rolls of a die, flips of a coin, or counts of A/C/G/T at some position in a DNA sequence. From these counts, we want to estimate the probabilities of these events. The problem of estimating probability parameters from observed count data is fundamental to computational biology.

Maximum likelihood (ML) parameter estimation

The simplest way to estimate parameters from count data is just

$$\hat{p}_x = f_x = \frac{c_x}{\sum_y c_y}.$$

That is, the frequency of an event is the *maximum likelihood (ML) estimate* of the probability of the event. ML estimation works well if the counts are a large, unbiased sample. ML estimation converges on the true probability as the size of the (unbiased) samples grow towards infinity.

For example, our estimates of f_a and f_b for background amino acid or nucleotide probabilities are exactly this: frequencies in a huge sample of millions (even billions) of observed residues.

Laplace's law of succession

ML estimation is often not sufficient, because we don't always have a large sample.

The most glaring problem occurs when we haven't seen an event at all yet, so that its observed count $c_x = 0$. Should we assume that $p_x = 0$? Probably not.

... Laplace plus-one here ... MP estimator under flat Dirichlet prior ...

Pseudocounts

And of course the problem of *small sample statistics* arises even for nonzero counts. For instance, with a fair coin, I might flip the coin 100 times and get 48 heads and 52 tails; so I observe a frequency of $f_H = .48$. Does this mean that the probability of flipping a head on the *next flip* is 0.48? Probably not; more likely, that probability, p_H , is 0.5, because we expect that the coin is fair. Since we've observed a frequency of 0.48 heads, intuitively we might easily conclude that that frequency's consistent with $p_H = 0.5$.

Indeed, if one were estimating parameters and betting based on our expectations given those parameters, a sensible person will generally win on average against an ML estimator. The person uses information that the ML estimator does not: that is, the fact that in general we expect a coin to be fair, though we would eventually allow the data to override that prior belief if it became clear that the coin was clearly flipping too many tails or heads.

... pseudocounts presented here as an ad hoc procedure with the right behavior ...

That is, we will often be interested in estimating unobserved parameters p_x of a probabilistic model, based on a finite number of observations that will often be in the form of counts, c_x . When we infer these p 's, we may want to also take into account *prior knowledge* about our problem - that is, knowledge we have before we look at the data - such as the idea that coins are usually fair. We will have to formalize this notion so that we know how to combine prior knowledge with the data. We don't yet have enough machinery in our repertoire to deal with this, so we'll delay further discussion 'til after we've seen more about Bayes' theorem, posterior probabilities, and prior probabilities.

The rest of a fragmentary outline on parameter estimation

- We've seen how to infer what model is correct, but where did the models come from? In particular, how do I infer the best parameters to use in a model?
- Example: a single coin and lots of flips. The intuition is to use observed frequencies as probability parameters. This intuition is largely correct; that's the maximum likelihood estimate.
- Parameter estimation viewed as yet another inference problem: $P(\theta|D) = P(D|\theta)P(\theta)/P(D)$.
- Sidebar: $P(\theta|D)$ is our first example of a continuous probability function. Definition of distributions and densities.
- Back to the coin example: plot $P(D|\theta)$. Pick the peak; "maximum likelihood" estimation.
- A problem with ML estimation: if we flip 49 heads and 51 tails, would you really bet that $P(\text{head}) = 0.49$? Or would you bet that $P(\text{head}) = 0.5$? 0.49 is the more "likely" choice based on the observed data, but you would probably lose a bet on this.

- Another (related) problem: what if I observe 0 of some event. Do I really infer $P(\text{event}) = 0$? Leads to rejection of hypothesis by a single rare observation.
- A second estimation approach: maximum a posteriori (MAP) estimation. Inclusion of prior information in parameter estimation. The Laplace "plus-one" prior. Pseudocounts and their intuitive properties. Necessity for formal prior distributions; concept of a "conjugate prior"...

ML estimation revisited more formally?; proof of ML

- Our goal: show the derivation of "optimal scores", and show that they have probabilistic meanings.
- Consider a set of counts (rolls of a die, or observed amino acids in an aligned column).
- The multinomial distribution as $P(D|M)$.
- Find parameters that maximize $P(D|M)$.
- maximizing $\log P(D|M)$ becomes convenient (additive). (Sidebar: View $\log p_x$ as "scores" to add up.)
- The derivation; differential calculus; constrained optimization and Lagrange multipliers – there's a reason why calculus is a prerequisite for comp bio...
- ta-da: the simple ML estimate $p_x = \frac{c_x}{\sum_y c_y}$.

