Finding Over-Represented 6-mers in Upstream Regions

Suppose we have a set of genes with similar expression profiles. Interpret “upstream region” as the 1000 bp just before the translation start site. (We might want to omit any part that is in another coding region.) The goal is to identify all 6-mers that occur in these upstream regions more frequently than can be expected by chance.

1. How many possible 6-tuples are there?
2. How difficult is it to write a program that counts how many times each 6-tuple occurs in the given sequences?
3. How long will the program take to run?

Counting 6-mers

1. There are $4^6 = 4096$ DNA 6-mers.
2. Writing a program that counts all of them in a given set of sequences would take an experienced programmer at most a day.
3. The program would run in a second.

Frequently, a binding site controls translation when in either orientation. Thus, let's consider a 6-mer equivalent to its reverse complement, e.g., $\texttt{AACCGT} \equiv \texttt{ACGGTT}$.

How many *distinct* 6-mers are there?

Counting Distinct 6-mers

There are $4^3 = 64$ palindromes of length 6. The $4096 - 64 = 4032$ non-palindromes fall into 2016 pairs of equivalent 6-mers.

Answer: $2016 + 64 = 2080$.

Once we’ve counted how frequently each 6-mer occurs in the set of upstream sequences, we could simply report the 6-mers with the highest counts.

What is wrong with that approach?

6-mer Composition

We can expect that different 6-mers will occur at quite different frequencies. For instance, the yeast genome is over 60% C+T, so we expect to see, e.g., CTCTCT much more frequently than AGAGAG. In organisms with methylation, CGCGCG will be extremely rare. Also, 6-mers occur at different frequencies in coding vs. non-coding regions. (Gene-prediction programs rely on this difference.)

We need a statistical model for the expected frequency that each 6-mer occurs by chance in upstream regions.

How can we estimate these “background” frequencies?

Background 6-mer Frequencies

We’ll take all upstream regions in the genome (not just those in our potentially co-regulated genes). We estimate the frequency of a particular 6-mer, say w, as $freq(w) = W/N$, where w occurs W times and there are N occurrences of 6-mers altogether.

Suppose the 6-mer w occurs n times in the set of upstream regions from genes with similar expression profiles.

How do we decided if n is too large to be explained by chance?

Approximating the Exact Probability

For now, we will think of the occurrences of 6-mers as independent. This is only an approximation, since if you know one 6-mer, then there are only 4 possibilities for the next one (shifting over one nucleotides). We take the probability that a particular 6-mer is w to be the observed frequency $freq(w)$.

Suppose that there are altogether T 6-mers in the upstream regions of the putatively co-regulated genes.

For a fixed $n \leq T$, what is the probability that w occurs precisely n times?

Approximating the Correct Probability II

If each occurrence of w has probability $freq(w)$, the the probability that precisely n of the T total occurrences are w is:

$$P(\#w = n) =$$
$$\frac{T!}{n!(T-n)!} \times freq(w)^N \times (1 - freq(w))^{(T-n)}$$

What is the probability that w occurs n or more times?

Approximating the Correct Probability III

Since $P(\#w = n)$ denotes the probability that w occurs precisely n times, the probability that w occurs n or more times is:

$$P(\#w \geq n) = \sum_{j=n}^{T} P(\#w = j) = 1 - \sum_{j=0}^{n-1} P(\#w = j)$$

Testing for Statistical Significance

Scenario: We're given a 6-mer w that occurs n times in the upstream regions of our putatively co-regulated genes, and $P(\#w \geq n) = 10^{-3}$. Can we conclude that a 6-mer with this many occurrences is unlikely to happen by chance alone?

Testing for Statistical Significance II

Since there are 2080 6-mers (counting a 6-mer as equivalent to its reverse complement), we can expect there to be two 6-mers w such that

$$P(\#w \geq n_w) = 10^{-3},$$

where w occurs n_w times.

Let $D = 2080$ and set:

$$sig = -\log_{10}[P(\#w \geq n) \times D]$$

One expects to find at random a 6-mer with $sig \geq 1$ every 10 clusters, a 6-mer with $sig \geq 2$ every 100 clusters, etc.

Finding the Exact Probability

Here, we consider an extremely simply case. Consider strings of `0` or `1`. Fix a “pattern” string w of length 2 and a “random” string s of length 5. For each n, what is the probability that w occurs in s precisely n times?

Finding the Exact Probability II

s	#00	#01
00000	4	0
00001	3	1
00010	2	1
...		
11111	0	0

n	0	1	2	3	4
00	13	10	6	2	1
01	6	20	6	0	0

For both words, the average number of occurrences in a string of length 5 is 1, but the variance is much larger for 00 than for 01.

Finding the Exact Probability III

Bottom line: Finding the exact probability that word w occurs in a random string s precisely n times is difficult. Much of the complexity comes from “self-overlapping words”. (Note that two matches to 00 can overlap, as in 000, but two matches to 01 can’t.) The problem has been solved for simple models of “random string s”, such as Markov models.