Brief overview of genome sequencing

BIOL 8803 Bioinformatics

Georgia Tech

Nov 13, 2003

Russell Hanson

Sequencing projects

- Human Genome Project (divided work among three labs)
  - Sanger Center – John Sulston (Brixton, UK)
  - Whitehead Institute – Eric Lander (Cambridge, MA)
  - WUSTL – Bob Waterston (St. Louis, MO)
- Private Projects
  - Celera Genomics, a small company, with a lot of assets, and a recent interest in creating synthetic life, base-by-base.
- Financing
  - Wellcome Trust, a private trust
  - US Government DOE/NIH (~$\$3*10^9$)
  - Venture Capital, 1/10th of amount spent publicly (~$\$3*10^8$)
- Publishing
  - “Simultaneous” with Celera, sequence must be deposited in public database of Nature/Science magazines.

How to sequence in a couple easy steps, with a fat check book, and a taste for repetition

- Eric Lander’s paper International Human Genome Sequencing Consortium. *Initial sequencing and analysis of the human genome*. Nature, 15:860--921, 2001.

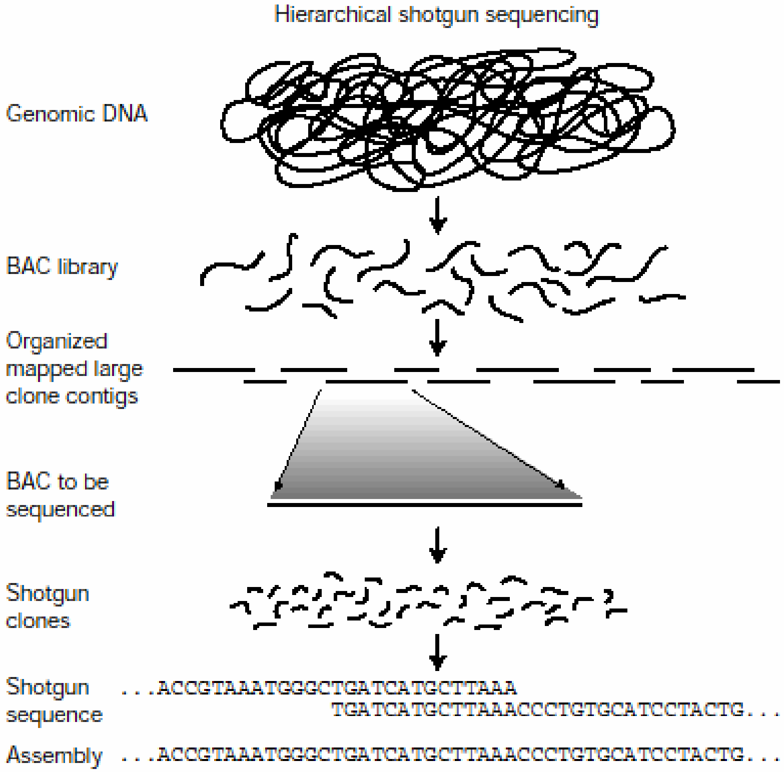

Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

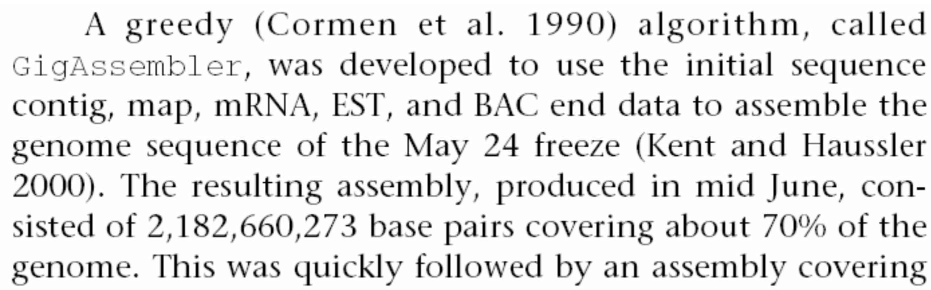
Figure 2 Idealized representation of the hierarchical shotgun sequencing strategy. A library is constructed by fragmenting the target genome and cloning it into a large-fragment cloning vector; here, BAC vectors are shown. The genomic DNA fragments represented in the library are then organized into a physical map and individual BAC clones are selected and sequenced by the random shotgun strategy. Finally, the clone sequences are assembled to reconstruct the sequence of the genome.