MAP estimation and Dirichlet priors

- Remember our simple "Laplace plus-one" and "pseudocount" rules. Now let's justify them formally.
- Bayes: $P(M|D)$ proportional to $P(D|M)P(M)$; MAP estimate for p_x are p_x that maximize $P(c|p)P(p)$.
- Wait; what is $P(p)$? A probability distribution over probability distributions. Analogy to "the dice factory" (MacKay). What does such a distribution look like?
- Another factor in choosing this distribution: the notion of a "conjugate" prior. After the calculus above, it's easy to imagine that we want a prior that doesn't greatly complicate our differential calculus.
- The Dirichlet distribution defined.
- Important properties of the Dirichlet to fix one's intuition: mean; peaky behavior for coefficients < 1 ; "flat" if all coefficients = 1; variance inversely proportional to sum of coefficients.

- Maximization of $P(c|p)P(p)$: $p_x = c_x + a_x - 1$, normalized.
- OK, almost there. Now, mean posterior estimation; $p_x = c_x + a_x$.
- Now we can review Laplace plus-one and pseudocounts. Plus-one corresponds to a flat Dirichlet prior. Pseudocounts correspond to a Dirichlet, specifying a certain expected mean and variance.
- A slight variation: estimation using mixture Dirichlets. Illustration using dice factories and amino acid distributions.

Expectation maximization

We will often be faced by a problem that involves *missing data*. That is, my probabilistic model will be dependent on some factor that I don't know *a priori* and must infer. That is, I can't calculate $P(\text{data}|\text{model})$ directly; I need to calculate it by marginalizing away some hidden nuisance variable ν :

$$P(\text{data}|\text{model}) = \sum_{\nu} P(\text{data}|\nu, \text{model})P(\nu|\text{model})$$

For example, in the case of ribosome binding sites, my model for $P(S|M)$ was expressed in terms of a conditional probability $P(S|k, M)$, because the overall probability of an RBS was dependent on the position of the SD motif. Given known probability parameters for the model M , I could infer a posterior distribution for the correct position k in a RBS. Or, given known motif positions k in a set of example sequences, I could extract and align the motifs, count the occurrence of bases in each aligned column, and estimate the probability parameters of the model.

What if I don't know either the parameters or the positions of the motifs? That is, what if I'm faced with the following biological problem of general interest: I'm given a set of N unaligned sequences of length L . Each sequence contains a conserved motif of length W . Can I simultaneously infer both the positions of the motifs and a probabilistic model of the motif consensus, starting from nothing but the unaligned data?

A very powerful and simple algorithm for this problem is *expectation maximization* (EM). In outline, EM works as follows:

- Initialize the model M to random parameters.
- While not converged:
 - Estimate the posterior probability distribution for the hidden variable. (Assuming the parameters are correct, what is the motif position k ?)
 - Collect expected counts from the data, given the estimated posterior distribution.
 - Given those expected counts, estimate new maximum likelihood parameters for the model. (Assuming the motif positions k are estimated correctly, what are the best model parameters?)

This iterative algorithm is guaranteed to converge to a local (not global) optimum.

- another example: $P(\text{seq}|\text{model}) = \sum_{\text{alignments}} P(\text{seq}, \text{alignment}|\text{model})$
- Toy example: two dice.
- Expectation step: using current parameters, estimate expected counts.
- Maximization step: using current expected counts, re-estimate new parameters.
- simple ways to get around local optimality: running from multiple randomly chosen starting points.
- Related algorithms include "Gibbs sampling", "simulated annealing", "Markov chain Monte Carlo" (MCMC).

Statistical inference

... what I mean by "inference"...

Bayesian inference

Let's consider the interesting case where we have a random variable D that represents possible observed outcomes in a data, and a random variable M that represents a set of models (hypotheses) that we want to test.

In science, we're usually interested in distinguishing between alternative models and finding the most likely one, given some data. That is, we are interested in $P(M|D)$, the probability of two or more alternative models given the data.

If we're lucky, we will know enough about the hypotheses we're considering that we will be able to specify them as *probabilistic models*, which means that we can calculate $P(D|M)$ – the probability of obtaining particular observations if we assume that a particular hypothesis is true.

more fragmentary notes on Bayesian statistical inference

- Consider the AUG initiator codon problem. Suppose we are given a triplet, and asked to infer, is this an initiator or not?
- Another example: consider the occasionally dishonest casino. We choose a die, roll it three times, and every roll comes up a 6. Did we pick a loaded die?
- Both questions involve statistical inference. We want to know the probability of a *hypothesis*, H , given some *data*, D . Either hypothesis can explain the data, so we cannot know for sure, but one hypothesis may be more probable so that we can assign betting odds.
- Bayes' rule: $P(H|D) = P(D|H)P(H)/P(D)$
- $P(D)$ is a marginal probability; $P(D) = \sum_H P(D|H)P(H)$

- Algebraic derivation of Bayes.
- Definition of posterior probability $P(M|D)$, prior probability $P(M)$, and model likelihood $P(D|M)$.
- Worked example for the casino; 21% chance that the die is a loaded one.
- "Bayesian statistical inference"
- Subjectivity of priors; example: the use of Bayesian inference in the recovery of a lost American nuclear weapon off the coast of Spain in the 1960's.

