Motif Discovery and Statistical Analysis
Scribe notes for CS284A

Notes by Nathanael Lenart
based on a lecture by Xiaohui Xie

October 7, 2008

1 Statistics Review

1.1 Generating Functions

We must first define the terms mean and variance. Mean in statistical terms is defined as the expected value of a random variable, given by

$$\mu = E[K] = \sum_{k=0}^{N} kP(K = k), \tag{1}$$

where K is a random variable from 0 to N, and k is a particular value of K. Variance is defined as a measure of statistical dispersion [2], essentially measuring how far out from the mean observed results go. It is given by

$$\sigma^2 = E[(k - \mu)^2] = E[k^2] - \mu^2 \tag{2}$$

For any known distribution, the generating function is

$$G(s) = E[S^K] \tag{3}$$

where G is the function and s is a variable. Therefore, $S^K \epsilon \{S^0, S^1, S^2, ..., S^N\}$, corresponding to $K = \{0, 1, 2, ..., N\}$. This moment generating function works for *all discrete distributions*. It has a number of very interesting properties. For example,

1. Firstly, applying the parameter $s = 1$ to the generating function yields the sum of all probabilities in the distribution, which must sum to 1:

$$G(1) = E[1^K] = \sum_{k=0}^{N} 1 * P(K = k) = 1$$

2. We can take the derivative of the generating function as well:

$$G'(s) = \frac{d}{ds}\left[\sum_{k=0}^{N} s^k P(K = k)\right]$$

$$= \sum_{k=0}^{N} \frac{d}{ds}\left[s^k P(K = k)\right]$$

$$= \sum_{k=0}^{N} k s^{k-1} P(K = k)$$

Using the derivative of the generating function and applying the parameter $s = 1$ to it, we see that

$$G'(1) = \sum_{k=0}^{N} k P(K = k) = E[K] = \mu, \tag{4}$$

the mean of a probability distribution.

3. Similarly, we can take the second derivative of the generating function:

$$G''(s) = \frac{d}{ds}\left[\sum_{k=0}^{N} k s^{k-1} P(K = k)\right]$$

$$= \sum_{k=0}^{N} k(k-1) s^{k-2} P(K = k)$$

Again, applying the parameter $s = 1$ to the generating function, we see that

$$G''(1) = \sum_{k=0}^{N} k(k-1) P(K = k) = E[K^2] - E[K]. \tag{5}$$

This is almost the variance of a probability distribution, known as σ^2. To actually get σ^2, we need to add $G'(1)$ ($= E[K]$) back to $G''(1)$, then subtract $(E[K])^2$ to it, thus obtaining $\sigma^2 = E[K^2] - (E[K])^2$.

1.2 Binomial Distribution

For a binomial distribution, in which we pick k out of N items, we say that the probability is:

$$P(K = k) = \binom{N}{k} p^k (1-p)^{N-k} \tag{6}$$

Therefore, the generating function associated with the binomial distribution is:

$$G(s) = E[s^K] = \sum_{k=0}^{N} s^k \binom{N}{k} p^k (1-p)^{N-k} \tag{7}$$

$$= \sum_{k=0}^{N} \binom{N}{k} (sp)^k (1-p)^{N-k}$$

(Remember that the expansion of $(x+y)^N = \sum_{k=0}^{N} \binom{N}{k} x^k y^{N-k}$.) So, if we set $x = sp$ and $y = 1-p$, we can see that

$$G(s) = (sp + 1 - p)^N \tag{8}$$

and then after setting $s = 1$, we see that $G(1) = 1$, as we expected from our discussion of generating functions in section 1.1.
We can then take the derivative of G with respect to s, in order to obtain the mean for the binomial distribution:

$$G'(s) = N(sp + 1 - p)^{N-1} * p = Np(sp + 1 - p)^{N-1}. \tag{9}$$

Substituting $s = 1$ into equation (9), we see that

$$G'(1) = Np(p + 1 - p)^{N-1} = Np. \tag{10}$$

Thus, for any binomial distribution, $G'(1) =$ the mean, $\mu = Np$.
Taking the second derivative and setting $s = 1$ yields the variance (σ^2) of the binomial distribution:

