

9 Physical Mapping (Knut Reinert)

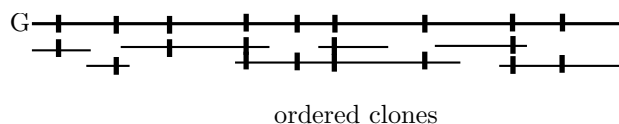
This exposition is based on the following sources, which are all recommended reading:

1. Pevzner, Computational Molecular Biology, MIT Press, 2000, chapter 2,3.
2. Setubal und Meidanis, Introduction to Computational Molecular Biology, PWS Publishing, 1997, chapter 5.
3. Gusfield, Algorithms on Strings, Trees, and Sequences, Cambridge University Press, 1997, chapter 16.
4. Michael S. Waterman, Introduction to computational biology, Chapman and Hall, 1995, chapters 2,3,6.
5. Böckenhauer, Bongartz, Algorithmische Grundlagen der Bioinformatik, Teubner Verlag, 2003, chapter 7

9.1 Physical maps

A *physical map* of a genome G tells us the location of certain *markers* along G . The markers are used for navigation. For example, given a piece of DNA T , if it contains some known markers, then one can use them to locate T in G , thus obtaining the genomic context of T for further exploration.

Since DNA sequences are usually stored in clone libraries, the correct order of the markers also implies an order of the clones.



The markers could be either relatively short nucleotide sequences (ranging from a couple of basepairs to several hundred) or restriction sites. In any case we want to position these markers along the DNA and define:

Definition 1. Let D be a DNA sequence. A physical map consists of a set M of markers and a function $p : M \rightarrow 2^{\mathbb{N}}$ which assigns to each marker in M a position in D .

We can distinguish two different families of methods for constructing a physical map:

1. restriction site mapping
Here we use restriction enzymes to digest the DNA and then use *the lengths of the restriction fragments* to reconstruct the positions of the restriction sites along the sequence.
2. fingerprint mapping
In these techniques one constructs *overlapping* clones. For each clone a *fingerprint* is constructed using restriction enzymes and hybridization experiments. Overlapping clones should have the same (or a very similar) fingerprint. This overlap information is used to order the markers (and the clones).

9.2 Restriction maps

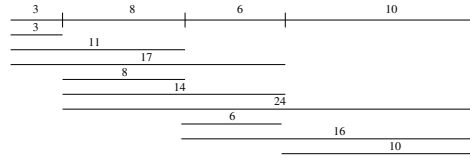
To build a restriction map, different *biochemical* techniques are used to derive information about the map and then *combinatorial* methods are used to reconstruct the map from that data.

The restriction map approach involves first *digesting* the given target sequence with one or more restriction enzyme(s) and then solving a variant of the following problem:

For a set X of points on the line, let $\Delta X = \{|x_1 - x_2| : x_1, x_2 \in X\}$ denote the multiset of all pairwise distances between points in X . In the *restriction mapping problem*, a subset $E \subseteq \Delta X$ (of experimentally obtained fragment lengths) is given and the task is to reconstruct X from E .

9.3 Partial digest problem

For the *partial digest problem* (PDP), the experiment provides data about *all* pairwise distances between restriction sites i.e. $E = \Delta X$



For example the above PDP problem has $\Delta X = \{0, 3, 6, 8, 10, 11, 14, 16, 17, 24\}$

- No polynomial time algorithm for the PDP is known.
- PDP is not a popular mapping method since it is difficult to reliably produce *all* pairwise distances.
- However, S. Skiena devised a simple backtracking algorithm that performs well in practice, although it might still need exponential time.

The input to Skienas algorithm is the multiset $E = \Delta X$ with $\binom{k}{2}$ elements of $\mathbb{N} - \{0\}$. Define $\delta(y, X) = \{ |x - y| \mid x \in X \}$. Then the algorithm proceeds as follows:

Algorithm 2.

```

1 PDP(E)
2  $X = \emptyset$ ; // initially the solution is empty
3  $y_{max} = \max E$ ;
4  $X = X \cup \{y_{max}, 0\}$ ; // this must be in every solution
5  $E = E - y_{max}$ ; // update the distance set
6 if placemax( $X, E$ ) then // compute the rest of X
7     output X;
8     else
9         output no solution;
10 fi

```

Here the code for the procedure placemax:

Algorithm 3.

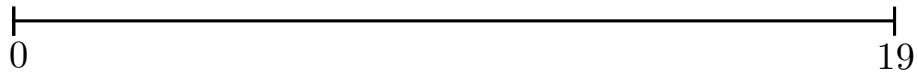
```

1 bool placemax( $X, E$ );
2 if  $E = \emptyset$  then return true; fi
3  $y = \max E$ ;
4 if  $\delta(y, X) \subseteq E$ 
5     then  $E = E - \delta(y, X)$ ;  $X = X \cup \{y\}$ ;
6     if placemax( $X, E$ )
7         then return true;
8         else  $E = E \cup \delta(y, X) - \{0\}$ ;  $X = X - \{y\}$ ;
9     fi
10 fi
11 if  $\delta(y_{max} - y, X) \subseteq E$ 
12     then  $E = E - \delta(y_{max} - y, X)$ ;  $X = X \cup \{y_{max} - y\}$ ;
13     if placemax( $X, E$ )
14         then return true;
15         else  $E = E \cup \delta(y_{max} - y, X) - \{0\}$ ;  $X = X - \{y_{max} - y\}$ ;
16     fi
17 fi
18 return false;

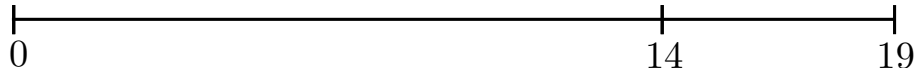
```

Consider the following example of the PDP algorithm: $E = \{1, 2, 3, 4, 5, 5, 7, 7, 9, 9, 10, 10, 12, 14, 19\}$.

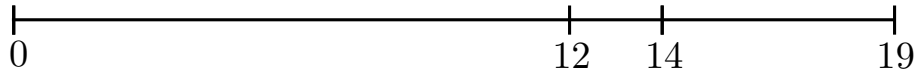
Before the first call of `placemax` we have the following situation:



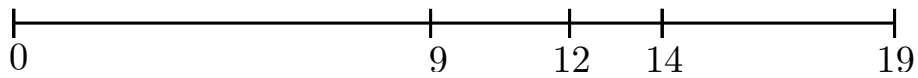
a) $X = \{0, 19\}$, $E = \{1, 2, 3, 4, 5, 5, 7, 7, 9, 9, 10, 10, 12, 14\}$,
 $y = 14$; $\delta(14, X) = \{5, 14\} \subseteq E$. \Rightarrow place 14 at left border.



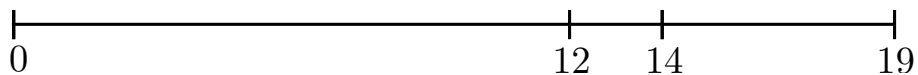
b) $X = \{0, 14, 19\}$, $E = \{1, 2, 3, 4, 5, 7, 7, 9, 9, 10, 10, 12\}$,
 $y = 12$; $\delta(12, X) = \{2, 7, 12\} \subseteq E$. \Rightarrow place 12 at left border.



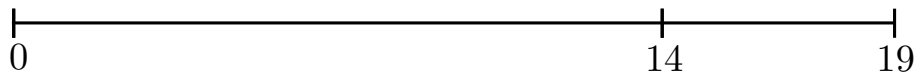
c) $X = \{0, 12, 14, 19\}$, $E = \{1, 3, 4, 5, 7, 9, 9, 10, 10\}$,
 $y = 10$; $\delta(10, X) = \{2, 4, 9, 10\} \not\subseteq E$,
 $\delta(19 - 10, X) = \{3, 5, 9, 10\} \subseteq E$. \Rightarrow place 10 at right border.



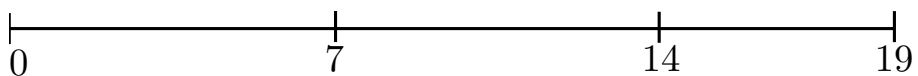
d) $X = \{0, 9, 12, 14, 19\}$, $E = \{1, 4, 7, 9, 10\}$,
 $y = 10$; $\delta(10, X) = \{1, 2, 4, 9, 10\} \not\subseteq E$,
 $\delta(19 - 10, X) = \{0, 3, 5, 9, 10\} \not\subseteq E \Rightarrow$ backtrack step c).



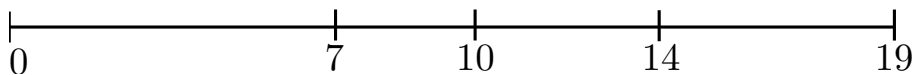
e) $X = \{0, 12, 14, 19\}$, $E = \{1, 3, 4, 5, 7, 9, 9, 10, 10\}$,
 $y = 10$; $\delta(10, X) = \{2, 4, 9, 10\} \not\subseteq E$,
 $\delta(19 - 10, X) = \{3, 5, 9, 10\} \subseteq E$. We already placed 10. \Rightarrow backtrack to b).