More on HGP sequencing

- **Scaffold:** The result of connecting contigs by linking information from paired-end reads from plasmids, paired-end reads from BACs, known messenger RNAs or other sources. The contigs in a scaffold are ordered and oriented with respect to one another.
- **BAC clone:** Bacterial artifcial chromosome vector carrying a genomic DNA insert, typically 100±200 kb. Most of the large-insert clones sequenced in the project were BAC clones.
- **Fingerprint clone contigs:** Contigs produced by joining clones inferred to overlap on the basis of their restriction digest fingerprints.
- The BAC library is constructed by fragmenting the original genome and cloning it into large-fragment cloning vectors. The genomic DNA fragments in the BAC clones are then organized into a physical map (often with the aid of fingerprint scafolding). Individual BAC clones are then selected and sequenced by an automated process using the random shotgun method. After the BACs are sequenced, the sequences are assembled reconstructing the sequence of the genome.

Assembly

- James Kent (UC Santa Cruz) writes GigAssembler which assembles the highly fragmented BACs, ESTs, contigs after the sequencing freeze. This algorithm uses "rafts" and "bridges" to group and merge pieces of assembly. (Kent WJ, Haussler D. Genome assembly of the working draft of the human genome with GigAssembler. Genome Res. 2001 Sep;11(9):1541-8)



A greedy (Cormen et al. 1990) algorithm, called GigAssembler, was developed to use the initial sequence contig, map, mRNA, EST, and BAC end data to assemble the genome sequence of the May 24 freeze (Kent and Haussler 2000). The resulting assembly, produced in mid June, consisted of 2,182,660,273 base pairs covering about 70% of the genome. This was quickly followed by an assembly covering

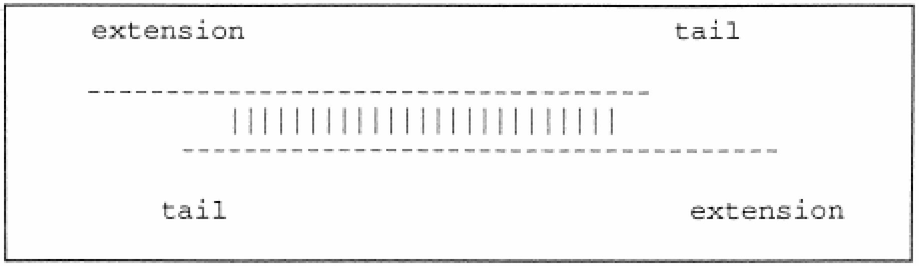
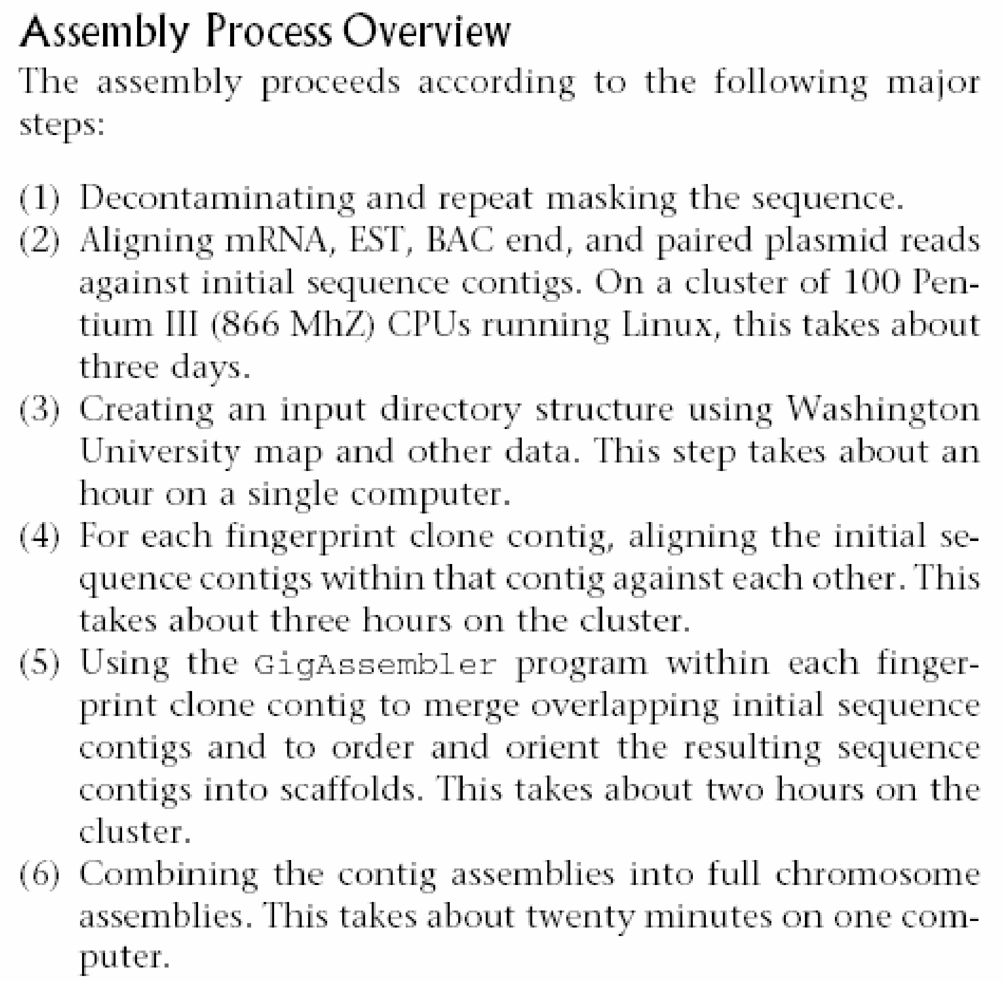


