DNA Sequence Assembly

Miklós Csűrös

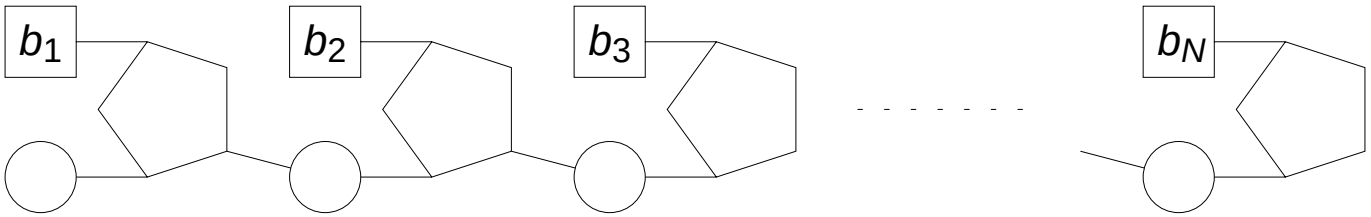
Département d'informatique et de recherche opérationnelle
Université de Montréal
csuros@iro.umontreal.ca

http ://www.iro.umontreal.ca/ csuros/

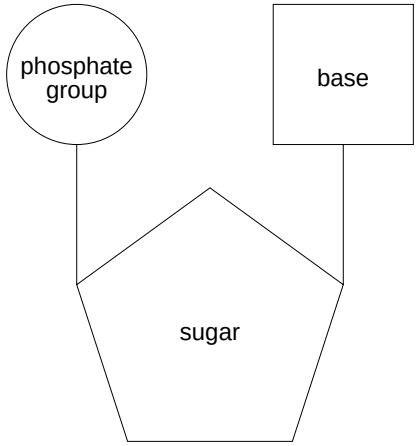
What is DNA ?

Polynucleotide :

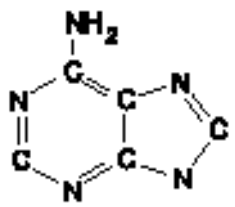
polymer (ie, chain) of [deoxyribo-]nucleotides

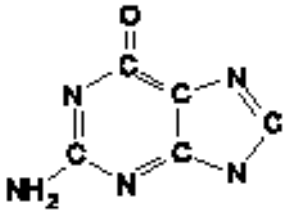


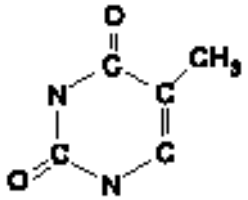
Nucleotides

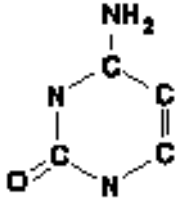


4 bases :