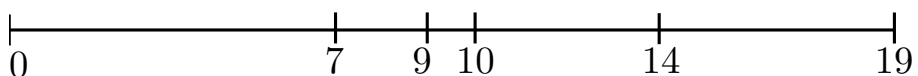
f) $X = \{0, 14, 19\}$, $E = \{1, 2, 3, 4, 5, 7, 7, 9, 9, 10, 10, 12\}$,
 $y = 12$; $\delta(12, X) = \{2, 7, 12\} \subseteq E$. We backtracked the placement of 12 at the left border.
 $\delta(19 - 12, X) = \{7, 7, 12\} \subseteq E$. \Rightarrow place 12 at right border.



g) $X = \{0, 7, 14, 19\}$, $E = \{1, 2, 3, 4, 5, 9, 9, 10, 10\}$,
 $y = 10$; $\delta(10, X) = \{3, 4, 9, 10\} \subseteq E \Rightarrow$ place 10 at left border.



h) $X = \{0, 7, 10, 14, 19\}$, $E = \{1, 2, 5, 9, 10\}$,
 $y = 10$; $\delta(10, X) = \{0, 3, 4, 9, 10\} \not\subseteq E$,
 $\delta(19 - 10, X) = \{1, 2, 5, 9, 10\} \subseteq E \Rightarrow$ place 10 at right border.



Now we are done. $P = \{0, 7, 9, 10, 14, 19\}$ is a feasible solution to the partial digest problem with input $E = \{1, 2, 3, 4, 5, 5, 7, 7, 9, 9, 10, 10, 12, 14, 19\}$. Note that $\bar{P} = \{0, 5, 9, 10, 12, 19\}$, the 'inverse' of P , is also feasible.

The algorithm always places the biggest leftover distance at the left border of the interval $[0..y_{max}]$. Then it

checks whether this was a valid placement. It can locally test for validity by checking whether the induced distances are in the distance set. If not, it tries to place the element at the right border. If this is not possible it backtracks. If no backtracking is possible, no solution exists.

Theorem 4. *Let E be an input to PDP. If a reconstruction of X exists, then the algorithm computes a solution.*

Proof: exercise. Hint: We only check placement at the right and left end of the interval. Argue indirectly that we cannot place a distance in the middle if we always choose the maximal distance.

What is the running time of this algorithm? It is clear that the backtracking will result in the worst case in an exponential running time. More specifically the following holds:

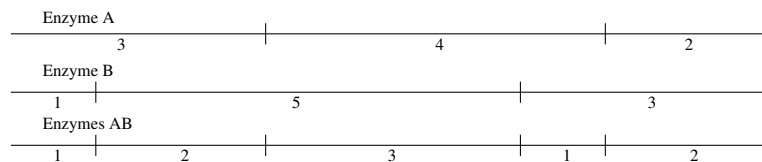
Theorem 5. *Algorithm 2 has a worst case running time of $O(2^k \cdot k \log k)$ for an input of $\binom{k}{2}$ elements.*

Proof: exercise.

9.4 Double digest problem

For the *double digest problem* (DDP), the experiment provides data about the *complete* digest, i.e. *all consecutive* restriction sites for two different restriction enzymes A and B applied alone and in combination yielding ΔA , ΔB and ΔAB . Hence the set of differences E contains not all pairwise distances.

Example:



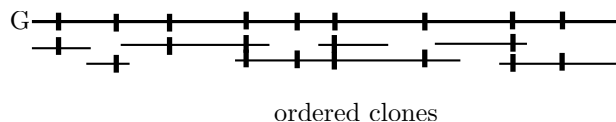
Here we have: $\Delta A = \{2, 3, 4\}$, $\Delta B = \{1, 3, 5\}$ and $\Delta AB = \{1, 1, 2, 2, 3\}$.

Exercise: Can you construct an example for which the solution is not unique?

- DDP is NP-complete.
- All algorithms have problems with more than 10 restriction sites for each enzyme.
- Solution is not unique and number of solutions grows exponentially.
- However, in contrast to the PDP, DDP experiments are easy to conduct.

9.5 Fingerprint mapping

Fingerprint mapping makes use of the fact that in most cases DNA is physically stored in clones that *overlap* to guarantee a complete coverage of the genome. If we place markers along the genome, then the clones that overlap should share the same markers in the region of overlap, that means they have the same *fingerprint*.



Fingerprints could be derived using:

1. Restriction maps of the clones.
2. Restriction fragment sizes. (If a significant fraction of the sizes is the same we assume an overlap).
3. Hybridization experiments. Here we can distinguish between *unique* markers like STS *Sequence Tagged Sites* and *non-unique* markers.

The use of *unique* and *non-unique* markers give rise to different algorithmic problems. We will concentrate on unique probes.

Given a set of unique probes, two protocols are commonly used, *STS content mapping* and *radiation hybrid mapping*.

9.6 STS content mapping

An *STS* is a short (200-500 bp) DNA sequence that occurs exactly once in the given genome. An *EST* (*expressed sequence tag*) is an STS that was derived from a cDNA.

Given a set $P = \{p_1, \dots, p_m\}$ of unique probes (i. e. markers), for example a set of STSs, and given a set of DNA fragments $S = \{S_1, \dots, S_n\}$ sampled from a common genomic region. Let $P(S_i)$ denote the set of probes that are *contained in* (that is, hybridize to) fragment S_i .

Problem:

Find a permutation π of the probe set P such that for every fragment S_i we have

$$P(S_i) = \{p_{\pi(j)}, p_{\pi(j+1)}, \dots, p_{\pi(k-1)}, p_{\pi(k)}\},$$

for some $1 \leq j \leq k \leq m$. That means π has to place the elements of $P(S_i)$ in one contiguous block.

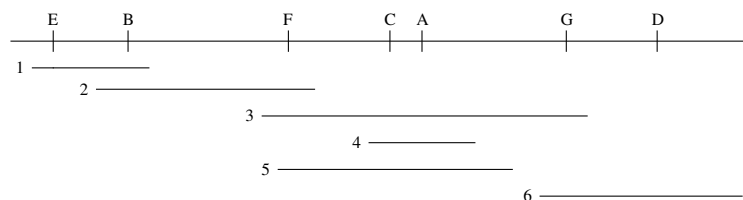
For example, given the following incidence matrix, where an entry in line i and row j is 1, if clone i hybridizes to probe j :

	probe						
clone	A	B	C	D	E	F	G
1	0	1	0	0	1	0	0
2	0	1	0	0	0	1	0
3	1	0	1	0	0	1	1
4	1	0	1	0	0	0	0
5	1	0	1	0	0	1	0
6	0	0	0	1	0	0	1

The probes A, \dots, G can be permuted as follows:

	probe						
clone	E	B	F	C	A	G	D
1	1	1	0	0	0	0	0
2	0	1	1	0	0	0	0
3	0	0	1	1	1	1	0
4	0	0	0	1	1	0	0
5	0	0	1	1	1	0	0
6	0	0	0	0	0	1	1

This implies the following layout:



Now all probes are consecutive for each clone and we say that the matrix has the *consecutive ones property*. The solution(s) can be computed in linear time and represented in a data structure called a *PQ-tree*.

Not only have we thus ordered all clones, but we have also determined an ordering of the probes (STSs).

Unfortunately, the hybridization experiments are very error-prone, usually suffering from:

- *false positives*: reporting that a clone contains a specific probe, when in fact it does not,

- *false negatives*: reporting that a clone does not contain a specific probe, when in fact it does, and
- *chimeras*: these are false clones built from different pieces of DNA that come from unrelated and distant parts of the genome and thus falsely bring together distant probes.

The following matrix depicts a correctly ordered probe set with a false negative in clone 3, a false positive in clone 1, and a possible chimeric clone 6.

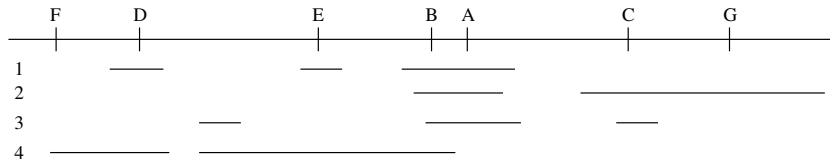
clone	probe						
	E	B	F	C	A	G	D
1	1	1	0	0	1	0	0
2	0	1	1	0	0	0	0
3	0	0	1	0	1	1	0
4	0	0	0	1	1	0	0
5	0	0	1	1	1	0	0
6	1	0	0	0	0	1	1

Before we discuss how to handle such errors, we describe the second method, radiation hybrid mapping, since it yields similar data.

9.7 Radiation hybrid mapping

In *radiation hybrid mapping*, a target (e.g. human) chromosome is irradiated and broken into a small number of fragments. These non-overlapping fragments are fused into a e.g. hamster cell and then replication produces a cell line. Subsequently, each cell line contains a pool of 5 – 10 large, disconnected, non-overlapping fragments of target DNA.

This is repeated several times using different random irradiation results.



Finally, it is determined which cell lines hybridizes to which probes. This is very similar to STS-content mapping, except that we do not know how many fragments a cell line contains or to which fragment a given probe actually hybridizes to.

The following matrix show the data from the above depicted radiation hybrid experiment:

cell line	probe						
	E	B	F	C	A	G	D
1	1	1	0	0	1	0	1
2	0	1	0	1	1	1	0
3	0	1	0	1	1	0	0
4	1	1	1	0	0	0	1

What is a sensible objective function to find the correct permutation of probes?

We can assume that probes that lie close to each other in the target genome are more likely to be contained in the same fragment (within a pool). Thus, we aim to minimize the total number of blocks of consecutive ones.

This problem is NP-hard!