Figure 2 Two sequences with tails. The nonaligning regions on either side can be classified into 'extensions' and 'tails.' Short tails are fairly common even when two sequences should be joined into a contig because of poor quality sequence near the ends and occasional chimeric reads. Long tails, however, are generally a sign that the alignment is merely due to the sequences sharing a repeating element.

GigAssembler's assembly process



Assembly Process Overview

The assembly proceeds according to the following major steps:

(1) Decontaminating and repeat masking the sequence.

(2) Aligning mRNA, EST, BAC end, and paired plasmid reads against initial sequence contigs. On a cluster of 100 Pentium III (866 MhZ) CPUs running Linux, this takes about three days.

(3) Creating an input directory structure using Washington University map and other data. This step takes about an hour on a single computer.

(4) For each fingerprint clone contig, aligning the initial sequence contigs within that contig against each other. This takes about three hours on the cluster.

(5) Using the `GigAssembler` program within each fingerprint clone contig to merge overlapping initial sequence contigs and to order and orient the resulting sequence contigs into scaffolds. This takes about two hours on the cluster.

(6) Combining the contig assemblies into full chromosome assemblies. This takes about twenty minutes on one computer.

This computer shotgunned fragment assembly stuff wasn’t totally new, i.e. back at the ranch…

- Gene Myers, at Colorado then Arizona CS Depts, had been quietly working on multiple fragment assembly throughout the late 80s (Kececioglu as well). Eventually he/his team got the job of writing the *Celera Assembler*, the pipeline for which is pictured at right. (Myers, E.W. et al. A Whole-Genome Assembly of Drosophila 2000. Science 287: 2196-2204)
- An early software package, UniFak (UNIX suite for fragment assembly)

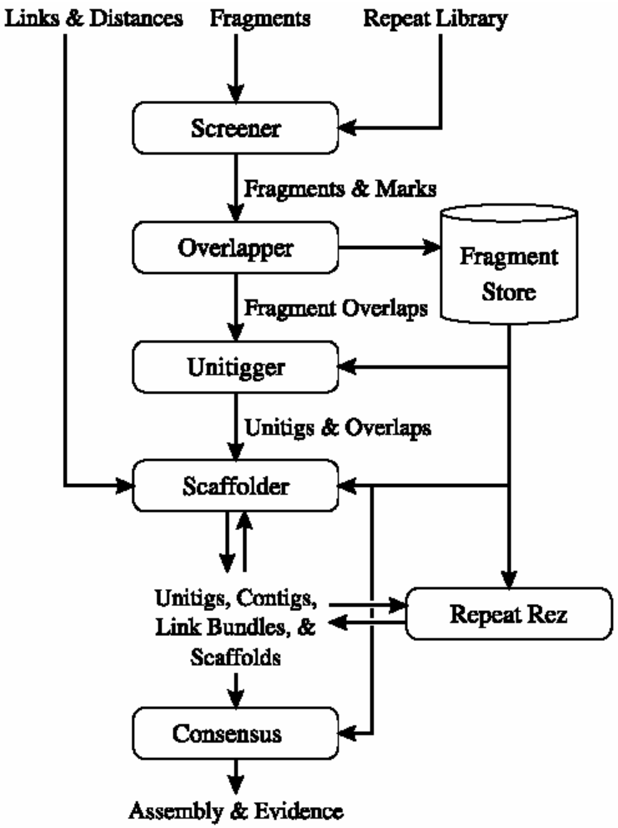
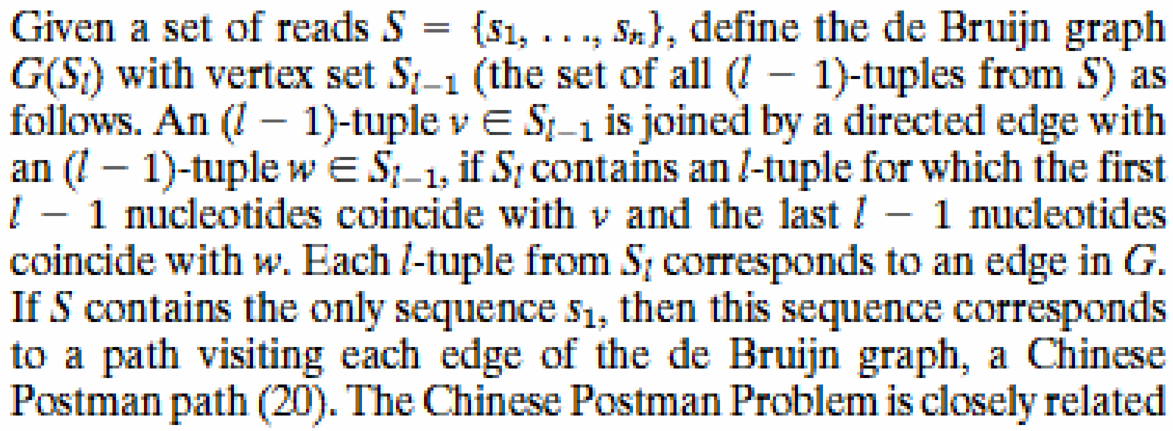


