BE.011/2.772J
Statistical Thermodynamics of Biomolecular Systems
Spring 2004
Griffith/Hamad-Schifferli

Problem Set #1 Solutions
Due: 2/11/04

1.) Dill 1.1

(a) The simplest way to solve this problem is to recall that when probabilities are *independent*, and you want the probability of events A AND B, you can *multiply* them. When events are *mutually exclusive* and you want the probability of events A OR B, you can *add* the probabilities. Therefore we try to structure the problem into an AND and OR problem. We want the probability of getting into J or DSM or UCSF. But this doesn't help because these events are not mutually exclusive (mutually exclusive means that if one happens, the other cannot happen). So we try again. The probability of acceptance somewhere, P(a), is P(a) = 1- P(r), where P(r) is the probability that you're rejected everywhere. (You're either accepted somewhere or you're not.) But this probability can be put in the above terms. P(r) = the probability that you're rejected at J AND at DSM AND at UCSF. These events are independent, so we have the answer. The probability of rejection at H is p(rH) = 1-0.5 = 0.5. Rejection at DSM is p(rDSM) = 1-0.3 = 0.7. Rejection at UCSF is p(rUCSF) = 1-0.1 = 0.9. Therefore P(r) = (0.5)(0.7)(0.9) = 0.315. Therefore the probability of at least one acceptance = **P(a) = 1-P(r) = 0.685**.

(b) This is the intersection of two independent events:

p(aH)p(aDSM) = (0.5)(0.3) = **0.15**.

2.) Dill 1.2

(a) Each base occurs with probability ¼. The probability of an A in position 1 is ¼, of a in position 2 is ¼, of A in position 3 is ¼, of T in position 4 is ¼ and so on. There are 9 bases. The probability of this specific sequences is $\mathbf{(1/4)^9 = 3.8 \times 10^{-6}}$**.**

(b) Same answer as (a) above.

(c) Each specific sequence has the probability given above, but in this case there are many possible sequences which satisfy the requirement that we have 4 A's, 2 T's, 2 G's, and 1 C. How many are there? We start as we have done before, by assuming all nine objects are distinguishable. There are 9! Arrangements of nine distinguishable objects in a linear sequence. (The first one can be in any of nine places, the second in any of the remaining eight places, and so on.) But we can't distinguish the four A's's, so we have over-counted by a factor of 4!, and must divide this out. We can't distinguish the two T's, so we have over-counted by 2! , and must also divide this out. And so on. So there are $\left[\frac{9!}{4!2!2!1!}\right]$ sequences having this composition. In parts a and b we found that each such sequence occurs with probability 0.25^9. So the overall probability of 4 A's, 2 T's, 2 G's, and 1 C is $3780*(0.25)^9 =$ **0.0144**

3.) Dill 1.6

(a) We want to find *a* such that $\int_0^1 p(x)dx = 1$. So we can plug in for *p(x)* and integrate:

$$\int_0^1 p(x)dx = \int_0^1 ax^n dx = \frac{a}{n+1}x^{n+1} \Big|_0^1 = \frac{a}{n+1} = 1$$

So, for normalization, we require that *a* = *n* + 1.

(b) To find the mean <x> we use the standard formula, and plug in for *p(x)*.

$$\langle x \rangle = \int_0^1 xp(x)dx = \int_0^1 ax^{n+1}dx = \frac{a}{n+2}x^{n+2} \Big|_0^1 = \frac{a}{n+2} = \frac{n+1}{n+2}$$

(c) To find the variance, again use the formula, and plug in for *p(x)*.

$$\sigma^2 = \int_0^1 x^2 p(x)dx - < x >^2$$

$$= \int_0^1 ax^{n+2}dx - \left[\frac{n+1}{n+2}\right]^2 = \frac{a}{n+3}x^{n+3} \Big|_0^1 - \left[\frac{n+1}{n+2}\right]^2$$

$$= \frac{n+1}{n+3} - \left[\frac{n+1}{n+2}\right]^2 = \frac{n+1}{(n+3)(n+2)^2}$$