9.8 TSP solution for consecutive ones with gaps

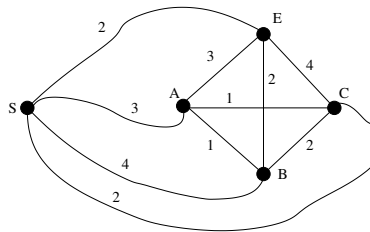
We reduce the problem of finding the probe permutation with the minimum number of consecutive ones to the traveling salesman problem as follows:

1. Define a weighted graph $G = (V, E)$ with $V = \{s, p_1, \dots, p_k\}$ where p_i is a node for each probe i and s is a special node.
2. E contains an edge from s to each p_i and an edge for each pair of probes.
3. The weight of the edges from s to the p_i is the number of ones in the corresponding column of the matrix.
4. The weight of any other edge (p_i, p_j) is the *Hamming distance* between the columns corresponding to p_i and p_j .

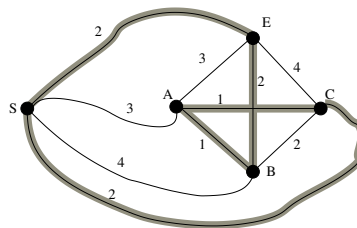
For the sake of exposition we look at a submatrix of the above example:

c/p	E	B	C	A
1	1	1	0	1
2	0	1	1	1
3	0	1	1	1
4	1	1	0	0

This translates into the following graph G :



An optimal tour is s, C, A, B, E :



c/p	C	A	B	E
1	0	1	1	1
2	1	1	1	0
3	1	1	1	0
4	0	0	1	1

This tour indeed gives the correct ordering and the blocks of ones happen to be gap-free.

Theorem 6. *The TSP tour of weight w corresponds to a probe permutation with exactly $\frac{w}{2}$ blocks of consecutive ones.*

Proof: Each TSP tour corresponds to a probe permutation. Except for the edges incident to s , a tour is charged the Hamming distance if it traverses edge (p_i, p_j) . For the combination $(0, 1)$ it is charged for the *beginning* of a new block induced by ordering p_i before p_j , and for the combination $(1, 0)$ it is charged for the *ending* of a block. Hence each block is charged 1 for its begin and end. The weights from s to each node counts the number of blocks ending in the rightmost resp. starting in the leftmost column. □

9.9 Summary

Physical mapping comes in two flavors:

1. Restriction mapping. Here restriction enzymes are used to digest the target into smaller pieces. Using the partial or double digest protocol certain sets of distances between restriction sites are constructed. The goal is to explain all distances. Restriction mapping is also used to determine whether two clones overlap (by using the restriction map as a fingerprint).
2. Fingerprint mapping. The goal is here to determine the order of *overlapping* clones. Fingerprints are constructed using hybridization experiments or restriction enzyme information.

There are different protocols to determine a map, each suitable in different situations. Each protocol has its associated algorithmic problem. Most of them are in exact form already NP-hard. Errors need to be taken into account.

You should know:

- The partial digest problem
- The double digest problem
- STS content mapping
- Radiation Hybrid mapping

We talked about a solution to the PDP problem using a backtracking algorithm and about solving the STS content mapping and radiation hybrid mapping by reducing it to the TSP problem.

10 Sequence Assembly (Daniel Huson)

The exposition is based on the following sources, which are all recommended reading:

1. Michael S. Waterman, Introduction to computational biology, Chapman and Hall, 1995. (Chapter 7)
2. Eugene W. (Gene) Myers *et al.*, A Whole-Genome Assembly of *Drosophila*, Science, 287:2196-2204, 24 March 2000.
3. Venter *et al.*, The sequence of the Human Genome, Science, 291:1304-1351, 16 February 2001.
4. Daniel Huson, Knut Reinert and Eugene Myers, The Greedy-Path Merging Algorithm for Sequence Assembly, RECOMB 2001, 157-163, 2001.

10.1 Genome Sequencing

Using a method that was basically invented in 1980 by Sanger, current sequencing technology can only determine 500 – 1000 consecutive base pairs of DNA in any one read. To sequence a larger piece of DNA, *shotgun sequencing* is used.

Originally, shotgun sequencing was applied to small viral genomes and to 30 – 40kb segments of larger genomes.

In 1994, the 1.8Mb genome of the bacteria *H.influenzae* was assembled from shotgun data.

At the beginning of 2000, an assembly of the 130Mb *Drosophila* genome was published.

At the beginning of 2001, two initial assemblies of the human genome were published.

10.2 The big picture – From molecule to sequence

Whole Genome Shotgun sequencing	illustration	Clone by clone sequencing
Source sequence (target) (\approx 3000 Mbp for human)	<pre>ACGTTGCAC TAGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA</pre>	Source sequence (target)
Not done in WGS	<pre> ACGTTGCAC TAGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA GACTACGACTACGACTCAGCA AGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA CGCTATATCGACTACGACTACGACTCAGCA ACGTTGCAC TAGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA RCBTTGC ACTAGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA CACTAGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA TACGACTACGACTACGACTCAGCA </pre>	is broken into smaller pieces (150–1000kbp)
Not done in WGS	<pre> ACGTTGCAC TAGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA ACGTTGCAC TAGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA CACTAGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA AGCACAGCGCGCTATATCGACTACGACTACGACTCAGCA GACTACGACTACGACTACGACTCAGCA </pre>	Big pieces are selected to tile the target (minimum tiling least costly but most difficult) \Rightarrow Physical mapping

Big source sequence is copied many times...

```
ACGTTGCAGTACGACAGGCGCTATATCGACTACGACTACGACTCAGCA
ACGTTGCAGTACGACAGGCGCTATATCGACTACGACTACGACTCAGCA
ACGTTGCAGTACGACAGGCGCTATATCGACTACGACTACGACTCAGCA
ACGTTGCAGTACGACAGGCGCTATATCGACTACGACTACGACTCAGCA
ACGTTGCAGTACGACAGGCGCTATATCGACTACGACTACGACTCAGCA
ACGTTGCAGTACGACAGGCGCTATATCGACTACGACTACGACTCAGCA
ACGTTGCAGTACGACAGGCGCTATATCGACTACGACTACGACTCAGCA
ACGTTGCAGTACGACAGGCGCTATATCGACTACGACTACGACTCAGCA
```

all source sequences are copied many times (e.g. 40000 for human)

and randomly broken into fragments, e.g. using *sonication* or *nebulation*, ...

```
AGCGGCTATATCGACTACG ACGACTCAGC ACTAGCACAGCGCGA
CGCTATATCGACTACGACG CGCTATATCGACTACGACG TTTTTT
ACGTTGCAGTACGACAGGCGCT CGCTATATCGACTACGACG TGGTG
TAGACTACGACTACGACG
ACTAGCACAGCGCGA AA ACTAGCACAGCGCGA ACGACTCAGC
TGCAGTACGACAGGCGCTATATCGACT ACCTATATCGACTACGACG
ACGACTCAGC ACGTTGCAGTACGACAGGCGCT
TAGACTACGACTACGACG ACGC TAGACTACGACTACGACG
```

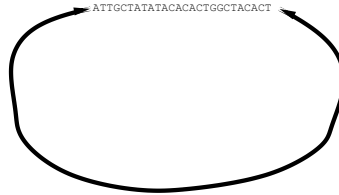
each sequence is randomly broken into fragments

that are then size selected, size e.g. 2kb, 10kb, 50kb or 150kb, ...

```
ACCGTGTCACACACGGTAGCAGCAGCAGCAGCAGCAGC
TGTTGTGCTGCTGTATATACACTGGCTACACT
ACCGTGTCACACACGGTAGCAGCAGCAGCAGCAGCAGC
TGTTGTGCTGCTGTATATACACTGGCTACACT
TGTCACACACGGTAGCAGCAGCAGCAGCAGCAGCAGC
ACCGTGTCACACACGGTAGCAGCAGCAGCAGCAGCAGC
ATTGTTATATACACTGGCTACACT
ACCGGAGCAGCAGCAGCAGCAGCAGCAGC
ATTGCTATATACACTGGCTACACT
ATATATACACTGGCTACACT
AGCAGCAGCAGCAGCAGCAGCAGCAGC
TATACACTGGCTACACT
ATTGTTGTGCTGCTG
ACTGGCTACACT
TATACACTACT
ATTGCTATATACACTGGCTACACT
```

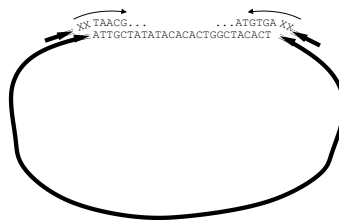
that are then size selected

and inserted into cloning vectors.



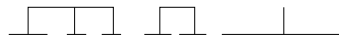
and all inserted into cloning vectors.

In *double barrel shotgun sequencing*, each clone is sequenced from both ends, to obtain a *mate-pair* of reads, each read of average length 550 with $\approx 1\%$ error.



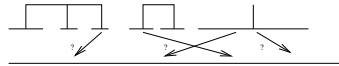
first approaches did not use double barrel, later they did.

Result of assembly is a collection of *scaffolds* for the *whole genome*.



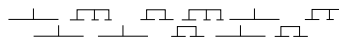
Each clone is a collection of scaffolds.

Ordering is quite difficult, since small pieces are hard to *map* back to the genomic axis



Local ordering is relatively easy.

Not done in WGS



The sequence of *all* clones has to be assembled according to the physical map and sequence overlaps. Due to repeats and assembly errors this is hard.

10.3 Shotgun sequencing data

Given an unknown DNA sequence $a = a_1a_2 \dots a_L$.

Shotgun sequencing of a produces a set of reads

$$\mathcal{F} = \{f_1, f_2, \dots, f_R\},$$

of average length 550 (at present).

Essential characteristics of the data:

- Incomplete coverage of the source sequences.
- Sequencing introduces errors at a rate of about 1% for the first 500 bases, if carefully performed.