Fig. 1. Assembly pipeline. From an engineering perspective, sequences of messages flow from one stage to the next. Each stage performs work on its input stream, producing a stream of output messages reflecting its transformational function. The text gives the function of each stage.

Eulerian superpath assembly

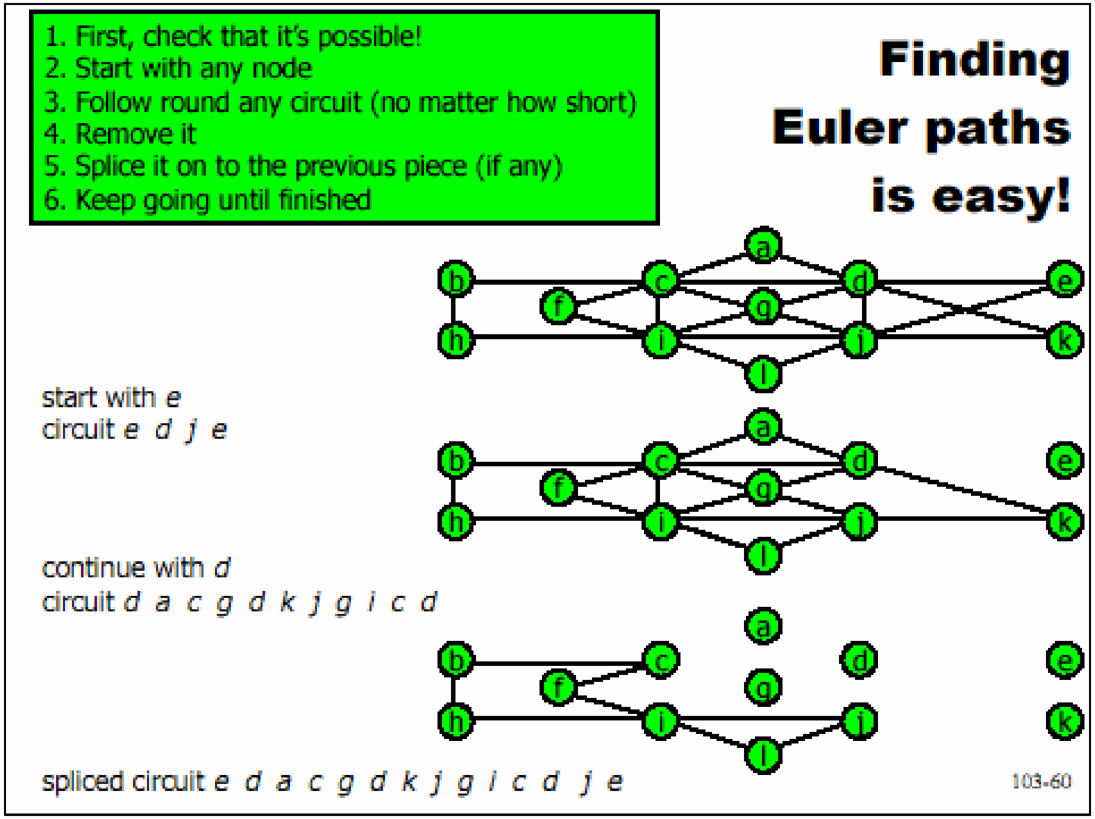
- Find a path visiting every EDGE exactly once (vertices may be repeated): Eulerian path problem (Pevzner et al. PNAS August 14, 2001 vol. 98 no. 17)



