

Shotgun sequencing

From Wikipedia, the free encyclopedia

In genetics, **shotgun sequencing**, also known as **shotgun cloning**, is a method used for sequencing long DNA strands. It is named by analogy with the rapidly-expanding, quasi-random firing pattern of a shotgun.

Since the chain termination method of DNA sequencing can only be used for fairly short strands (100 to 1000 basepairs), longer sequences must be subdivided into smaller fragments, and subsequently re-assembled to give the overall sequence. Two principal methods are used for this: chromosome walking, which progresses through the entire strand, piece by piece, and shotgun sequencing, which is a faster but more complex process, and uses random fragments.

In shotgun sequencing,^{[1][2]} DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain *reads*. Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing. Computer programs then use the overlapping ends of different reads to assemble them into a continuous sequence.^[1]

Shotgun sequencing was one of the precursor technologies that was responsible for enabling full genome sequencing.

Contents

- 1 Example
- 2 Whole genome shotgun sequencing
 - 2.1 Coverage
- 3 Next-generation sequencing
- 4 References
- 5 External links

Example

For example, consider the following two rounds of shotgun reads:

Strand	Sequence
Original	AGCATGCTGCAGTCATGCTTAGGCTA
First shotgun sequence	AGCATGCTGCAGTCATGCT----- -----TAGGCTA
Second shotgun sequence	AGCATG----- -----CTGCAGTCATGCTTAGGCTA
Reconstruction	AGCATGCTGCAGTCATGCTTAGGCTA

In this extremely simplified example, none of the reads cover the full length of the original sequence, but the

four reads can be assembled into the original sequence using the overlap of their ends to align and order them. In reality, this process uses enormous amounts of information that are rife with ambiguities and sequencing errors. Assembly of complex genomes is additionally complicated by the great abundance of repetitive sequence, meaning similar short reads could come from completely different parts of the sequence.

Many overlapping reads for each segment of the original DNA are necessary to overcome these difficulties and accurately assemble the sequence. For example, to complete the Human Genome Project, most of the human genome was sequenced at 12X or greater *coverage*; that is, each base in the final sequence was present, on average, in 12 reads. Even so, current methods have failed to isolate or assemble reliable sequence for approximately 1% of the (euchromatic) human genome.

Whole genome shotgun sequencing

Whole genome shotgun sequencing for small (4000 to 7000 basepair) genomes was already in use in 1979.^[1] Broader application benefited from pairwise end sequencing, known colloquially as *double-barrel shotgun sequencing*. As sequencing projects began to take on longer and more complicated DNAs, multiple groups began to realize that useful information could be obtained by sequencing both ends of a fragment of DNA. Although sequencing both ends of the same fragment and keeping track of the paired data was more cumbersome than sequencing a single end of two distinct fragments, the knowledge that the two sequences were oriented in opposite directions and were about the length of a fragment apart from each other was valuable in reconstructing the sequence of the original target fragment. The first published description of the use of paired ends was in 1990^[3] as part of the sequencing of the human HGPRT locus, although the use of paired ends was limited to closing gaps after the application of a traditional shotgun sequencing approach. The first theoretical description of a pure pairwise end sequencing strategy, assuming fragments of constant length, was in 1991.^[4] At the time, there was community consensus that the optimal fragment length for pairwise end sequencing would be three times the sequence read length. In 1995 Roach et al.^[5] introduced the innovation of using fragments of varying sizes, and demonstrated that a pure pairwise end-sequencing strategy would be possible on large targets. The strategy was subsequently adopted by The Institute for Genomic Research (TIGR) to sequence the genome of the bacterium *Haemophilus influenzae* in 1995^[6], and then by Celera Genomics to sequence the *Drosophila melanogaster* (fruit fly) genome in 2000,^[7] and subsequently the human genome.

To apply the strategy, high-molecular-weight DNA is sheared into random fragments, size-selected (usually 2, 10, 50, and 150 kb), and cloned into an appropriate vector. The clones are then sequenced from both ends using the chain termination method yielding two short sequences. Each sequence is called an *end-read* or *read* and two reads from the same clone are referred to as *mate pairs*. Since the chain termination method usually can only produce reads between 500 and 1000 bases long, in all but the smallest clones, mate pairs will rarely overlap.

The original sequence is reconstructed from the reads using sequence assembly software. First, overlapping reads are collected into longer composite sequences known as *contigs*. Contigs can be linked together into *scaffolds* by following connections between mate pairs. The distance between contigs can be inferred from the mate pair positions if the average fragment length of the library is known and has a narrow window of deviation. Depending on the size of the gap between contigs, different techniques can be used to find the sequence in the gaps. If the gap is small (5-20kb) then the use of PCR to amplify the region is required, followed by sequencing. If the gap is large (>20kb) then the large fragment is cloned in special vectors such as BAC (Bacterial artificial chromosomes) followed by sequencing of the vector.

Proponents of this approach argue that it is possible to sequence the whole genome at once using large arrays of sequencers, which makes the whole process much more efficient than more traditional approaches. Detractors

argue that although the technique quickly sequences large regions of DNA, its ability to correctly link these regions is suspect, particularly for genomes with repeating regions. As sequence assembly programs become more sophisticated and computing power becomes cheaper, it may be possible to overcome this limitation^[*citation needed*].

Coverage

Coverage is the average number of reads representing a given nucleotide in the reconstructed sequence. It can be calculated from the length of the original genome (*G*), the number of reads(*N*), and the average read length(*L*) as $N \times L/G$. For example, a hypothetical genome with 2,000 base pairs reconstructed from 8 reads with an average length of 500 nucleotides will have 2x redundancy. This parameter also enables one to estimate other quantities, such as the percentage of the genome covered by reads (sometimes also called coverage). A high coverage in shotgun sequencing is desired because it can overcome errors in base calling and assembly. The subject of DNA sequencing theory addresses the relationships of such quantities.