$$G''(1) = Np(1-p). \tag{11}$$

2 Regulatory Motif Discovery

In regulatory motif discovery, we are looking for certain sequences of DNA (commonly referred to as k-mers, for a DNA string of length k) in much longer

sequences. Given a set of N sequences $S = \{S_0, S_1, S_2, ..., S_{N-1}\}$, each of length L, we should be able to obtain the probability that a particular k-mer would appear in each of these sequences. If we find that the k-mer appears many more times than we expected, we can say that k-mer is *overrepresented*. This is a significant indication that the sequence we looked for is involved in regulatory processes for the sequences we looked in.[4]

2.1 A Look at Probability and Sequential Data

In the world of DNA, our alphabet consists of the set of 4 letters, $\{A, C, G, T\}$. For simplicity, we will assume that each base is equally likely; that is, $P_A = P_C = P_G = P_T = \frac{1}{4}$. Now, we will look at a specific sequence, S, which is a random variable of length L. Let us now consider a particular k-mer of length w, for example, the sequence "ACGTAC."

1. What is the probability that the k-mer will appear in S?
$$p = 1 - \left(1 - \frac{1}{4^w}\right)^{L-w+1}$$

2. Suppose we have N random variables, $S = \{S_0, S_1, S_2, ..., S_{N-1}\}$. Then the probability of the k-mer appearing in k of the N sequences is binomial:
$$P(K = k) = \binom{N}{k} p^k (1-p)^{N-k}$$

3. Suppose that, in these N sequences, we observed the k-mer in a total of y sequences. Now we can calculate the significance ("surprise") of the observation by using *p-values*. A p-value is the probability of observing a result as least as extreme as the one actually observed [1]. In other words, it is a measure of the observed value's distance from the mean. The farther from the mean the result is found, the more surprising (significant) that result is. A p-value is given by:
$$\alpha = P(K = y) + P(K = y + 1) + ... + P(K = N) = \sum_{k=y}^{N} P(K = k)$$

   The lower a p-value is (in absolute value), the more significant the result. The line marking where the cutoff for significance is somewhat arbitrary, but $\alpha = 0.05$ is a commonly accepted value, meaning the occurrence of a result of 5% or less probability is a significant find.

2.2 Enumeration-Based Method of Motif Discovery

Using an enumeration-based method of motif discovery, we would just list all the possible k-mers out in a table. For each k-mer in our table, we can see how many sequences it is found in as a subsequence. If there are N sequences, then we can model it as a binomial distribution, where y_i is a particular k-mer and p_i is its associated probability:

$$p_i = \sum_{k=y_i}^{N} \binom{N}{k} p^k (1-p)^{N-k}.$$

However simple this may seem in theory, it can be quite problematic in practice. For instance, consider k-mers only of length 6. Since we have 4 bases that can each fill 6 positions, we quickly see that there are $4^6 = 4096$ different k-mers of length 6! Clearly, then, the formula above for calculating is too computationally-intensive to actually be of much value. (Remember that $\binom{N}{k}$ evaluates to $\frac{N!}{k!(N-k)!}$, so just for 6-mers, the formula given above would work out to $4096 * 3 = 12288$ factorials!)

2.3 Normal Distribution Approximation

To alleviate the problems of the enumeration-based method above, we note that the binomial distribution is well modeled by the normal distribution, for large values of N. The formula for a normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \tag{12}$$

Using the normal distribution defined above, we can calculate z-scores, which are standardized versions of raw data[3]. Using standardized data, we can then compare our obtained values to the values we expected to see more easily in order to view its significance. A z-score is quite similar to a p-value; in fact, a z-score can be converted into a p-value.
If we say that X is a random variable, then the z-score is given by the equation

$$Z = \frac{X - \mu}{\sigma}. \tag{13}$$

Recall that for the binomial distribution, $\mu = E[X] = Np$ and $\sigma^2 = E\left[(k-\mu)^2\right] = Np(1-p)$, so the z-score for a given k-mer is:

$$z_{y_i} = \frac{y_i - Np}{\sqrt{Np(1-p)}}$$

This is indicative of a standard normal distribution, shifted and scaled such that $\mu = 0$ and $\sigma^2 = 1$.