- The reads are sampled from both strands of the source sequence and thus the orientation of any given read is unknown.

10.4 The fragment assembly problem

The input is a collection of reads (or *fragments*) $\mathcal{F} = \{f_1, f_2, \dots, f_R\}$, that are sequences over the alphabet $\Sigma = \{A, C, G, T\}$.

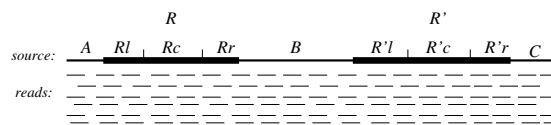
An ϵ -*layout* of \mathcal{F} is a string S over Σ and a collection of R pairs of integers $(s_j, e_j)_{j \in \{1, 2, \dots, R\}}$, such that

- if $s_j < e_j$ then f_j can be aligned to the substring $S[s_j, e_j]$ with less than $\epsilon \cdot |f_j|$ differences, and
- if $s_j > e_j$ then f_j can be aligned to the substring $\overline{S[e_j, s_j]}$ with less than $\epsilon \cdot |f_j|$ differences, then
- $\cup_{j=1}^R [\min(s_j, e_j), \max(s_j, e_j)] = [1, |S|]$.

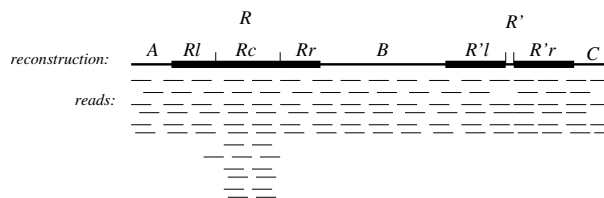
The string S is the reconstructed source string. The integer pairs indicate where the reads are placed and the order of s_i and e_i indicate the orientation of the read f_i , i.e. whether f_i was sampled from S or its complement \overline{S} .

The set of all ϵ -layouts models the set of all possible solutions. There are many such solutions and so we want a solution that is in some sense *best*. Traditionally, this has been phrased as the *Shortest Common Superstring Problem (SCS)* of the reads within error rate ϵ . Unfortunately, the SCS Problem often produces overcompressed results.

Consider the following source sequence that contains two instances R, R' of a high fidelity repeat and three stretches of unique sequence A, B and C :



The shortest answer isn't always the best and the interior part $R_c \approx R'_c$ of the repeat region is *overcompressed*:



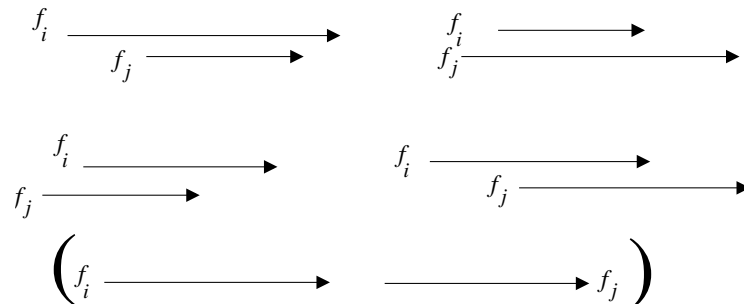
10.5 Sequence assembly in three stages

Traditional approaches to sequence assembly divides the problem into three phases:

1. In the *overlap* phase, every read is compared with every other read, and the overlap graph is computed.
2. In the *layout* phase, the pairs (s_j, e_j) are determined that position every read in the assembly.
3. In the *consensus* phase, a multialignment of all the placed reads is produced to obtain the final sequence.

10.6 The overlap phase

For a read f_i , we must calculate how it overlaps any other read f_j (or its reverse complement, $\overline{f_j}$). Holding f_i fixed in orientation, f_i and f_j can overlap in the following ways:



The number of possible relationships doubles, when we also consider $\overline{f_j}$.

The overlap phase is the computational bottleneck in large assembly projects. For example, assembling all 27 million human reads produced at Celera requires

$$2 \cdot \binom{27000000}{2} \approx 1458000000000000 \approx 1.5 \cdot 10^{15}$$

comparisons.

For any two reads a and b (and either orientation of the latter), one searches for the overlap alignment with the highest alignment score, based on a similarity score $s(a, b)$ on Σ and an indel penalty $g(k) = k\delta$.

Let $S(a, b)$ be the maximum score over all alignments of two reads $a = a_1a_2 \dots a_m$ and $b = b_1b_2 \dots b_n$, we want to compute:

$$A(|a|, |b|) = \max \left\{ S(a_k, a_{k+1} \dots a_i, b_l, b_{l+1} \dots b_j) \mid \left\{ \begin{array}{l} 1 \leq k \leq i \leq m, \\ 1 \leq l \leq j \leq n, \\ \text{and } i = m \text{ or } j = n \text{ holds} \end{array} \right\} \right\}.$$

10.7 Overlap alignment

This is a standard pairwise alignment problem (similar to local alignment, except we don't have a 0 in the recursion) and we can use dynamic programming to compute:

$$A(i, j) = \max\{S(a_k, a_{k+1} \dots a_i, b_l, b_{l+1} \dots b_j) \mid 1 \leq k \leq i \text{ and } 1 \leq l \leq j\}.$$

Algorithm (Overlap alignment)

Input: $a = a_1a_2 \dots a_n$ and $b = b_1b_2 \dots b_m$, $s(\cdot, \cdot)$ and δ

Output: $A(i, j)$

begin

$A(0, j) = A(i, 0) \leftarrow 0$ for $i = 1, \dots, n, j = 1, \dots, m$

for $i = 1, \dots, n$:

for $j = 1, \dots, m$:

$$A(i, j) \leftarrow \max \left\{ \begin{array}{l} A(i-1, j) - \delta, \\ A(i, j-1) - \delta, \\ A(i-1, j-1) + s(a_i, b_j) \end{array} \right\}$$

end

Runtime is $O(nm)$.

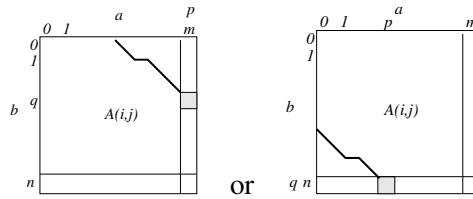
Given two reads $a = a_1a_2 \dots a_m$ and $b = b_1b_2 \dots b_n$. For the matrix $A(i, j)$ computed as above, set

$$(p, q) := \arg \max\{A(i, j) \mid i = m \text{ or } j = n\}.$$

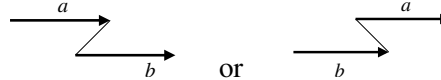
There are two cases:

$$p = m \quad \text{or} \quad q = n$$

The trace-back paths look like this:



The alignments look like this:



10.8 Faster overlap detection

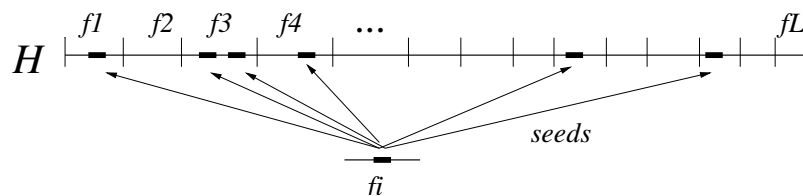
Dynamic programming is too slow for large sequencing projects. Indeed, it is wasteful, as in assembly, only high scoring overlaps with more than e.g. 96% identity, play a role.

One can use a *seed and extend* approach (as used in BLAST):

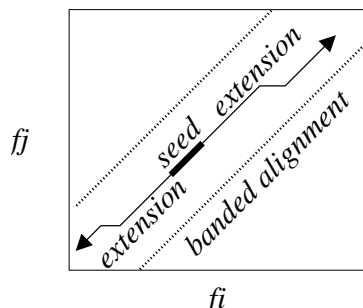
1. Produce the concatenation of all input reads $H = f_1 f_2 \dots f_L$.
2. For each read $f_i \in \mathcal{F}$: Find all *seeds*, i.e. exact matches between k -mers of f_i and the concatenated sequence H . (Merge neighboring seeds.)
3. Compute *extensions*: Attempt to extend each (merged) seed to a high scoring overlap alignment between f_i and the corresponding read f_j .

(A k -mer is a string of length k . In this context, $k = 18 \dots 22$)

Computation of seeds:



Extension of seeds using *banded* dynamic programming (running time is linear in the read length):

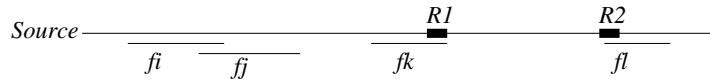


10.9 True and repeat-induced overlaps

Assume that we have found a high quality overlap o between f_i and f_j . There there are three possible cases:

- The overlap o corresponds to an overlap of f_i and f_j in the source sequence. In this case we call o a *true* overlap.

- The reads f_i and f_j come from different parts of the source sequence and their overlapping portions are contained in different instances of the same repeat, this is called a *repeat-induced* overlap.
- The overlap exists by chance. To avoid short random overlaps, one requires that an overlap is at least 40bp long.



True overlap between f_i and f_j , repeat induced overlap between f_k and f_l .

10.10 Avoiding repeat-induced overlaps

To avoid the computation of repeat-induced overlaps, one strategy is to only consider seeds in the seed-and-extend computation whose k -mers are not contained inside a repeat. In this way we can ensure that any computed overlap has a significant unique part.