4.) Dill 1.10

(a) You have 20 distinguishable amino acids, and each sequence is *n* letters long. You have one test sequence and *s* sequences in a database that you are testing against. The chance of matching is the chance of seeing a specific amino acid from your test sequence, so this part is like problem 2a since each position is independent. The probability of a perfect match of all *n* residues is p^n. To find how many perfect matches there will be, just multiply by the number of sequences you are trying to match to, *s*. (Since real numbers are easier, if the probability of a perfect match was 0.25, of 100 sequences you would expect to see 25 matches.) So, the average number of matches in *s* sequences = sp^n.

(b) When there is one mismatch in a sequence, this means *n* – 1 positions match. You need to know how many ways you can choose these (n-1) matches out of n amino acids. And for the probabilities, you have a probability p for each of the (n-1) matches and a probability (1-p) for the one mismatch. The answer is given by the binomial distribution!

p(one mismatch in a single comparison) = $\binom{n}{n-1} p^{n-1}(1-p)^1 = np^{n-1}(1-p)$

To get the average number of single mismatches, again multiply by *s* as in part (a). Average number of single mismatches in *s* comparisons = **snp^{n-1}*(1-p)***.

Note, in general, for k matches:

$P(k) = sp^k(1\text{-}p)^{n\text{-}k} \dfrac{n!}{k!(n-k)!}$

5.) Dill 1.11
First, this problem assumes that all cysteines will be paired. In a real protein, although an unpaired cysteine is less favorable than a paired cysteine, your protein can still fold and function nicely even if there are a couple unpaired cysteines in it. But, in this problem, assume all cysteines must be paired.

There are two ways of doing this problem. The first way: number the individual sulfhydryl groups along the chain. The first sulfhydryl along the sequence can bond to any of the other n-1. This removes two sulfhydryls from consideration. The third sulfhydryl can then bond to any of the remaining n-3. Four sulfhydryls are now removed from consideration. The fifth can now bond to any of the remaining n-5 sulfhydryls, etc., until all n/2 bonds are formed. Thus the total possible number of arrangements of disulfide bonds is a product of n/2 terms:
= (n-1)(n-3)(n-3)…1. So in this case of 6 CYS residues, the answer is 5*3*1 = **15**.

Another way: we write the names of all the cysteines on separate pieces of paper and put them in a bag. We generate different possible pairings by drawing two pieces of paper from the bag at a time until all of the papers have been drawn. (so, “n choose 2”, then “n-2 choose 2”, etc.) The number of ways we can do this is:

$$\binom{n}{2}\binom{n-2}{2}\binom{n-4}{2}\cdots\binom{4}{2}\binom{2}{2}$$

where

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

expresses the number of possible ways we can choose k items from a pool of n distinguishable items without regard to order. We can simplify this expression.

$$\binom{n}{2}\cdots\binom{2}{2}=\frac{n!}{2!(n-2)!}\frac{(n-2)!}{2!(n-4)!}\cdots\frac{6!}{2!4!}\frac{4!}{2!2!}\frac{2!}{2!0!}=\frac{n!}{2!}\frac{1}{2!}\cdots\frac{1}{2!}\frac{1}{2!2!}=\frac{n!}{2^{n/2}}$$

This is not our final answer! This expression actually overcounts the number of arrangements because pulling out Cys7-Cys26 as the first pair, for instance, counts separately from pulling it out as the tenth pair. We have overcounted by a factor of (n/2)! which is the number of possible orderings of n/2 pairs. Therefore, the number of arrangements is equal to:

$$\frac{n!}{2^{n/2}}/\left(\frac{n}{2}\right)!=\frac{n!}{2^{n/2}\left(\frac{n}{2}\right)!}$$

Although these two equations were derived in very different ways, they are numerically identical for all n! Pretty neat.

