Probabilities and Statistics in Shotgun Sequencing

Shotgun Sequencing.

Before any analysis of a DNA sequence can take place it is first necessary to determine the actual sequence itself, at least as accurately as is reasonably possible. Unfortunately, technical considerations make it impossible to sequence very long pieces of DNA all at once. Current sequencing technologies allow accurate reading of no more than 500 to 800bp of contiguous DNA sequence. This means that the sequence of an entire genome must be assembled from collections of comparatively short subsequences. This process is called DNA sequence "*assembly*".

One approach of sequence assembly is to produce the sequence of a DNA segment (called as a "contig", or perhaps a genome) from a large number of randomly chosen sequence reads (many overlapping small pieces, each on the order of 500-800 bases). One difficulty of this process is that the locations of the fragments within the genome and with respect to each other are not generally known. However, if enough fragments are sequenced so that there will be many overlaps between them, the fragments can be matched up and assembled. This method is called "*shotgun sequencing*."

Shotgun sequencing approaches, including the whole-genome shotgun approach, are currently a central part of all genome-sequencing efforts. These methods require a high level of automation in sample preparation and analysis and are heavily reliant on the power of modern computers. There is an interplay between substrates to be sequenced (genomes and their representation in clone libraries), the analytical tools for generating a DNA sequence, the sequencing strategies, and the computational methods. The key underlying determinant is that we can obtain high-quality continuous sequence reads of up to 500 to 800 bases with current technology. This represents a tiny fraction of either a prokaryotic or eukaryotic genome. The problem in large measure is defined by the need to assemble a larger whole from a large number of small parts. Therefore, a large number of randomly generated sequence reads should be used to generate sequences at appropriately large levels of "*sequence coverage*"

Coverage Theorem.

Sequence coverage is the average number of times any given genomic base is represented in sequence reads.

Definition--It is customary to say that a-times coverage (or aX coverage) is obtained if, when the length of original long sequence is G, the total length of the fragments sequenced is aG.

Theorem 1--Assuming that there are N fragments of length L each and the length of original long sequence is G, the coverage is $a=NL/G$. Then in order for the mean proportion of the original long sequence covered by at lease one fragment to be 0.99, it is necessary to have at least 4.6X coverage.

Before proving the above theorem, let us see the distribution of the location of left-hand end of fragments. If we assume that the fragments are taken at random from the original full-length sequence, so that, ignoring end effects, the position of the left-hand end of any fragment is uniformly distributed in $(0, G)$, and thus falls in an interval $(x, x+L)$ on the original sequence with probability L/G. The number of fragments whose left-hand end falls in this interval has a binomial

distribution with mean *NL/G*. If *N* is large and *L* is small, the discussion on "Poisson is Ultimate Binomial" above (1) shows that this distribution is approximately Poisson with mean *NL/G*.

Proof of theorem 1: The mean proportion of the genome covered by one or more fragments is the probability that a point chosen at random is covered by at least one fragment. This is the probability that at least one fragment has its left-hand end in the interval of length *L* immediately to the left of this point, which is approximately $1-P(X=0)=1-\frac{(NL/G)^0 e^{-NL/G}}{0!}=1-e^{-NL/G}$. In order for $1-e^{-NL/G}$ to be 0.99 it is necessary to have *NL/G*=4.6, so that the sum of the fragment lengths is then not quite five times the genome length.

Note that since the human genome is approximately 3×10^9 nucleotides, 4.6X coverage will still miss approximately 30,000,000 of them.

Mean number of contigs

Each contig has a unique rightmost fragment, so that the mean number of contigs is the number of fragments *N* multiplied by the probability that a fragment is the rightmost member of a contig. This latter probability is the probability that no other fragment has its left-hand end point on the fragment in question. From Poisson probability function, this probability is $e^{-NL/G}$. Thus mean number of contigs = $Ne^{-NL/G}$.