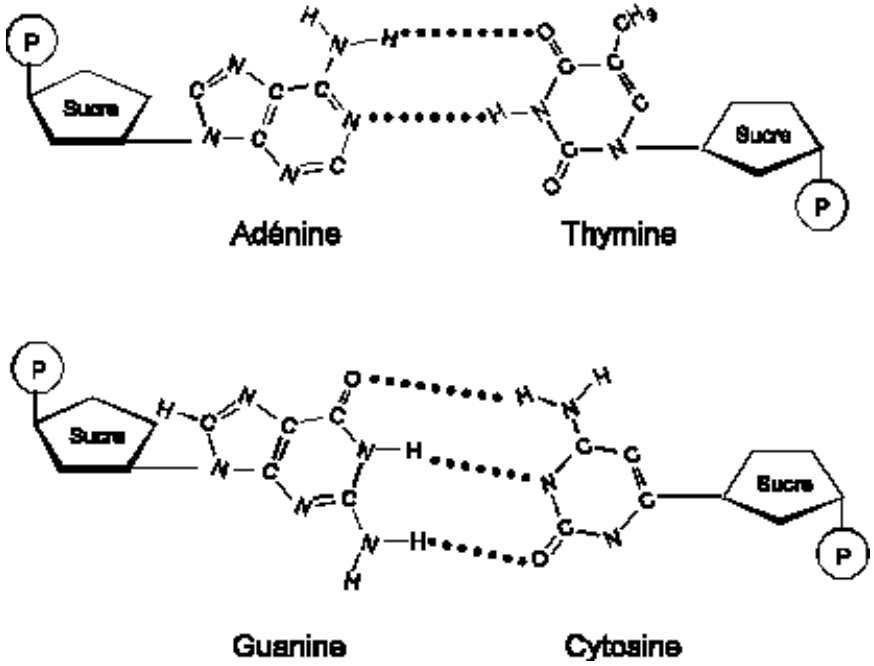

Adénine


Guanine


Thymine


Cytosine

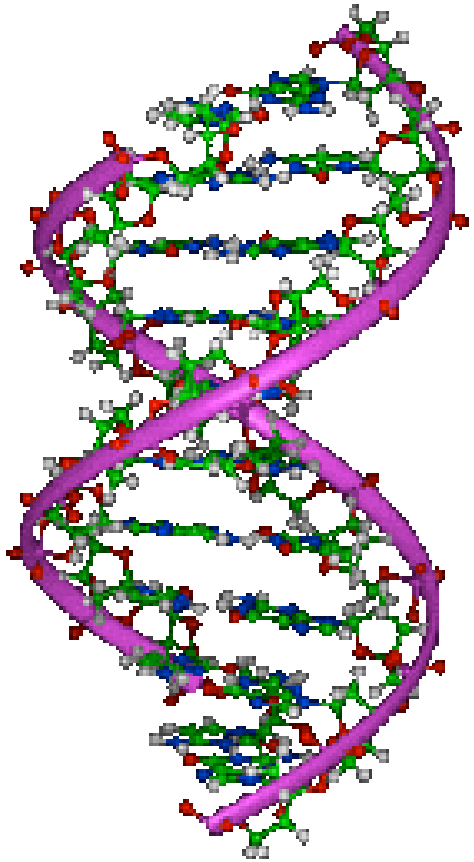
Base pairing



hydrogen bonds between A-T and G-C

DNA is formed by two polynucleotide strands

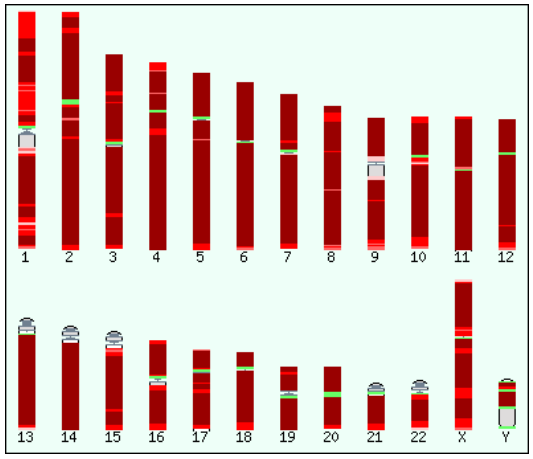
Watson et Crick : double helix



Genome

nuclear genome : in cell nucleus, arranged in chromosomes : every chromosome contains one DNA molecule

humans : 22 diploid chromosomes (1–22) +2 sex chromosomes (X and Y)
sizes between 50 and 250 million base pairs, total length $3 \cdot 10^9$ bp



The Problem

the longest piece of DNA we can determine the sequence of : cca. 1000 nucleotides (600 is more typical) — put together the whole HG from these short sequences

puzzle :

size of the HG : 3000 characters per page : book of a million pages (cca 6 cm per 1000 pages : 60 m)

pieces of 9–10 lines

$\Rightarrow$ It is impossible.

The End.

Completion of the human genome

draft : June 2000

For Immediate Release June 25, 2000

PRESIDENT CLINTON ANNOUNCES THE COMPLETION OF THE FIRST SURVEY OF THE ENTIRE HUMAN GENOME
Hails Public and Private Efforts Leading to This Historic Achievement

> Today, we are learning the language in which God created life. We are gaining ever more awe for the complexity, the beauty, the wonder of God's most divine and sacred gift.

finished : April 2003 (no press conf at White House)

How to do it

algorithms and biotechnology

: boundaries between often become blurred

Outline

- technologies : restriction enzymes, PCR, cloning, shotgun sequencing

- hierarchical and whole-genome shotgun approaches

- physical maps : fingerprints and STS

- assembly : overlap-layout-consensus

- SBH and Euler-path assembly

- pooling

Playing with strands

– double $\longrightarrow$ single : break the hydrogen bonds [eg, heat]
– single $\longrightarrow$ double :

1. hybridization attaching 2 complementary ssDNA
2. enzymes (polymerase) to complement a ssDNA template

How to cut DNA into smaller pieces

1. randomly (eg, sonication)

2. less randomly : restriction enzymes— cut DNA at a specific site

Name	Site
AluI	AG.CT
EcoRI	G.AATTC
HindIII	A.AGCTT
hundreds more . . .	

Restriction enzymes

RE site is often a palindrome (“Eva, can I stack cats in a cave?”)
— its inverse complement is the same

```
   ----->
5' GAATTC
   CTTAAG 3'
   <-----
```

sticky ends when cut is not in the middle
$\Rightarrow$ different pieces hybridize

Cloning

Idea :

1. fragments after cutting with RE
2. insert in a vector of a host (bacterium or virus)
3. let them grow

vectors : plasmids, phages ; BAC (*Bacterial Artificial Chromosome*), . . .

