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., ACCGTA ACCGTA.
There are $4096 = 4032$ non-palindromes and $64$ palindromes of length 6. Therefore, $64$ of equivalent 6-mers. The count of 2016 pairs of equivalent 6-mers falls into 2016 pairs. The count of 2080 pairs. 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?

How can we estimate these "background" frequencies?

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

this difference.

non-coding regions. (Gene-prediction programs rely on 6-mers occur at different frequencies in coding vs.
6-mers occur at different frequencies in coding vs.

methylation, CCGCG will be extremely rare. Also, much more frequently than AGAGA. In organisms with over 60% C+T, so we expect to see, e.g., CCGCG

different frequencies. For instance, the yeast genome is 4

6-mer Composition
We'll take all upstream regions in the genome (not just those in our potentially co-regulated genes). We'll take all upstream regions from genes with similar expression profiles.

Suppose the 6-mer $w$ occurs $n$ times in the set of upstream regions from genes with similar expression profiles. How do we decide if $n$ is too large to be explained by chance?

The frequency of a particular 6-mer, say $w$, as estimated the frequency of a particular 6-mer, say $w$, as $\frac{N}{M} = \binom{M}{n}$ be the number of occurrences of 6-mers altogether:

are occurrences of 6-mers altogether, where $m$ occurs $m$ times and there

Background 6-mer Frequencies
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 to be the observed frequency \( \text{freq}(w) \). Suppose that there are altogether \( \sum L \) 6-mers in the upstream regions of the putatively co-regulated genes.

For a fixed \( n \) 6-mers, what is the probability that \( m \) occurs precisely \( n \) times?

\[
\sum_{L \geq n} \binom{L}{n} \text{freq}(w)^n \cdot \left(1 - \text{freq}(w)\right)^{L-n}
\]
Approximating the Correct Probability for Each Occurrence of \( w \) with Probability \( \text{freq}(w) \)

If each occurrence of \( w \) has probability \( \text{freq}(w) \)

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

\[
(u - L) ((m) \beta \Gamma f - 1) \times N(m) \beta \Gamma f \times \frac{i(u-L)i u}{i L} = (u = m \#) P
\]
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:

$$P(w \geq n) = \sum_{j=n}^{\infty} P(w = j)$$

or more formally:

$$P(w = m \#) \sum_{j=0}^{\infty} - 1 = P(w = m \#) \sum_{j=1}^{\infty} = (u < m \#) \sum_{j=0}^{u}$$

Since $P$ denotes the probability that $w = m \#$
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?
One expects to find at random a 6-mer with \( w \) occurring \( n \) times. Let \( D = 2080 \) and set:

\[
\log_{10} \left( \frac{D \times (n \leq m\#)}{D} \right) = \frac{0.1 \log_{10} D}{D} = \text{sig}
\]

where \( w \) occurs \( n \) times, \( m \) such that there to be two 6-mers \( w \) such that equivalent to its reverse complement, we can expect since there are 2080 6-mers (counting a 6-mer as...
Finding the Exact Probability

Here, we consider an extremely simple case. Consider a "random" string $s$ of length 5. For each $n$, what is the probability that pattern string $w$ of length 2 and strings of 0 or 1. Fix a "pattern" string $w$ of length 2 and

probability that $w$ occurs in $s$ precisely $n$ times?
Finding the Exact Probability

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 that a 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 01 cannot overlap, as in 00, but two matches to 01 can. The problem has been solved for simple models of "random strings," such as Markov models.

Bottom line: Finding the exact probability that a word $w$ occurs in a random string $s$ precisely $n$ times is difficult.