There are two strategies for this:

- *Screening known repeats*: Each read is aligned against a database of known repeats, i.e. using Repeat-masker. Portions of reads that match a known repeat are labeled *repetitive*.
- *De novo screening*: For each k -mer contained in H , the concatenation of reads, we determine how many times it occurs in H and then label those k -mers as *repetitive*, whose number of occurrences is unexpectedly high.

10.11 Celera’s overlapper

The assembler developed at Celera Genomics employs an overlapper than compares up to 32 million pairs of reads per second.

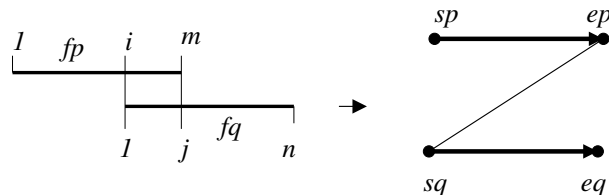
Overlapping all pairs of 27 million reads of human DNA using this program takes about 10 days, running on about 10-20 four processor machines (Compaq ES40), each with 4GB of main memory.

The input data file is about 50GB. To parallelize the overlap compute, each job grabs as many reads as will fit into 4GB of memory (minus the memory necessary for doing the computation) and then streams all 27 million reads against the ones in memory.

10.12 The overlap graph

The overlap phase produces an *overlap graph* OG , defined as follows: Each read $f_p \in \mathcal{F}$ is represented by a directed edge (s_p, e_p) from node s_p to e_p , representing the start and end of f_p , respectively. The *length* of such a *read edge* is simply the length of the corresponding read.

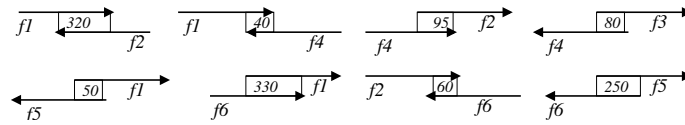
An overlap between $f_p = f_{p1}f_{p2} \dots f_{pm}$ and $f_q = f_{q1}f_{q2} \dots f_{qn}$ gives rise to an undirected *overlap edge* e between s_p , or e_p , and s_q , or e_q , depending on the orientation of the overlap, e.g.:



The label (or “length”) of the overlap edge e is defined to be -1 times the overlap length, e.g. $-(\frac{m-i+j-1}{2} + 1)$ in the figure.

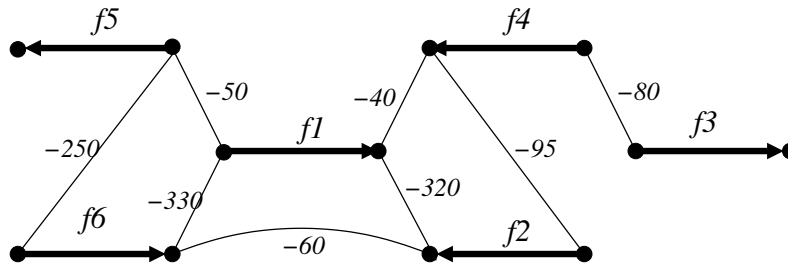
10.13 Example

Assume we are given 6 reads $\mathcal{F} = \{f_1, f_2, \dots, f_6\}$, each of length 500, together with the following overlaps:



Here, for example, the last 320 bases of read f_1 align to the first 320 bases of the reverse complement $\overline{f_2}$ of f_2 , whereas f_1 and $\overline{f_5}$ overlap in the first 50 bases of each.

We obtain the following overlap graph OG :

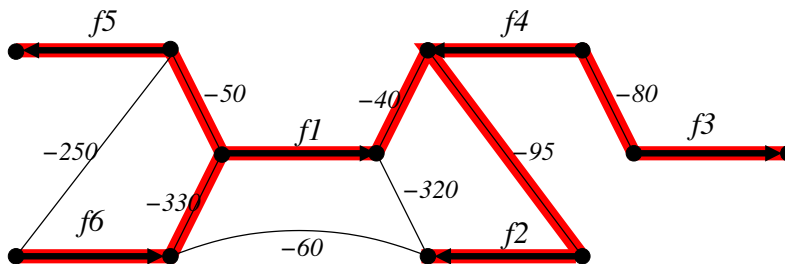


Each read f_p is represented by a read edge (s_p, e_p) of length $|f_p|$. Overlaps off the start s_p , or end e_p , of f_p are represented by overlap edges starting at the node s_p , or e_p , respectively. Each overlap edge is labeled by -1 times the overlap length.

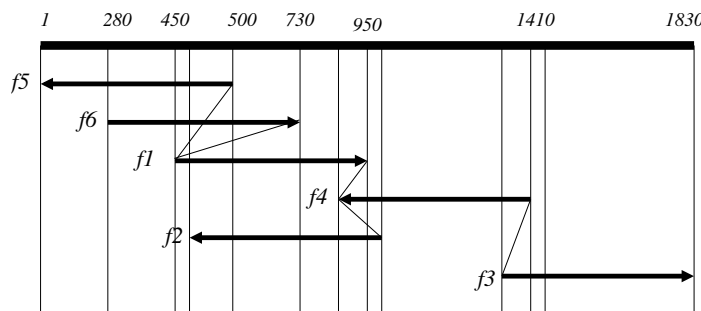
10.14 The layout phase

The goal of the layout phase is to arrange all reads into an approximate multi-alignment. This involves assigning coordinates to all nodes of the overlap graph OG , and thus, determining the value of s_i and e_i for each read f_i .

A simple heuristic is to select a *spanning forest* of the overlap graph OG that contains all read edges. (A spanning forest is a set F of edges such that any two nodes in the same connected component of OG are connected by a unique simple, unoriented path of edges in F .)



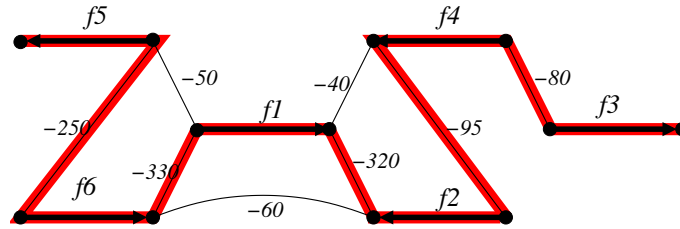
such a subset of edges positions every read with respect to every other, within a given connected component of the graph:



Such a putative alignment of reads is called a *contig*.

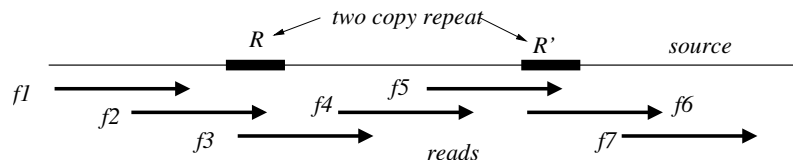
The spanning tree is usually constructed using a *greedy heuristic* in which the overlap edges are chosen in

decreasing overlap length (i.e., increasing edge “length”).

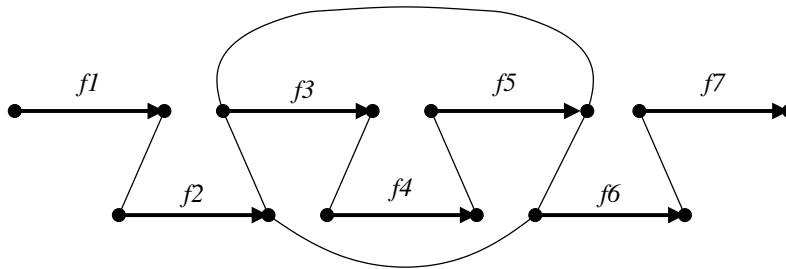


10.15 Repeats and the layout phase

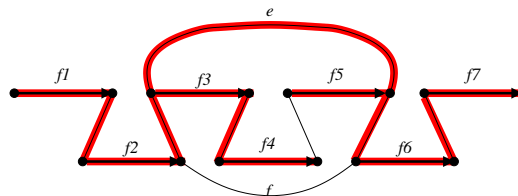
Consider the following situation:



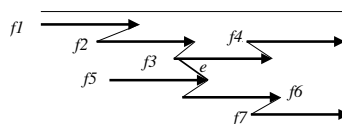
This gives rise to the following overlap graph:



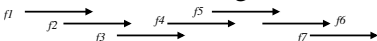
Consider this spanning tree:



A layout produced using the edge *e* or *f* does not reflect the true ordering of the reads and the obtained contig is called *misassembled*:



However, avoiding the repeat-induced edges *e* and *f*, one obtains a correct layout:



Note that *neither* of the two layouts is “consistent” with all overlap edges in the graph.

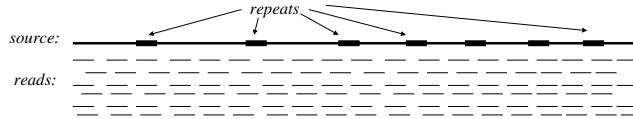
10.16 Unitigging

The main difficulty in the layout phase is that we can’t distinguish between true overlaps and repeat-induced overlaps. The latter produce “inconsistent” layouts in which the coordinate assignment induces overlaps that are not reflected in the overlap graph (e.g., reads *f4* and *f7* in the example above).

Thus, the layout phase proceeds in two stages:

1. *Unitigging*: First, all uniquely assemblable contigs are produced, as just described. These are called *unitigs*.
2. *Repeat resolution*: Then, at a later stage, one attempts to reconstruct the repetitive sequence that lies between such unitigs.