Given a set of reads $S = \{s_1, \ldots, s_n\}$, define the de Bruijn graph $G(S_l)$ with vertex set S_{l-1} (the set of all $(l-1)$-tuples from S) as follows. An $(l-1)$-tuple $v \in S_{l-1}$ is joined by a directed edge with an $(l-1)$-tuple $w \in S_{l-1}$, if S_l contains an l-tuple for which the first $l-1$ nucleotides coincide with v and the last $l-1$ nucleotides coincide with w. Each l-tuple from S_l corresponds to an edge in G. If S contains the only sequence s_1, then this sequence corresponds to a path visiting each edge of the de Bruijn graph, a Chinese Postman path (20). The Chinese Postman Problem is closely related

- This is different from the classical method of overlap-layout-consensus, which depends on the overlap graph. Instead of keeping the pieces, or contigs, whole, they are cut up into smaller *regular* pieces, changing the Layout Problem to the Euler Path Problem.

Euler path algorithm



1. First, check that it's possible!
2. Start with any node
3. Follow round any circuit (no matter how short)
4. Remove it
5. Splice it on to the previous piece (if any)
6. Keep going until finished

Finding Euler paths is easy!

start with *e*
circuit *e d j e*

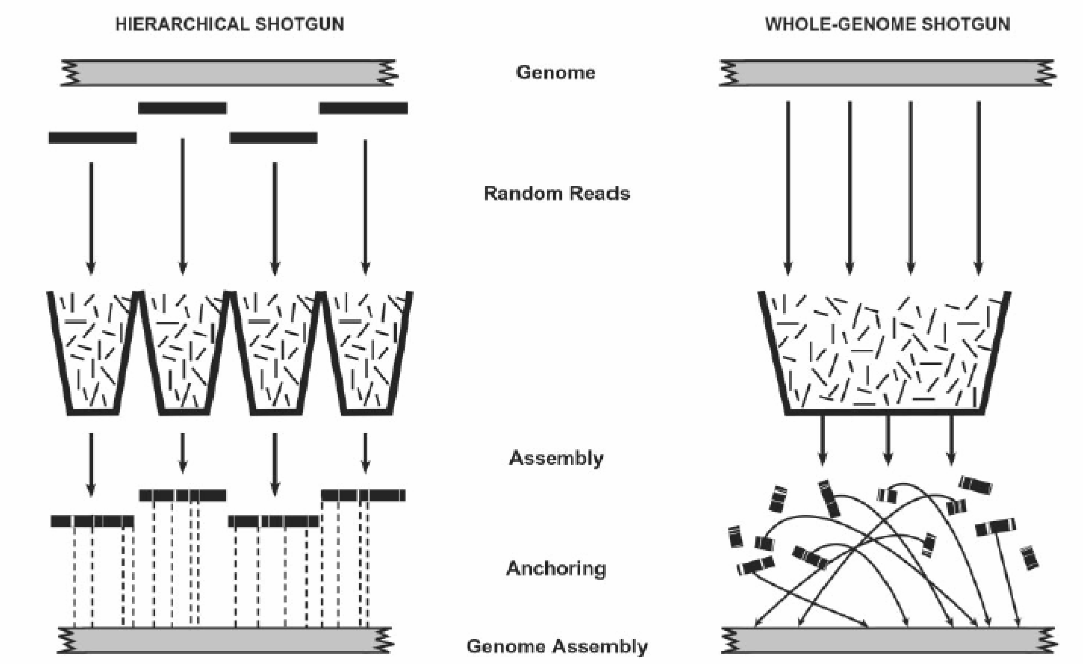
continue with *d*
circuit *d a c g d k j g i c d*