6.) Dill 1.23

(a) These chances of winning given in the problem are conditional probabilities, i.e. the probability of winning, given that the weather is good or bad. The approach here is to elucidate the four mutually exclusive and collectively exhaustive outcomes, winning and good weather, P(W,G), winning and bad weather, P(W,B), losing and good weather, P(L,G), and losing and bad weather, P(L,B). These joint probabilities can be related to the conditional probabilities, P(W|G), etc. and the weather probabilities, P(G) and P(B) by the following equations:

P(W,G) = P(W|G)P(G) = (0.7)(0.4) = 0.28
P(W,B) = P(W|B)P(B) = (0.2)(0.6) = 0.12
P(L,G) = P(L|G)P(G) = (0.3)(0.4) = 0.12
P(L,B) = P(L|B)P(B) = (0.8)(0.6) = 0.48

(Note that P(L|G) was computed from using the fact that P(W|G) + P(L|G) = 1.)

(b) P(B|L) = $\frac{P(L,B)}{P(L)}$

P(L) = P(L|G)P(G) + P(L|B)P(B) = (0.3)(0.4) + (0.8)(0.6) = 0.6
Therefore, P(B|L) = 0.48/0.6 = 0.8 – there is an 80% chance there was bad weather, given that they lost.

7.) You have 4 boxes and 3 different (i.e., distinguishable) balls, which are Red, Yellow, and Blue.

a) If there is no restriction on the number of balls per box, how many different ways can the balls be put in the boxes?
One might think this is like the problem we did in lecture on 2/6 – we could consider the 4 boxes as 3 walls that can be distributed between the balls. These walls are indistinguishable, however. So, we have 6 objects in 4 categories: wall, R, Y, B. And there are 3 walls, 1R, 1Y, 1B. So, $W = \frac{6!}{3!1!1!1!}$ = 120. BUT, this overcounts for many situations. For example, this considers RY in one box separately from YR in that box. Since I interpreted this problem such that one shouldn't be able to tell the difference between the arrangements RY and YR in a box, you would have to divide out for all pairs like this… And then for all triples like this… And this starts getting ugly. Why not think about it more simply? How many ways can you place the R ball? 4… How many ways can you place the Y ball? 4… The B ball? 4… So there are **4^3 = 64** ways the balls can be put into the boxes.

b) If we impose a restriction that the first box must contain 2 balls, and the second box must contain 1 ball, how many ways can this be achieved? First, write out all the possibilities. Then, write out the general formula you would use to calculate it, and see if it matches.
All the possibilities:

Boxes:	1	2	3	4 (boxes 3 and 4 always contain 0 balls because of the constraint)
	RY	B		
	RB	Y		
	BY	R		

General formula:
Consider that we have only two boxes available to our three balls, so this problem can be rewritten as: how many ways can you choose 2 balls to put in box 1 (out of a possible 3 balls)? OR how many ways can you choose 1 ball to put in box 2 (of a possible 3 balls)? The answer is given by "3 choose 2" or "3 choose 1" or $\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3$. ("3 choose 1" gives the same answer).

c) What is the probability that the situation in b) (2 balls in box 1, 1 in box 2) was achieved if the balls were thrown in randomly, as in a)?
In part (b) we discovered that there were 3 possible ways to achieve this situation. And in part (a) we found that there are 64 possible ways the balls could fall randomly. So, the probability of achieving this situation is 3/64 = **0.047.**

8.) You have a genome that is exactly 1×10^9 bases long. You would like to choose a DNA sequence out of the genome that is unique. What is the minimum length of the sequence such that it could be unique (i.e., it is possible that it does not occur anywhere else in the genome)?

You want the probability of seeing that exact sequence to be not greater than 1 in 1×10^9.

For a 14mer:
$¼^{14} = 3.725 \times 10^{-9}$ so this is not long enough

For a 15mer:
$¼^{15} = 9.313 \times 10^{-10}$ so this is long enough to be unique!