Reads are sampled from a source sequence that contains repeats:

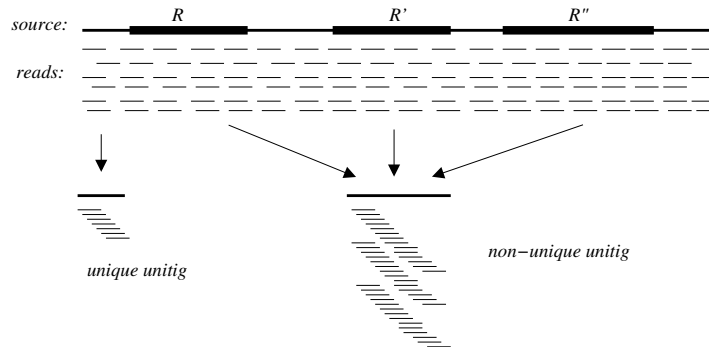


Reads that form consistent chains in the overlap graph are assembled into unitigs and the remaining “repetitive” reads are processed later:



10.17 Unique unitigs

As defined above, a “unitig” is obtained as a chain of consistently overlapping reads. However, a unitig does not necessarily represent a segment of unique source sequence. For example, its reads may come from the interior of different instances of a long (many copy) repeat:



Non-unique unitigs can be detected by virtue of the fact that they contain significantly more reads than expected.

10.18 Identifying unique unitigs

Under assumption that the sampling of reads from the target is done uniformly, the *arrival* of the fragments start positions mapped along the target sequence should have constant, low probability. Hence we can model this process using a Poisson distribution.

Let R be the number of reads and G the estimated length of the source sequence. We then expect a on average $\frac{R}{G}$ arrivals of fragments per base.

For a unitig with k reads and approximate length ρ , the probability of seeing the $k - 1$ start positions in the interval of length ρ is

$$\frac{e^{-c} c^k}{k!},$$

with $c := \frac{\rho R}{G}$, if the unitig is not oversampled, and

$$\frac{e^{-2c} (2c)^k}{k!},$$

if the unitig is the result of collapsing two repeats.

(see Mike Waterman's book, page 148, for details)

The *arrival statistic* used to identify *unique* unitigs is the (natural) log of the ratio of these two probabilities,

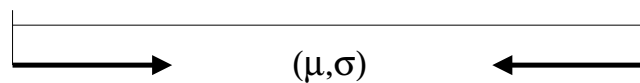
$$c - (\log 2)k.$$

A unitig is called *unique*, if it's arrival statistic has a positive value of 10 or more.

10.19 Mate pairs

Fragment assembly of reads produces contigs, whose relative placement and orientation with respect to each other is unknown.

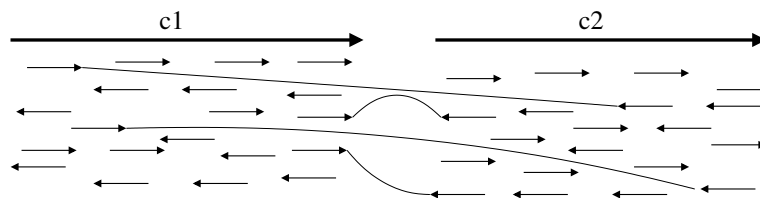
Recall that modern shotgun sequencing protocols employ a so-called *double barreled* shotgun. That is, longer clones of a given fixed length are sequenced from both ends and one obtains a pair of reads, a *mate pair*, whose relative orientation and mean μ (and standard deviation σ) length are known:



Typical clone lengths are $\mu = 2kb, 10kb, 50kb$ or $150kb$. For clean data, $\sigma \approx 10\%$ of μ . Mate pair mismatching is a problem and can effect 10 – 30% of all pairs.

10.20 Scaffolding

Consider two reconstructed contigs. If they correspond to neighboring regions in the source sequence, then we can expect to see mate pairs to span the gap between them:



Such mate pairs determine the relative orientation of both contigs, and we can compute a mean and standard deviation for the gap between them. In this case, the contigs are said to be *scaffolded*:



10.21 Determining the distance between two contigs

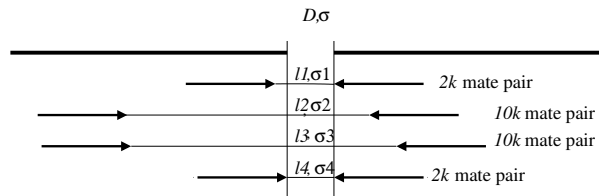
Given two contigs c_1 and c_2 connected by mate pairs m_1, m_2, \dots, m_k . Each mate pair gives an estimation of the distance between the two contigs.

These estimations can viewed as independent measurements $(l_1, \sigma_1), (l_2, \sigma_2), \dots, (l_k, \sigma_k)$ of the distance D between the two contigs c_1 and c_2 . Following standard statistical practice, they can be combined as follows:

Define $p := \sum \frac{l_i}{\sigma_i^2}$ and $q = \sum \frac{1}{\sigma_i^2}$. We set the distance between c_1 and c_2 to

$$D := \frac{p}{q}, \text{ with standard deviation } \sigma := \frac{1}{\sqrt{q}}.$$

Here is an example:



It is possible that the mate pairs between two contigs c_1 and c_2 lead to significantly different estimations of the distance from c_1 and c_2 . In practice, only mate pairs that *confirm* each other, i.e. whose estimations are within 3σ of each other are considered together in a gap estimation.

10.22 The significance of mate pairs

Given two contigs c_1 and c_2 . If there is only one mate pair between the two contigs, then due to the high error rates associated with mate pairs, this is not significant.

If, however, c_1 and c_2 are *unique unitigs*, and if there exist two different mate pairs between the two that give rise to the same relative orientation and similar estimations of the gap size between c_1 and c_2 , then this the scaffolding of c_1 and c_2 is highly reliable.

This is because that probability that two false mate pairs occur that confirm each other, is extremely small.

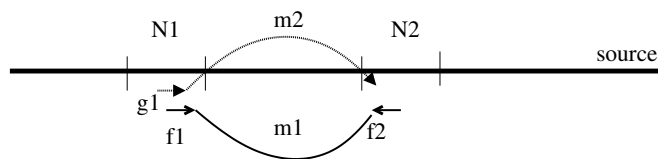
10.23 Example

Let the sequence length be $G = 120MB$, for example (Drosophila). For simplicity, assume we have 5-fold coverage of mate pairs, with a mean length of $\mu = 10kb$ and standard deviation of $\sigma = 1kb$.

Consider a false mate pair $m_1 = (f_1, f_2)$ with reads f_1 and f_2 . Let N_1 and N_2 denote the two intervals (in the source sequence) of length 3σ centered at the starts of f_1 and f_2 , respectively. Both have length $6kb$.

Consider a second false mate $m_2 = (g_1, g_2)$ with g_1 inside N_1 . The probability that g_2 lies in N_2 is roughly

$$\frac{6kb}{120MB} = \frac{1}{20000}.$$



Assume that the reads have length 600. Assume that 10% of all mate pairs are false. At 5-fold coverage, the interval N_1 is covered by about $5 \cdot \frac{6000}{600} = 50$ reads, of which ≈ 5 have false mates.

Hence, the probability that m_1 is confirmed by some second false mate pair m_2 is

$$\approx 5 \cdot \frac{1}{20000} = \frac{1}{4000} = 0.00025.$$

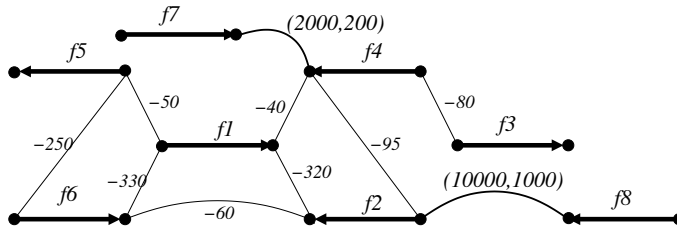
This does not take into account that N_1 certainly contains many reads with correct mate pairs.

10.24 The overlap-mate graph

Given a set of reads $\mathcal{F} = \{f_1, f_2, \dots, f_R\}$ and let G denote the overlap graph associated with \mathcal{F} .

Given one set (or more) $M_{\mu,\sigma} = \{m_1, \dots, m_k\}$ of mate pairs $m_k = (f_i, f_j)$, with mean μ and standard deviation σ .

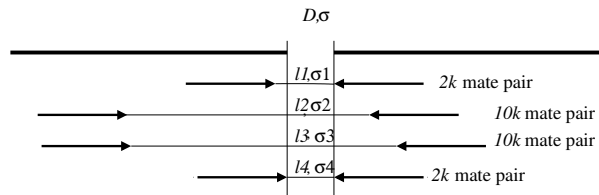
Let f_i and f_j be two mated reads represented by the edges (s_i, e_i) and (s_j, e_j) in G . We add an undirected *mate* edge between e_i and e_j , labeled (μ, σ) , to indicate that f_i and f_j are mates and thus obtain the *overlap-mate graph*:



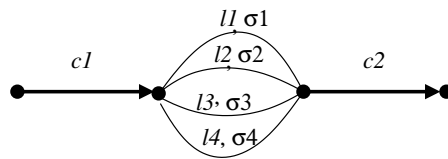
10.25 The contig-mate graph

Given a set of \mathcal{F} of fragments and a set of assembled contigs $\mathcal{C} = \{c_1, c_2, \dots, c_t\}$. A more useful graph is obtained as follows:

Represent each assembled contig c_i by a *contig edge* with nodes s_i and e_i . Then, add *mate edges* between such nodes to indicate that the corresponding contigs contain fragments that are mates:



Leads to:



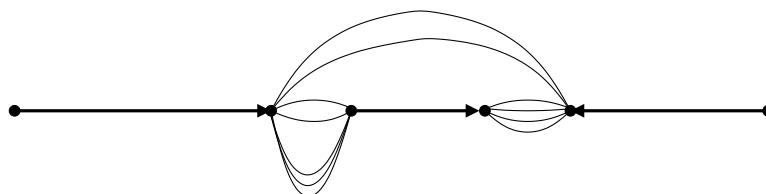
10.26 Edge bundling

Consider two contigs c_1 and c_2 , joined by mate pair edges m_1, \dots, m_k between node e_1 and s_2 . Every maximal subset of mutually confirming mate edges is replaced by a single *bundled mate edge* e , whose mean length μ and standard deviation σ are computed as discussed above. Any such bundled edge is labeled (μ, σ) .