spliced circuit *e d a c g d k j g i c d j e*

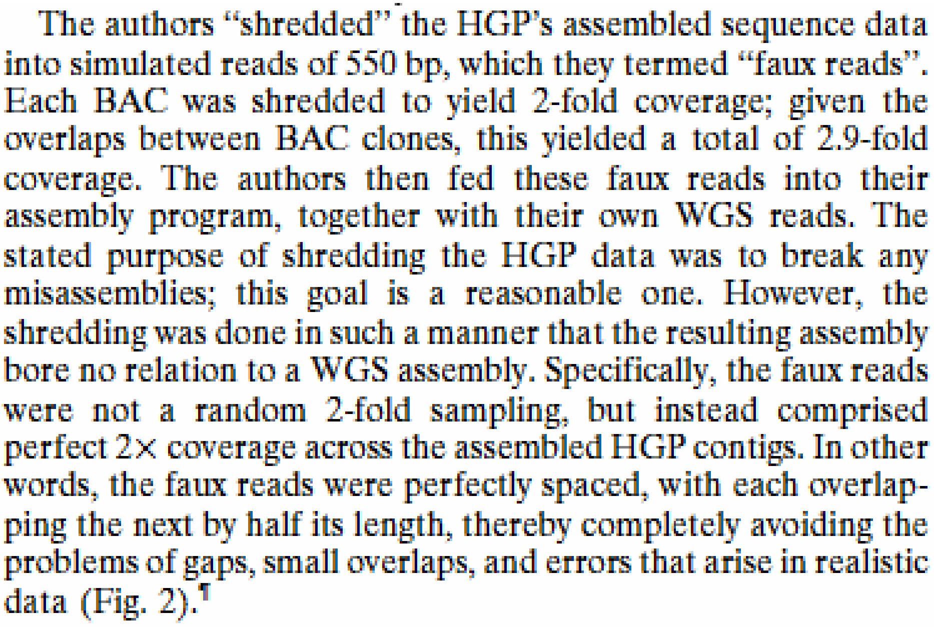
103-60

Celera’s paper used whole-genome shotgun, so dependent on HGP data, which used hierarchical shotgun assembly

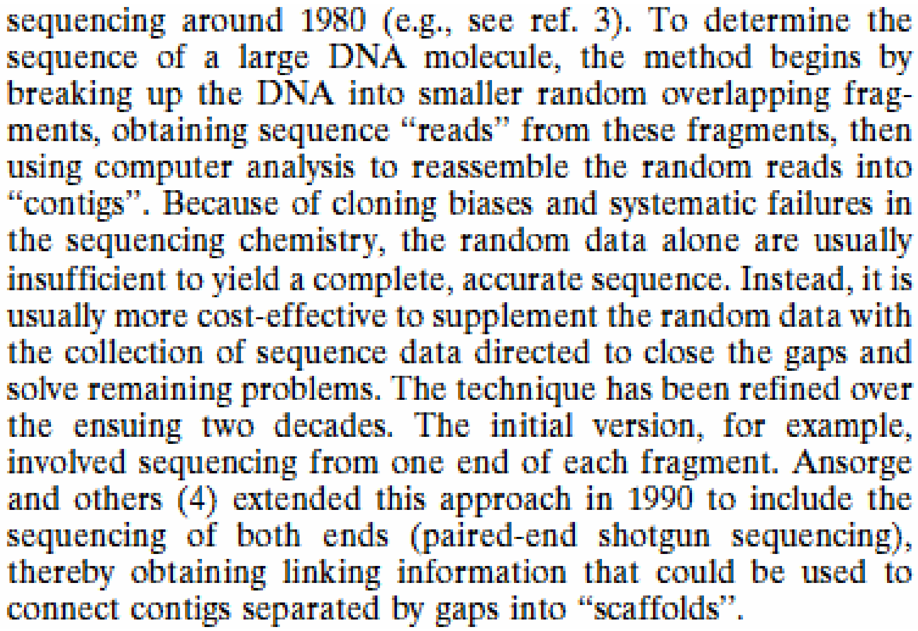
Waterston, “On the sequencing of the human genome,” 3712–3716 PNAS March 19, 2002 vol. 99 no. 6.



WGS vs. HGS cont.