Cloning : animation

Polymerase Chain Reaction

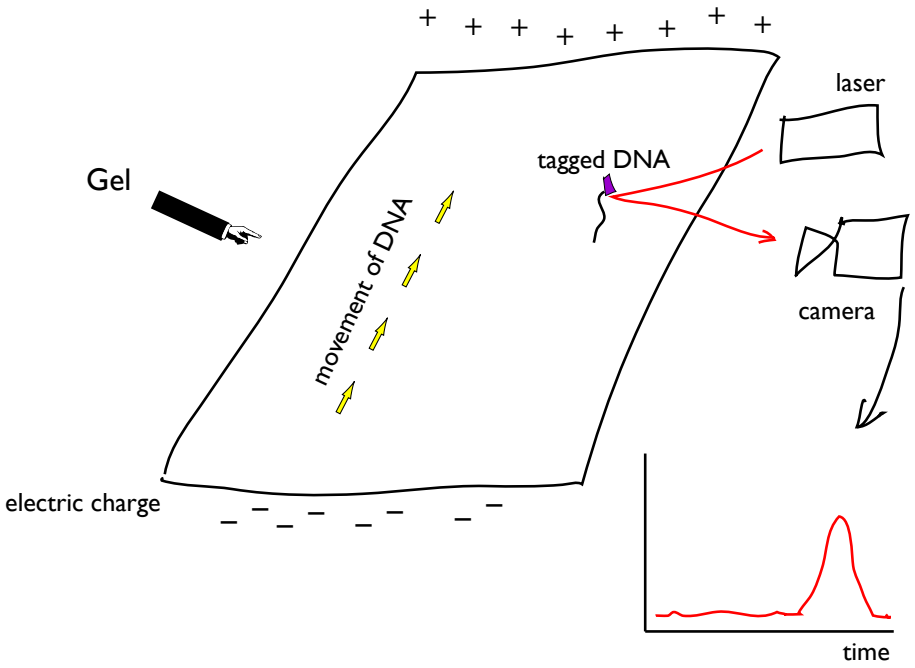
another method of copying DNA

Chain termination sequencing

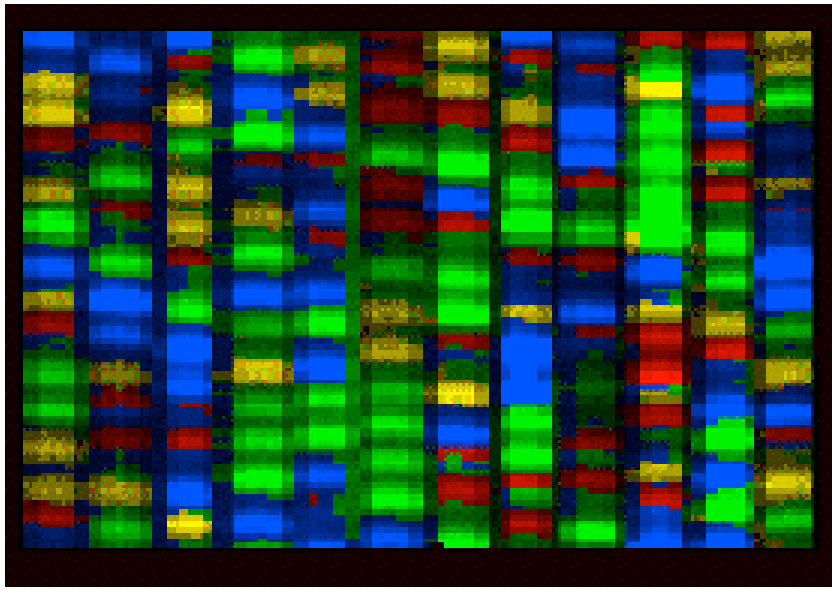
(aka Sanger sequencing) for fragments between about 50 and 1000 (typically 600) bp