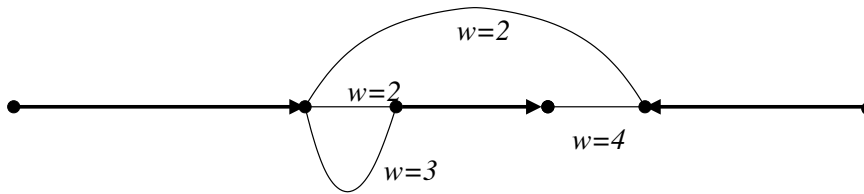
(A heuristic used to compute these subsets is to repeatedly bundle the median-length simple mate edge with all mate edges within three standard deviations of it, until all simple mate edges have been bundled.)

Additionally, we set the weight $w(e)$ of any mate edge to 1, if it is a simple mate edge, and to $\sum_{i=1}^k w(e_i)$, if it was obtained by bundling edges e_1, \dots, e_k .

Consider the following graph:

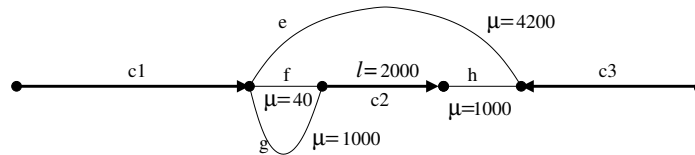


Assuming that mate edges drawn together have similar lengths and large enough standard deviation, edge bundling will produce the following graph:

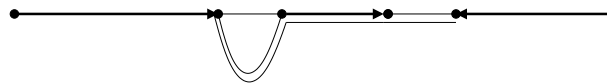


10.27 Transitive edge reduction

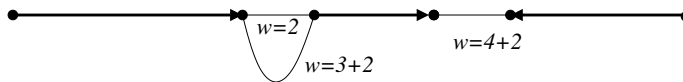
Consider the previous graph with some specific edge lengths:



The mate edge e gives rise to estimation of the distance from the right node of contig c_1 to the left node of c_3 that is similar to the one obtained by following the path $P=(g, c_2, h)$. Based on this *transitivity* property we can *reduce* the edge e on to the path p :



to obtain:



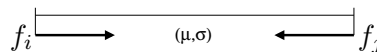
Consider two nodes v and w that are connected by an alternating path $P = (m_1, b_1, m_2, \dots, m_k)$ of mate-edges (m_1, m_2, \dots) and contig edges (c_1, c_2, \dots) from v to w , beginning and ending with a mate-edge. We obtain a mean length and standard deviation for P by setting $l(P) := \sum_{m_i} \mu(m_i) + \sum_{c_i} l(c_i)$ and $\sigma(P) := \sqrt{\sum_{m_i} \sigma(m_i)^2}$.

We say that a mate-edge e from v to w can be *transitively reduced* on to the path P , if e and P approximately have the same length, i. e., if $|\mu(e) - l(P)| \leq C \cdot \max\{\sigma(e), \sigma(P)\}$ for some constant C , typically 3. If this is the case, then we can *reduce* e by removing e from the graph and incrementing the weight of every mate-edge m_i in P by $w(e)$.

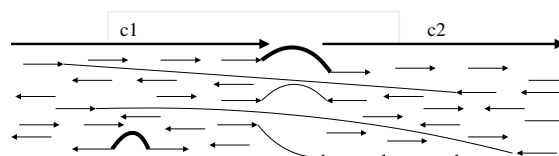
In the following, we will assume that any contig-mate graph considered has been edge-bundled and perhaps also transitively reduced to some degree.

10.28 Happy mate pairs

Consider a mate pair m of two reads f_i and f_j , obtained from a clone of mean length μ and standard deviation σ :



Assume that f_i and f_j are contained in the same contig or scaffold of an assembly. We call m *happy*, if f_i and f_j have the correct relative orientation (i.e., are facing each other) and are at approximately the right distance, i.e., $|\mu - |s_i - s_j|| \leq 3\sigma$. Otherwise, m is *unhappy*. Two unhappy mates are highlighted here:



10.29 Ordering and orientation of the contig-mate graph

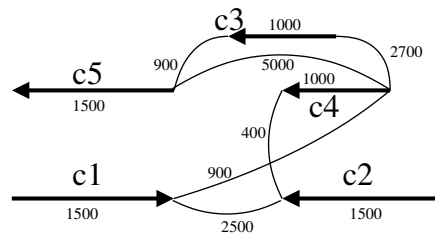
Given a collection of contigs $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$ constructed from a set of reads $\mathcal{F} = \{f_1, f_2, \dots, f_R\}$, together with the corresponding mate pair information M . Let $G = (V, E)$ denote the associated contig-mate graph.

An *ordering (and orientation)* of G (or \mathcal{C}) is a map $\phi : V \rightarrow \mathbb{N}$ such that $|\phi(b_i) - \phi(e_i)| = l(c_i)$ for all contigs $c_i \in \mathcal{C}$, in other words, an assignment of coordinates to all nodes that preserves contig lengths.

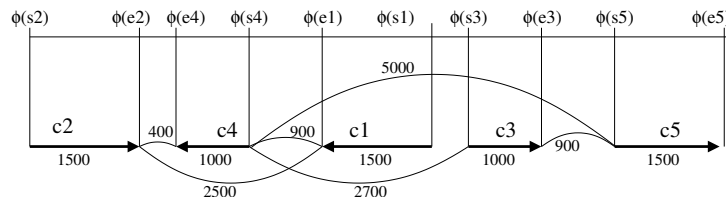
Additionally, we require $\{\phi(b_i), \phi(e_i)\} \neq \{\phi(b_j), \phi(e_j)\}$ for any two distinct contigs c_i and c_j .

10.30 Example

Given the following contig-mate graph:



An ordering ϕ assigns coordinates $\phi(v)$ to all nodes v and thus determines a layout of the contigs:



10.31 Happiness of mate edges

Let $G = (V, E)$ be a contig-mate graph and ϕ an ordering of G .

Consider a mate-edge e with nodes v and w . Let c_i denote the contig edge incident to v and let c_j denote the contig edge incident to w . Let v' and w' denote the other two nodes of c_i and c_j , respectively. We call e *happy* (with respect to ϕ), if c_i and c_j have the correct relative orientation, and if the distance between v and w is approximately correct, in other words, we require that either

1. $\phi(v') \leq \phi(v)$ & $|\phi(w) - \phi(v) - \mu(e)| \leq 3\sigma(e)$ & $\phi(w) \leq \phi(w')$, or
2. $\phi(w') \leq \phi(w)$ & $|\phi(v) - \phi(w) - \mu(e)| \leq 3\sigma(e)$ & $\phi(v) \leq \phi(v')$.

Otherwise, e is *unhappy*.

10.32 The Contig Ordering Problem

Given a collection of contigs $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$ constructed from a set of reads $\mathcal{F} = \{f_1, f_2, \dots, f_R\}$, together with the corresponding mate pair information M . Let $G = (V, E)$ denote the associated contig-mate graph.

Problem The *Contig Ordering Problem* is to find an ordering of G that maximizes the sum of weights of happy mate edges.

Theorem The corresponding decision problem is NP-complete.

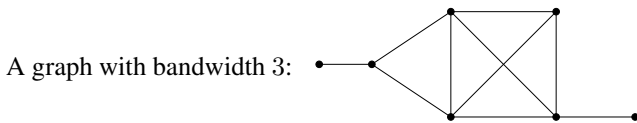
(The decision problem is: Given a contig-mate graph G , does there exist an ordering of G such that the total weight of all happy edges $\geq K$?)

10.33 Proof of NP-completeness

Recall: to prove that a problem X is NP-complete one must reduce a known NP-complete problem N to X . In other words, one must show that any instance I of N can be translated into an instance J of X in polynomial time such that I has the answer *true* iff J does.

We will use the following NP-complete problem:

BANDWIDTH: For a given graph $G = (V, E)$ with node set $V = \{v_1, v_2, \dots, v_n\}$ and number K , does there exist a permutation ϕ of $\{1, 2, \dots, n\}$ such that for all edges $\{v_i, v_j\} \in E$ we have $|\phi(i) - \phi(j)| \leq K$? (See Garey and Johnson 1979 for details.)



Problem is in NP: For a given ordering ϕ , we can determine whether the number of happy mate-edges exceeds the given threshold K in polynomial time by simple inspection of all mate edges.

Reduction of BANDWIDTH: Given an instance $G = (V, E)$ of this problem, we construct a contig graph $G' = (V', E')$ in polynomial time as follows:

First, set $V' := V$ and $E' := E$, and let these edges be the mate-edges, setting $\mu(e) := 1 + \frac{K-1}{2}$ and $\sigma(e) := \frac{K-1}{6}$ so as to obtain a happy range of $[1, K]$, and $w(e) := 1$, for every mate-edge e .