2.4 Assumptions

In the analyses we performed today, we made a number of assumptions to simplify our work. In reality, things are not that simple, so we must address the assumptions we made.

1. We assumed that the probability of each base was equal. In other words, we assumed,

   $$P_A = P_C = P_G = P_T = \frac{1}{4}$$

   but this is most likely not the case. To fix this assumption, let us assume that the sum of the probabilities still add up to 1, but they are not necessarily equal. This means that the $\frac{1}{4^w}$ term we had in our previous equation, $P = 1 - \left(1 - \frac{1}{4^w}\right)^{L-w+1}$ will need to change. We obtained $\frac{1}{4^w}$ by performing w multiplications of $\frac{1}{4}$, since all the probabilities were equal. We now need to use probabilities specific to the bases in the k-mer we are interested in, i.e. if the k-mer we are interested in is "ACGTAC," then the $\frac{1}{4^w}$ term changes to $P_A P_C P_G P_T P_A P_C$. Thus, the new formula for obtaining the probability of this k-mer in a sequence S is

   $$P = 1 - (1 - P_A P_C P_G P_T P_A P_C)^{L-w+1}$$

   Using this new definition of probability creates no problems in the use of the other formulas given earlier in section 2.1, since they only rely on the probability, which we have now updated.

2. We assumed that each base was independent of other bases. In a real scenario, it is quite possible, for instance, that almost every A encountered is followed immediately by a T, but our model does not account for this possibility.

3. When we checked to see if a particular k-mer was in a sequence S, we merely checked to see if that sequence appeared, returning *true* if it was found, and *false* if it wasn't. However, this does not keep track of whether a k-mer could have appeared multiple times in the same sequence.

4. We did all our calculations based on a particular k-mer, or k-mers of a certain length, but it is entirely possible that k-mers of other lengths exist.

5. We picked a k-mer to search for, and searched for *exactly* that pattern, not allowing any variation to occur. It is quite possible, for example, that the k-mer "CAACTG" could be substituted with the k-mer "CAAGTG" and provide exactly the same result. To account for this, one solution is using a *positional weight matrix* . This matrix would be of dimensions $4 \times L$, where L is the length of the k-mer. Each row in the matrix is for 1 of the 4 bases, and each column stands for each position in a sequence. The sum of each column adds to 1, since they are probabilities of bases occurring in that particular position. Here is an example positional weight matrix:

$$\begin{bmatrix} P_{A1} & P_{A2} & P_{A3} & P_{A4} & P_{A5} & P_{A6} \\ P_{C1} & P_{C2} & P_{C3} & P_{C4} & P_{C5} & P_{C6} \\ P_{G1} & P_{G2} & P_{G3} & P_{G4} & P_{G5} & P_{G6} \\ P_{T1} & P_{T2} & P_{T3} & P_{T4} & P_{T5} & P_{T6} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & .5 & 0 & 0 \\ 0 & 0 & 0 & .5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

This matrix would either correspond to the sequence "CACCTA" or the sequence "CACGTA."

References

[1] *Wikipedia, p-value*, `http://en.wikipedia.org/wiki/P-value`.

[2] *Wikipedia, variance*, `http://en.wikipedia.org/wiki/variance`.

[3] *Wikipedia, z-score*, `http://en.wikipedia.org/wiki/Z-score`.

[4] James Foulds, *Gene regulation and motif discovery*, `http://www.ics.uci.edu/~xhx/courses/CS284A/lectures/CS284A_Scribe_Lec2.pdf`.