Maximum likelihood inference

- Likelihood ratios; log likelihood ratios (LLR)
- Relationship of LLR and chi-squared test
- Relationship of LLR and Bayesian posterior
- Substitution matrices are log-likelihood ratios
- Derivation of BLOSUM matrices (Henikoff 1992 paper)
- Interpretation of arbitrary score matrices as probabilistic models (Altschul 1991 paper)

Frequentist inference

- What BLAST/FASTA P-values and E-values mean
- The meaning of $P(S \geq x)$ and why we want to know it.
- Expectation value, $E = NP(S \geq x)$
- BLAST's "P(n)"; Poisson assumption; $P(n) = 1 - \exp(-NP(S \geq x))$
- These are "frequentist" or sampling statistics
- "Significance" = rejection of null hypothesis; comment on logical fallacy of accepting a proposed hypothesis by rejecting a different one.

Probabilistic models of ungapped pairwise alignments

Information theory

(See also: "Information Theory Primer With an Appendix on Logarithms" by Thomas D. Schneider <http://www-lecb.ncifcrf.gov/toms/paper/index.html>.)

- "How conserved is this alignment column?" We need a justified measure for conservation.

- From Shannon, we have a definition of average uncertainty: the so-called "entropy" $H(X) = -\sum_x p_x \log p_x$
- Properties: maximal if p_x equiprobable; zero if one $p_x = 1.0$.
- Common to use \log_2 and express $H(X)$ in "bits": can be viewed as the number of yes/no questions necessary to resolve uncertainty.
- Connection to probabilistic modeling: recall $\log p_x$ viewed as a log likelihood "score" in the above section. Under such an additive scoring system, the expected (average) score is $\sum_x p_x \log p_x$ – the Shannon entropy.
- The meaning of "information": reduction in uncertainty. Viewed as the difference in uncertainty before and after an "informative" event. $I = H_{before} - H_{after}$.
- Example: entropy of random DNA
- Example: information content of an aligned column. Information content as a legitimate measure of conservation.
- Relative entropy (Kullback-Leibler "distance"). $H(P||Q) = \sum_x p_x \log(p_x/q_x)$. Identical to information content if Q is a uniform background distribution; two terms often used interchangeably.
- Connection to probabilistic modeling: relative entropy = log odds score, and log odds score can be viewed as a rearrangement of a Bayesian posterior.
- Mutual information $M(XY) = \sum_x y p_{xy} \log(p_{xy}/p_x p_y)$. Example: RNA sequence alignments.

Additional reading

Perhaps the best general reference on probability theory is Feller's *Introduction to Probability Theory and its Applications* [5, 6]. Cover and Thomas's *Elements of Information Theory* is my favored reference for Shannon information theory [1]. Anthony Edwards' *Likelihood* is an unusually lucid (and dogmatic) book on maximum likelihood inference [4]; it also particularly accessible to biologists, since Edwards is a geneticist.

A wonderfully eccentric book on Bayesian inference is the unpublished and unfinished *Probability Theory: The Logic of Science*, by Ed Jaynes [9]. Jaynes was a professor emeritus in the physics department here at Washington University in St. Louis until his recent death. I treasure a collection of some of his papers and manuscripts that came into my possession.

Bibliography

- [1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
- [2] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, pages 345–352. National Biomedical Research Foundation, Washington DC, 1978.
- [3] S. R. Eddy and D. J. C. MacKay. Is the Pope the Pope? *Nature*, 382:490, 1996.
- [4] A. W. F. Edwards. *Likelihood*. Johns Hopkins University Press, Baltimore, 1992.
- [5] William Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*. John Wiley & Sons, 1968.
- [6] William Feller. *An Introduction to Probability Theory and Its Applications, Volume 2*. John Wiley & Sons, 1971.
- [7] Steven Henikoff and Jorja G. Henikoff. Automated assembly of protein blocks for database searching. *Nucl. Acids Res.*, 19:6565–6572, 1991.
- [8] Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919, 1992.
- [9] E. T. Jaynes. *Probability Theory: The Logic of Science*. Available from <http://bayes.wustl.edu>., 1998.
- [10] J. Shine and L. Dalgarno. Determinant of cistron specificity in bacterial ribosomes. *Nature*, 254:34–38, 1975.