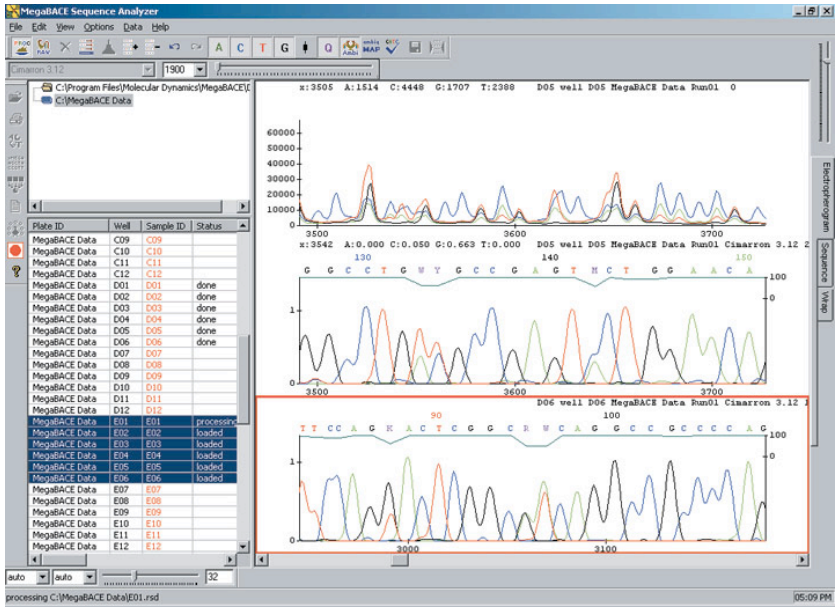
Electrophoresis

measures the size of a DNA fragment



Electrophoresis in sequencing

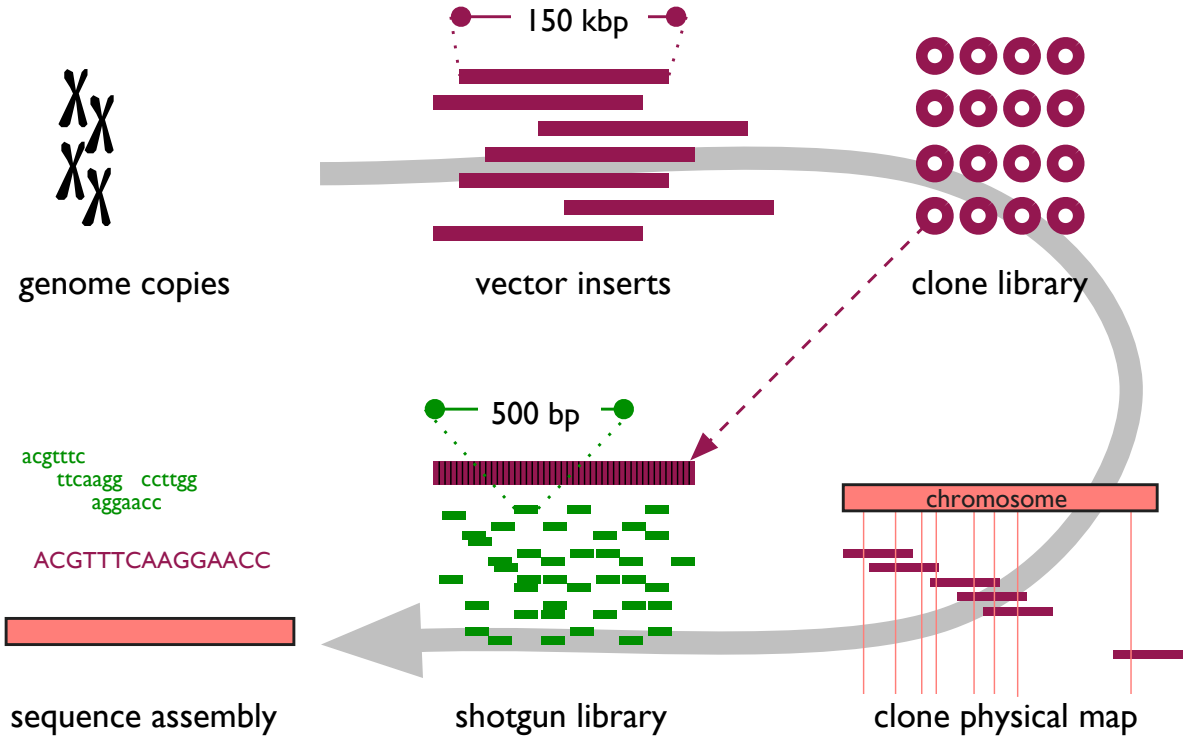




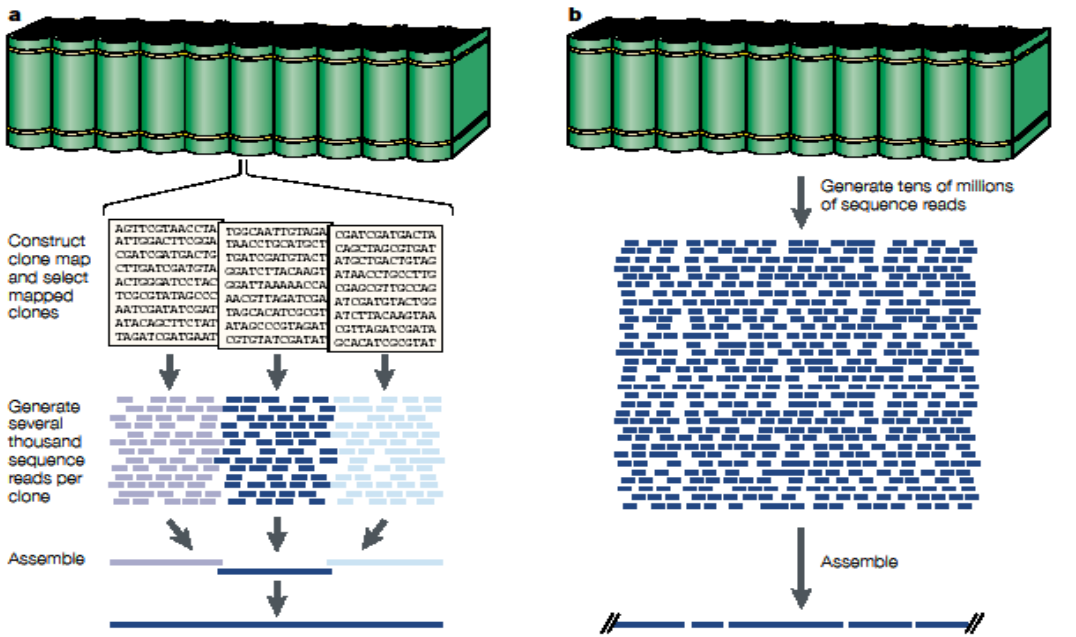
http ://www.megabace.com/

Sequencing a genome

hierarchical (clone-by-clone) method



Sequencing a genome



Clone-by-clone

Whole-genome shotgun

E. Green. Nature Reviews Genetics 2:573 (2001)

Sequencing a BAC clone

BAC insert length : around 150 thousand bp
(around 50 pages in the book)

random shotgun sequences from the BAC :
put them together based on overlaps

```
s1: AATGCC
s2:    GCCTTACAC