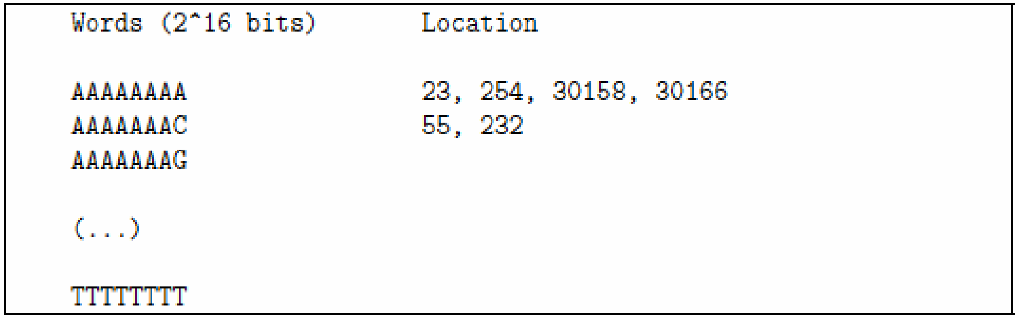
The authors “shredded” the HGP’s assembled sequence data into simulated reads of 550 bp, which they termed “faux reads”. Each BAC was shredded to yield 2-fold coverage; given the overlaps between BAC clones, this yielded a total of 2.9-fold coverage. The authors then fed these faux reads into their assembly program, together with their own WGS reads. The stated purpose of shredding the HGP data was to break any misassemblies; this goal is a reasonable one. However, the shredding was done in such a manner that the resulting assembly bore no relation to a WGS assembly. Specifically, the faux reads were not a random 2-fold sampling, but instead comprised perfect 2× coverage across the assembled HGP contigs. In other words, the faux reads were perfectly spaced, with each overlapping the next by half its length, thereby completely avoiding the problems of gaps, small overlaps, and errors that arise in realistic data (Fig. 2).[¶]



sequencing around 1980 (e.g., see ref. 3). To determine the sequence of a large DNA molecule, the method begins by breaking up the DNA into smaller random overlapping fragments, obtaining sequence “reads” from these fragments, then using computer analysis to reassemble the random reads into “contigs”. Because of cloning biases and systematic failures in the sequencing chemistry, the random data alone are usually insufficient to yield a complete, accurate sequence. Instead, it is usually more cost-effective to supplement the random data with the collection of sequence data directed to close the gaps and solve remaining problems. The technique has been refined over the ensuing two decades. The initial version, for example, involved sequencing from one end of each fragment. Ansorge and others (4) extended this approach in 1990 to include the sequencing of both ends (paired-end shotgun sequencing), thereby obtaining linking information that could be used to connect contigs separated by gaps into “scaffolds”.

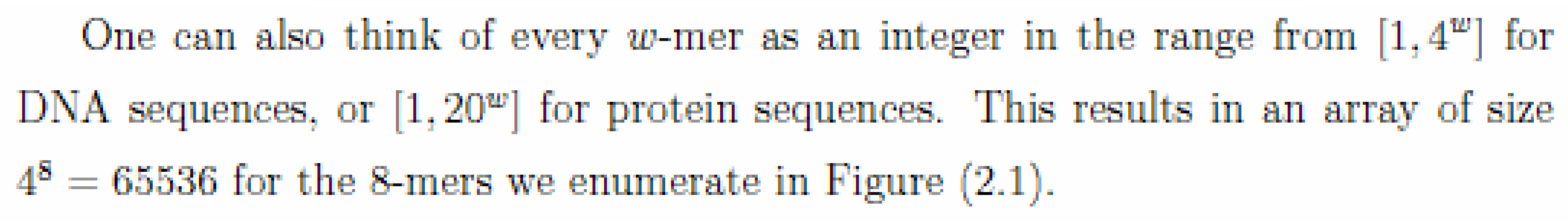
BLAST brute-force

- Hash the nucleic acid *w*-mers. Shift the frame. Record all the matches. Run some statistics (generate an *E* value). Observe that setting word size *w* equal to the database length, it will never finish.
- Only two bits for each letter, if there is a U or N, a random nucleotide is chosen (see the BLAST book).