Then, for each initial node $v \in V$, add a new auxiliary node v' to V' and join v and v' by a contig edge of length 0.

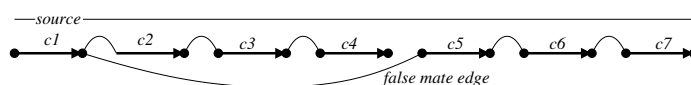
The answer to the BANDWIDTH question is *true*, iff the graph G' has an ordering ϕ such that all mate edges in G' are happy: □

$$\begin{aligned}
 &\text{A graph } G \text{ has BANDWIDTH } \leq K \\
 &\iff \\
 &\exists \text{ permutation } \phi \text{ such that } (v_i, v_j) \in E \text{ implies } |\phi(i) - \phi(j)| \leq K \\
 &\iff \\
 &\exists \text{ ordering } \phi \text{ such that } (v_i, v_j) \in E \text{ implies } 1 \leq |\phi(i) - \phi(j)| \leq K \\
 &\iff \\
 &\exists \text{ ordering } \phi \text{ such that } e = (v_i, v_j) \in E \text{ implies } \mu(e) - 3\sigma(e) \leq |\phi(i) - \phi(j)| \leq \mu(e) + 3\sigma(e) \\
 &\iff \\
 &\text{all mate-edges of } G' \text{ are happy.} \quad \square
 \end{aligned}$$

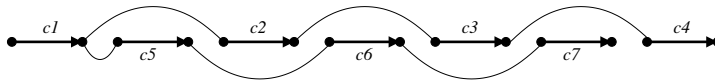
10.34 Spanning tree heuristic for the Contig Ordering Problem

An ordering ϕ that maximizes the number of happy mate edges is a useful scaffolding of the given contigs.

The simplest heuristic for obtaining an ordering is to compute a maximum weight spanning tree for the contig-mate graph and use it to order all contigs, similar to the read layout problem.



Unfortunately, this method does not work well in practice, as false mate edges lead to incorrect *interleaving* of contigs from completely different regions of the source sequence:



10.35 Representing an ordering in the mate-contig graph

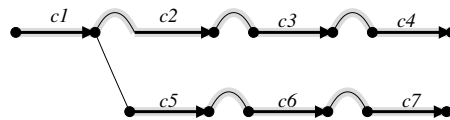
By the definition given above, an ordering is an assignment of coordinates to all nodes of the contig-mate graph that corresponds to a scaffolding of the contigs. When we are not interested in the exact coordinates, then the relative order and orientation of the contigs can be represented as follows:

Given a contig-mate graph $G = (V, E)$. A set $S \subseteq E$ of *selected* edges is called a *scaffolding* of G , if it has the following two properties:

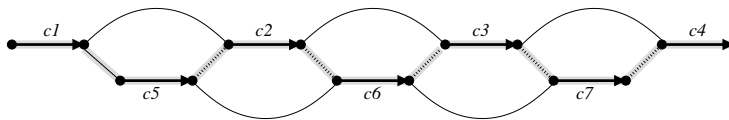
- every contig edge is selected, and
- every node is incident to at most two selected edges.

Thus, a scaffolding of G is a set of non-intersecting *selected paths*, each representing a scaffolding of its contained contigs.

The following example contains two chains of selected edges representing scaffolds $s_1 = (c_1, c_2, c_3, c_4)$ and $s_2 = (c_5, c_6, c_7)$:



However, to be able to represent the interleaved scaffolding discussed earlier, we need to add some *inferred* edges (shown here as dotted lines) to the graph:



10.36 Greedy path-merging

Given a contig-mate graph $G = (V, E)$. The greedy path merging algorithm is a heuristic for solving the Contig Ordering Problem. It proceeds “bottom up” as follows, maintaining a valid scaffolding $S \subseteq E$:

Initially, all contig edges c_1, c_2, \dots, c_k are selected, and none others. At this stage, the graph consists of k selected paths $P_1 = (c_1), \dots, P_k = (c_k)$.

Then, in ordering of decreasing weight we consider each mate edge $e = \{v, w\}$: If v and w lie in the same selected path P_i , then e is a chord of P_i and no action is necessary.

If v and w are contained in two different paths P_i and P_j , then we attempt to merge the two paths to obtain a new path P_k and accept such a merge, only if the increase of $H(G)$, the number of happy mate edges, is larger than the increase of $U(G)$, the number of unhappy ones.

10.37 The greedy path-merging algorithm

Algorithm Given a contig-mate graph G . The output of this algorithm is a node-disjoint collection of selected paths in G , each one defining an ordering of the contigs whose edges it covers.

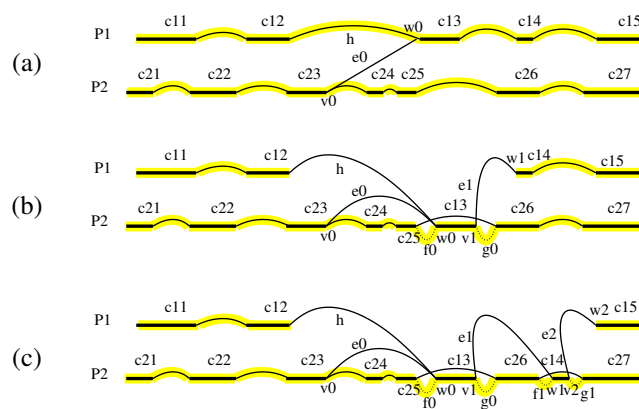
begin

Select all contig edges.

for each mate-edge e in descending order of weight:
if e is not selected:
 Let v, w denote the two nodes connected by e
 Let P_1 be the selected path incident to v
 Let P_2 be the selected path incident to w
if $P_1 \neq P_2$ **and** we can *merge* P_1 and P_2 (guided by e)
 to obtain P :
if $H(P) - (H(P_1) + H(P_2)) \geq U(P) - (U(P_1) + U(P_2))$:
 Replace P_1 and P_2 by P
end

10.38 Merging two paths

Given two selected paths P_1 and P_2 and a *guiding* unselected mate-edge e_0 with nodes v_0 (incident to P_1) and w_0 (incident to P_2). Merging of P_1 and P_2 is attempted as follows:

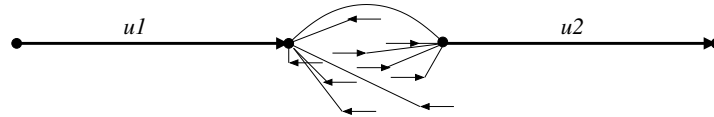


10.40 Repeat resolution

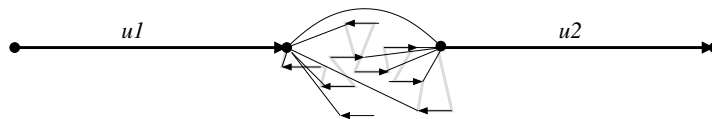
Consider two unique unitigs u_1 and u_2 that are placed next to each other in a scaffolding, due to a heavy mate edge between them:



We consider all non-unique unitigs and singleton reads that potentially can be placed between u_1 and u_2 by mate edges:



Different heuristics are used to explore the corresponding local region of the overlap graph in an attempt to find a chain of overlapping fragments that spans the gap and is compatible with the given mate pair information:



10.41 Multialignment

In a last step we have to compute a consensus sequence for each contig based on the layout of the fragments (this can also be done right after computing the contigs/unitigs).

```

R1          ACGCTCCAACCGCTAATACG
R2          ATCGCTAATCCACGCCCGCCCGC
R3    AAAC-CTCCAACCG
R4          TGC GCGCCCGCCCGAAACCGC
Consensus AAAC-CTCCAACCGCTAATGCGCGCCCGCCCGAAACCGC

```

10.42 Summary

Given a collection $\mathcal{F} = \{f_1, f_2, \dots, f_R\}$ of reads and mate pair information, sampled from a unknown source DNA sequence. Assembly proceeds in the following steps:

1. compute the overlap graph, e.g. using a seed-and-extend approach,
2. construct all unitigs, e.g. using the minimal spanning tree approach,
3. scaffold the unitigs, e.g. using the greedy-path merging algorithm,
4. attempt to resolve repeats between unitigs, and
5. compute a multi alignment of all reads in a given contig to obtain a consensus sequence for it.

Note that the algorithms for steps (2) and (3) that are used in actual assembly projects are much more sophisticated than ones described in these notes.

10.43 A WGS assembly of human (Celera)

Input: 27 million fragments of av. length 550bp, 70% paired:

5m	pairs of length 2kb
4m	pairs of length 10kb
0.9m	pairs of length 50kb
0.35m	pairs of length 150kb

Celera's assembler uses approximately the following resources:

Program	CPU hours		Max. memory
Screener	4800	2-3 days on 10-20 computers	2GB
Overlapper	12000	10 days on 10-20 computers	4GB
Unitigger	120	4-5 days on a single computer	32GB
Scaffolder	120	4-5 days on a single computer	32GB
RepeatRez	50	Two days on a single computer	32GB
Consensus	160	One day on 10-20 computers	2GB

Total: \approx 18000 CPU hours.

The size of the human genome is \approx 3Gb. An unpublished 2001 assembly of the 27m fragments has the following statistics:

- The assembly consists of 6500 scaffolds that span 2776Mb of sequence.
- The spanned sequence contains 150000 gaps, making up 148Mb in total.
- Of the spanned sequence, 99.0% is contained in scaffolds (or contigs?) of size 30kb or more.
- Of the spanned sequence, 98.7% is contained in scaffolds (or contigs?) of size 100kb or more.