s3:         ACACTG
s4:             CTGAAGG
----------------------
B :  AATGCCTTACACTGAAGG
```

Shortest superstring

Def. T of length $|T| = m$ is a substring of S iff for some j,

$$T[1] = S[j], T[2] = S[j+1], \ldots, T[m] = S[j+m-1].$$

Shortest Supersting Problem (SSP) :
Given sequences $\mathcal{F} = \{s_1, s_2, \ldots, s_n\}$, find the shortest string S such that every s_i is a substring of S.

SSP is NP-hard, and even APX-hard.

Greedy algorithm is at most 3 (2 ?) optimal
Best [provable] algorithm is at most 2.5 times optimal

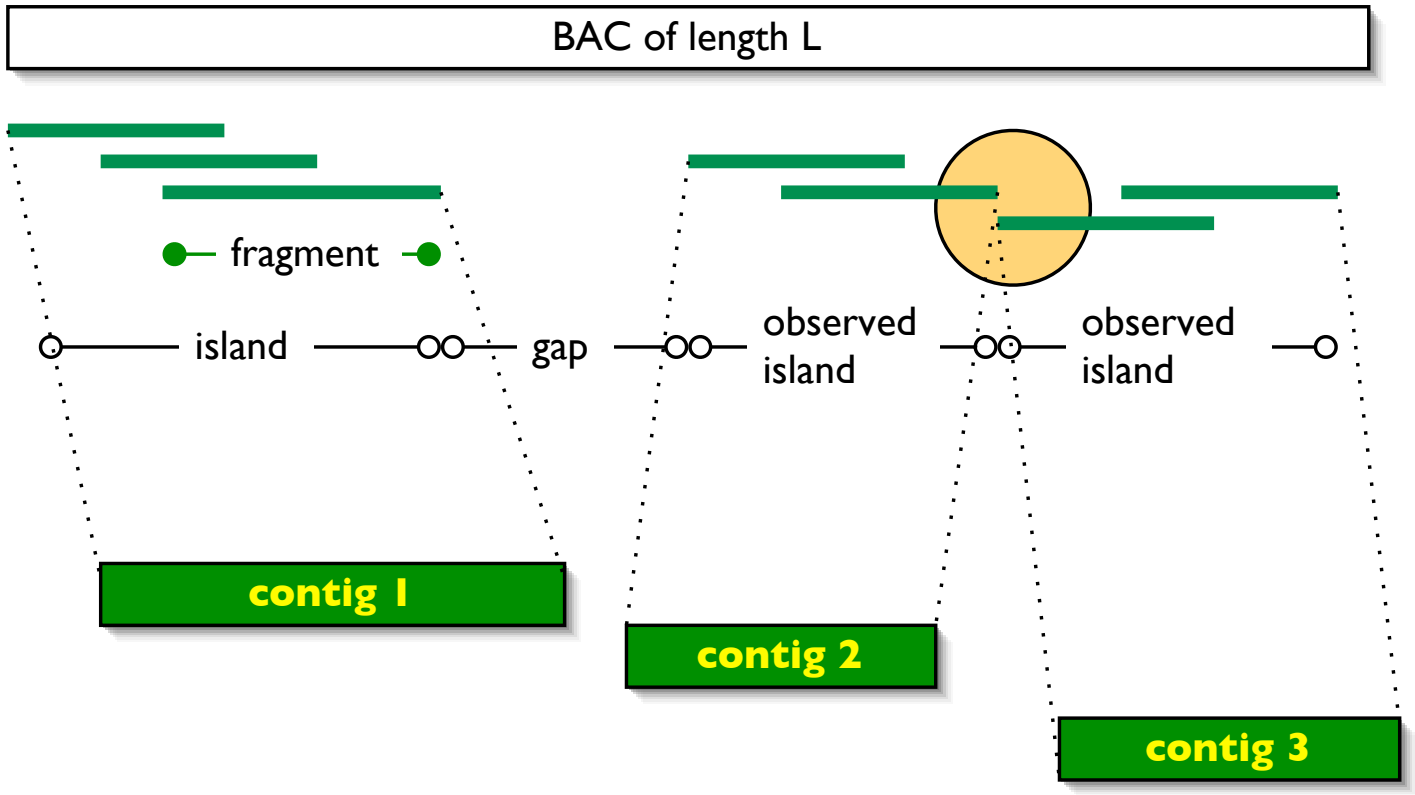
How many sequences are needed ?