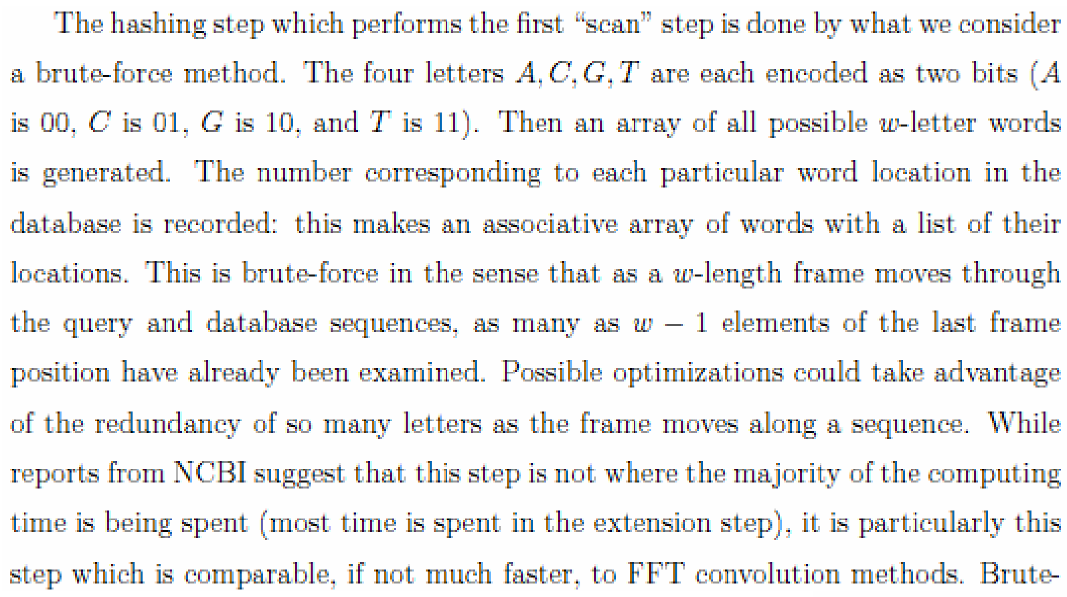
Words (2^16 bits)	Location
AAAAAAAA	23, 254, 30158, 30166
AAAAAAAC	55, 232
AAAAAAAG	
(...)	
TTTTTTTT	

Figure 2.1: BLAST brute-force table lookup

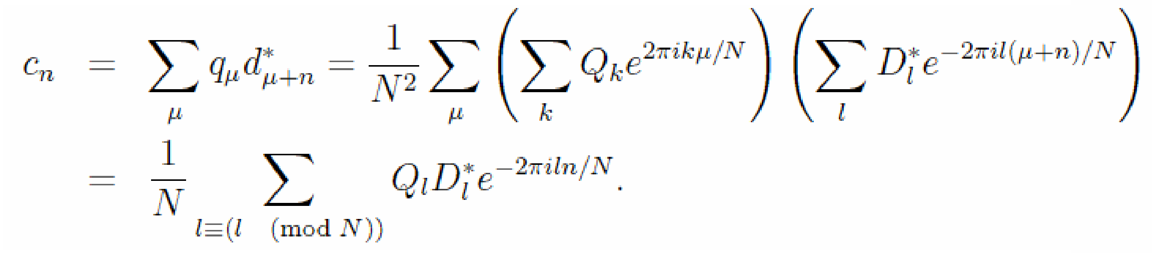


One can also think of every w-mer as an integer in the range from $[1, 4^w]$ for DNA sequences, or $[1, 20^w]$ for protein sequences. This results in an array of size $4^8 = 65536$ for the 8-mers we enumerate in Figure (2.1).

Bit-wise encoding explained



The hashing step which performs the first "scan" step is done by what we consider a brute-force method. The four letters A, C, G, T are each encoded as two bits (A is 00, C is 01, G is 10, and T is 11). Then an array of all possible w-letter words is generated. The number corresponding to each particular word location in the database is recorded: this makes an associative array of words with a list of their locations. This is brute-force in the sense that as a w-length frame moves through the query and database sequences, as many as $w-1$ elements of the last frame position have already been examined. Possible optimizations could take advantage of the redundancy of so many letters as the frame moves along a sequence. While reports from NCBI suggest that this step is not where the majority of the computing time is being spent (most time is spent in the extension step), it is particularly this step which is comparable, if not much faster, to FFT convolution methods. Brute-



$$\begin{aligned} c_n &= \sum_{\mu} q_\mu d^*_{\mu+n} = \frac{1}{N^2}\sum_{\mu}\left(\sum_k Q_k e^{2\pi i k\mu/N}\right)\left(\sum_l D^*_l e^{-2\pi i l(\mu+n)/N}\right) \\ &= \frac{1}{N}\sum_{l\equiv(l \quad (\mathrm{mod}\ N))} Q_l D^*_l e^{-2\pi i l n/N}. \end{aligned}$$

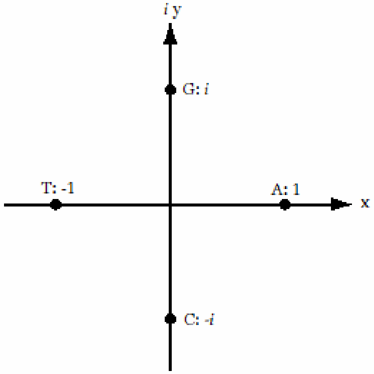


Figure 3.5: **1-Vector complex-plane base encoding**

End