Sometimes a distinction is made between sequence coverage and physical coverage. Sequence coverage is the average number of times a base is read (as described above). Physical coverage is the average number of times a base is read or spanned by mate paired reads.^[8]

Next-generation sequencing


Although shotgun sequencing was the most advanced technique for sequencing genomes from about 1995–2005, other technologies have surfaced, called next-generation sequencing. These technologies produce shorter reads (anywhere from 25–500bp) but many hundreds of thousands or millions of reads in a relatively short time (on the order of a day).^[9] This results in high coverage, but the assembly process is much more computationally expensive. These technologies are vastly superior to shotgun sequencing due to the high volume of data and the relatively short time it takes to sequence a whole genome. The major disadvantage is that the accuracies are usually lower (although this is compensated for by the high coverage).^[10]

References

- ↑ ^{*a b c*} Staden, R (1979). "A strategy of DNA sequencing employing computer programs". *Nucleic Acids Research* **6** (7): 2601–10. doi:10.1093/nar/6.7.2601 (http://dx.doi.org/10.1093%2Fnar%2F6.7.2601) . PMC 327874 (http://www.pubmedcentral.gov/articlerender.fcgi?tool=pmcentrez&artid=327874) . PMID 461197 (http://www.ncbi.nlm.nih.gov/pubmed/461197) .
- ↑ Anderson, S (1981). "Shotgun DNA sequencing using cloned DNase I-generated fragments". *Nucleic Acids Research* **9** (13): 3015–27. doi:10.1093/nar/9.13.3015 (http://dx.doi.org/10.1093%2Fnar%2F9.13.3015) . PMC 327328 (http://www.pubmedcentral.gov/articlerender.fcgi?tool=pmcentrez&artid=327328) . PMID 6269069 (http://www.ncbi.nlm.nih.gov/pubmed/6269069) .
- ↑ Edwards, A; Caskey, T (1991). "Closure strategies for random DNA sequencing". *Methods: A Companion to Methods in Enzymology* **3** (1): 41–47. doi:10.1016/S1046-2023(05)80162-8 (http://dx.doi.org/10.1016%2FS1046-2023%2805%2980162-8) .
- ↑ Edwards, A; Voss, H.; Rice, P.; Civitello, A.; Stegemann, J.; Schwager, C.; Zimmerman, J.; Erfle, H.; Caskey, T.; Ansorge, W. (1990). "Automated DNA sequencing of the human HPRT locus". *Genomics* **6** (4): 593–608. doi:10.1016/0888-7543(90)90493-E (http://dx.doi.org/10.1016%2F0888-7543%2890%2990493-E) . PMID 2341149 (http://www.ncbi.nlm.nih.gov/pubmed/2341149) .
- ↑ Roach, JC; Boysen, C; Wang, K; Hood, L (1995). "Pairwise end sequencing: a unified approach to genomic mapping and sequencing". *Genomics* **26** (2): 345–353. doi:10.1016/0888-7543(95)80219-C (http://dx.doi.org/10.1016%2F0888-7543%2895%2980219-C) . PMID 7601461 (http://www.ncbi.nlm.nih.gov/pubmed/7601461) .

6. ^ Fleischmann, RD; et al. (1995). "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.". *Science* **269** (5223): 496–512. doi:10.1126/science.7542800 (http://dx.doi.org/10.1126%2Fscience.7542800) . PMID 7542800 (http://www.ncbi.nlm.nih.gov/pubmed/7542800) .
 7. ^ Adams, MD; et al. (2000). "The genome sequence of *Drosophila melanogaster*". *Science* **287** (5461): 2185–95. doi:10.1126/science.287.5461.2185 (http://dx.doi.org/10.1126%2Fscience.287.5461.2185) . PMID 10731132 (http://www.ncbi.nlm.nih.gov/pubmed/10731132) .
 8. ^ http://www.nature.com/nrg/journal/v11/n10/fig_tab/nrg2841_F1.html
 9. ^ Karl, V; et al (2009). "Next Generation Sequencing: From Basic Research to Diagnostics". *Clinical Chemistry* **55** (4): 41–47. doi:10.1373/clinchem.2008.112789 (http://dx.doi.org/10.1373%2Fclinchem.2008.112789) . PMID 19246620 (http://www.ncbi.nlm.nih.gov/pubmed/19246620) .
 10. ^ Metzker, Michael L. (2010). "Sequencing technologies - the next generation". *Nat Rev Genet* **11** (1): 31–46. doi:10.1038/nrg2626 (http://dx.doi.org/10.1038%2Fnrg2626) . PMID 19997069 (http://www.ncbi.nlm.nih.gov/pubmed/19997069) .
- "Shotgun sequencing comes of age" (http://www.the-scientist.com/news/20021231/06) . *The Scientist*. http://www.the-scientist.com/news/20021231/06 . Retrieved December 31, 2002.
 - "Shotgun sequencing finds nanoorganisms - Probe of acid mine drainage turns up unsuspected virus-sized Archaea" (http://www.spaceref.com/news/viewpr.rss.html?pid=21532) . *SpaceRef.com*. http://www.spaceref.com/news/viewpr.rss.html?pid=21532 . Retrieved December 23, 2006.

External links

 *This article incorporates public domain material from the National Center for Biotechnology Information document "NCBI Handbook" (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=handbook.TOC&depth=2) .*

Retrieved from "http://en.wikipedia.org/wiki/Shotgun_sequencing"

Categories: Molecular biology

- This page was last modified on 2 May 2011 at 22:43.
 - Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. See Terms of Use for details.
- Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.