number of fragments : n
length of each fragment : ℓ
length of BAC : L

coverage : $c = n\ell/L$

minimum overlap for joining fragments : $\theta\ell$ where $0 < \theta < 1$

Terminology



Coverage model — cont.

Thm. The probability that a particular position in the BAC is covered by at least one fragment is cca. $(1 - e^{-c})$.

Proof. Probability that a fixed fragment covers the position : $\ell/(L-\ell+1) \approx \ell/L$

Probability that none of them does : $\left(1 - \frac{\ell}{L}\right)^n$.

Approximation : $(1 - a/x)^x \approx e^{-a}$.

Model — cont.

Thm. The number of gaps is cca. $ne^{-c(1-\theta)} = \frac{\ell}{L}ce^{-c(1-\theta)}$.

Proof. Probability that a fixed fragment is the rightmost fragment of an observed island :

$$p = \left(1 - \frac{(1-\theta)\ell}{L}\right)^{n-1}.$$

+approximation as before

Expected number of gaps= np.

Model — cont.

Position of fragments defined by their right-hand end on BAC : random variables $X_1, X_2, \ldots, X_n$.

Set $h == 1 - \theta$.

Let's fix a fragment (X_1). What is the position Y_1 of the first fragment after X_1 ? Probability that $Y_1 > X_1 + h\ell$ equals

$$J(h) = \left(1 - \frac{h\ell}{L}\right)^{n-1} \approx \left(1 - \frac{ch}{n}\right)^{n} \approx e^{-ch}.$$

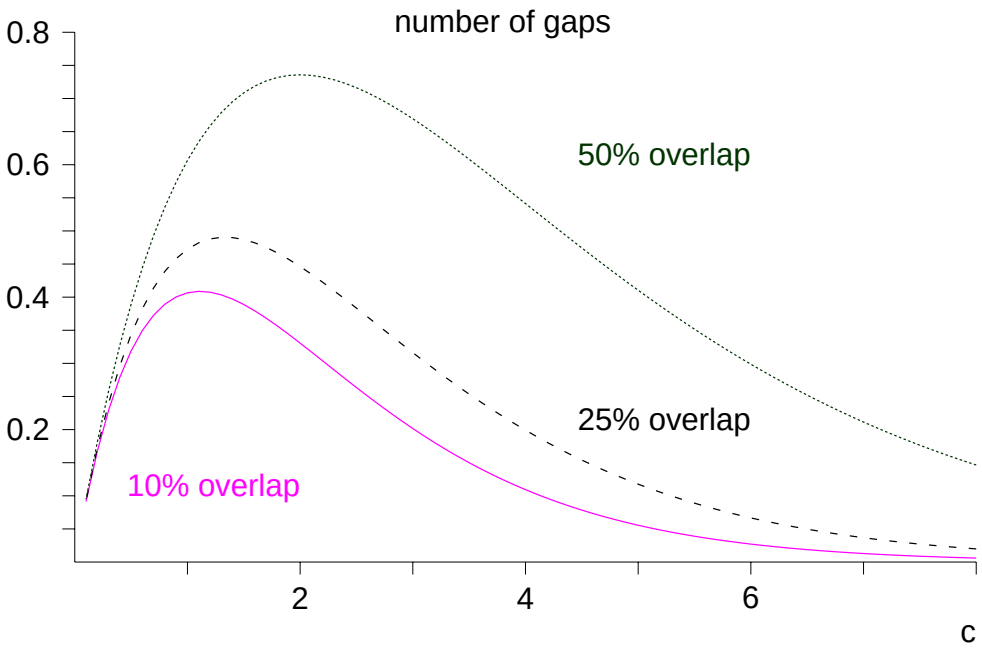
Model — cont.

Thm. The number of fragments in an observed island is cca. $e^{c(1-\theta)}$.

Proof. Let M be the number of fragments in the island.
Consider the first fragment in the island. Probability that the island is a singleton ($M = 1$) : $p_1 = J(1-\theta)$.
Probability that $M = k$: $p_k = \Big(1 - J(1-\theta)\Big)^{k-1} J(1-\theta)$ — geometric distribution. Expectation of M is $1/J(1-\theta)$.

Modelization by Poisson process.

Model — cont.



typical coverages : 5 (“half shotgun”) 10 (“full shotgun”)
for $\ell = 500, L = 150000, n = 1500$ or $n = 3000$.

Physical maps

Given a random BAC library, select a minimal overlapping set for complete sequencing.

15-fold coverage for human genome $\approx$ 300 thousand clones.

How can you do the selection without sequencing each BAC ?
$\Rightarrow$ physical mapping : BAC overlaps detected based on shared *features*

features :
1. restriction maps : fingerprint consisting of fragment sizes after digestion with a restriction enzyme

2. STS maps : containing specific substrings : Sequence Tagged Site (verified through hybridization)

Physical maps

Map characteristics (in increasing order of difficulty)

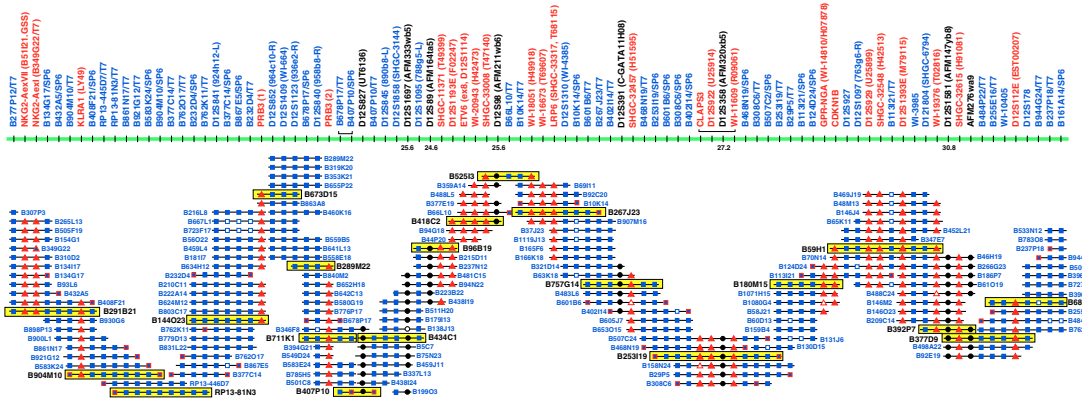
1. clone ordering by physical locations

2. clone coordinates (in base pairs)

3. links to chromosome locations

STS map : clones can be linked to chromosome locations
(STS can be located on the chromosome, eg by FISH : fluorescent in-situ hybridization)

A resolved map



STS map

Probes $\mathcal{P} = \{p_1, p_2, \cdots, p_n\}$,
clones $\mathcal{C} = \{c_1, \ldots, c_m\}$

Hybridization matrix : binary matrix $\mathbf{M}$ of size $m \times n$ such that $\mathbf{M}[i, j] = 1$ iff clone c_i hybridizes with probe p_j.
(Ie, clone c_i contains the substring corresponding to p_j.)

Problem : find correct permutation for the clones and probes.

STS map — no errors

The correct row and column permutation of $\mathbf{M}$ corresponds to a matrix with the consecutive ones property.

Consecutive ones [in a row] : for every row i there exists a, b such that $\mathbf{M}[i, j] = 1$ iff $a \leq j \leq b$.

Why ?

Permutaion found in linear time (PQ-tree).

STS map — errors

Hybridization errors in the matrix. Best permutation minimizes the number of implied errors.

Traveling salesman formulation (for ordering the rows) :
vertices $u, v_1, \ldots, v_m$, where v_i corresponds to row i.

Weighted edges : edge weight between v_i and $v_{i'}$ equals the number of columnes in which they differ ;
edge weight between u and v_i equals the number of “1” entries in row i.

Now a Hamiltonian path with smallest weight (Traveling salesman) gives the best ordering :
it minimizes the number of gaps within rows.

Why ?

STS map — errors

Maximum likelihood formulation : probabilistic model with
- model for probe locations (Poisson process)
- probability of false negatives (missed hybridization)
- probability of false positives (false hybridization)

Probability for a given permutation *and* clone placement.

Find best permutation and clone placement that minimizes the likelihood.

(Not very easy but works well.)

Pooling

Combinatorial group testing : 1940s

Wasserman-test for $n = k^2$ individuals, at most one of them is infected.

Arrange the blood samples in a $k \times k$ matrix, test samples in rows and columns together.

Positive in row i and column $j \Rightarrow$ the sample in cell $[i, j]$ is positive.

$\rightarrow 2\sqrt{n}$ tests instead of n.

(Minimum number of tests : $\lceil \log_2 n \rceil$)

Pooling 2

Testing a clone (or set of clones) : array probes and pool them by rows and columns. If clone hybridizes with row i and column j, then it contains the substring for probe in cell $[i, j]$.

	C_1	C_2
R_1	B_{11}	B_{12}
R_2	B_{21}	B_{22}

if hits from R_1, R_2, C_1, and C_2 : B_{11} and B_{22} or B_{12} and B_{21} ?

Transversal design

Let q be a prime. A transversal design for q^2 elements is the following :

Number the items as $B_{a,b}$: $a, b = 0, \ldots, q-1$.
Define pools $P_{x,y}$: $x, y = 0, \ldots, q-1$.

Pool $P_{x,y}$ contains $B_{a,b}$ iff $y \equiv a + bx \pmod{q}$ holds.

For an array layout : define pool sets $\mathcal{P}_i = \{P_{i,y} : y = 0, \ldots, q-1\}$ for $i = 0, \ldots q-1$.
Assign the pool sets as rows and columns.

Transversal design 3

10	15	20	25	5
14	19	24	4	9
18	23	3	8	13
22	2	7	12	17
1	6	11	16	21

8	11	19	22	5
15	18	21	4	7
17	25	3	6	14
24	2	10	13	16
1	9	12	20	23

Transversal design 4

Thm. For two different pools $P_{x_1,y_1} \cap \mathcal{P}_{x_2,y_2}$ has exactly one element if $x_1 \neq x_2$ and is empty if $x_1 = x_2$.

Proof An item $B_{a,b}$ is included in both iff

$$y_1 = ax_1 + b \quad \text{and} \quad y_2 = ax_2 + b.$$

Now if $x_1 = x_2$ then $y_1 = y_2$, otherwise there is exactly one solution for (a, b).

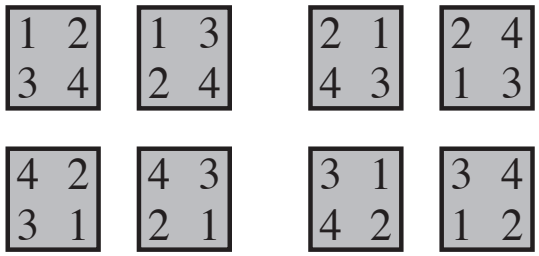
Other properties : every item is included in exactly once in every pool set two items appear in the same row or column at most once

$\Rightarrow$ every item is identifiable by two of the pools it's included in ; up to $(q-2)$ items are identifiable by the set of pools they are included in.

Random shuffling

method : random shuffling

rectangle : two rows and two columns
preserved rectangle : 4 items in a rectangle after and before shuffling with the same items on the diagonals



Theorem. Expected number of preserved rectangles is approximately $1/2$.

Random shuffling

Proof. Probability that a particular rectangle is preserved :

$$p = \frac{8\binom{m}{2}^2}{(m^2)(m^2-1)(m^2-2)(m^2-3)}.$$

Expected number of preserved rectangles : $\binom{m}{2}^2 p = \frac{1}{2} + \frac{2}{m}(1 + o(1))$.

works also if non-square array (unlike transversal designs)

expected number of preserved rectangles on shuffled $m \times m'$ array

$$\frac{1}{2} + \frac{m + m'}{mm'}(1 + o(1))$$

A more practical view of assembly

- sequencing errors

- repeats

- unknown orientation

Sequencing errors

Phred : base quality values

$$q = -10 \cdot \log_{10} p$$

where p is the error probability.

Calculated from trace taking intoo account various specifics of the sequencing machine.

Sequences are then typically trimmed at the beginning and the end (where low quality values are).

Errors

Def. T is an ϵ-approximate substring of S iff S has a substring S' such that the edit distance b/w T and S' is less than $\epsilon|T|$.

RECONSTRUCTION Problem :
Given error rate ϵ and sequences $\mathcal{F} = \{s_1, s_2, \ldots, s_n\}$, find the shortest string S such that every s_i is an ϵ-approximate substring of S.

NP-hard (take $\epsilon = 0$ to get SSP).

Overlaps

Comparing a large number of shotgun sequences to find overlaps :

Hashtable of k-long sequences appearing in sequences : quick location of seeds for alignment

Extension of alignments in a greedy fashion or by banded dynamic programming.

Build overlap graph : vertices correspond to sequences, edges indicate overlap.

Grow a path in the graph in a greedy fahsion.

Repeats

Double-barreled shotgun : mate-pair sequences
pairs of shotgun sequences with a “known” distance between them.

WGA assembler : scaffolds (contigs with known orientation and disatance between them)

Hybrid assembly

Orienting BAC sequences using a set of (WGS) mate pair reads.

Overlap graph : BAC endpoint vertices. Goal : assign integer values to vertices.

Directed edges :

1. edge between BAC endpoint vertices with known length

2. if sequences of a mate pair match two different BACs (implies orientation and distance), then add mate link edge between BAC endpoints : has length and standard deviation σ.

Hybrid assembly 2

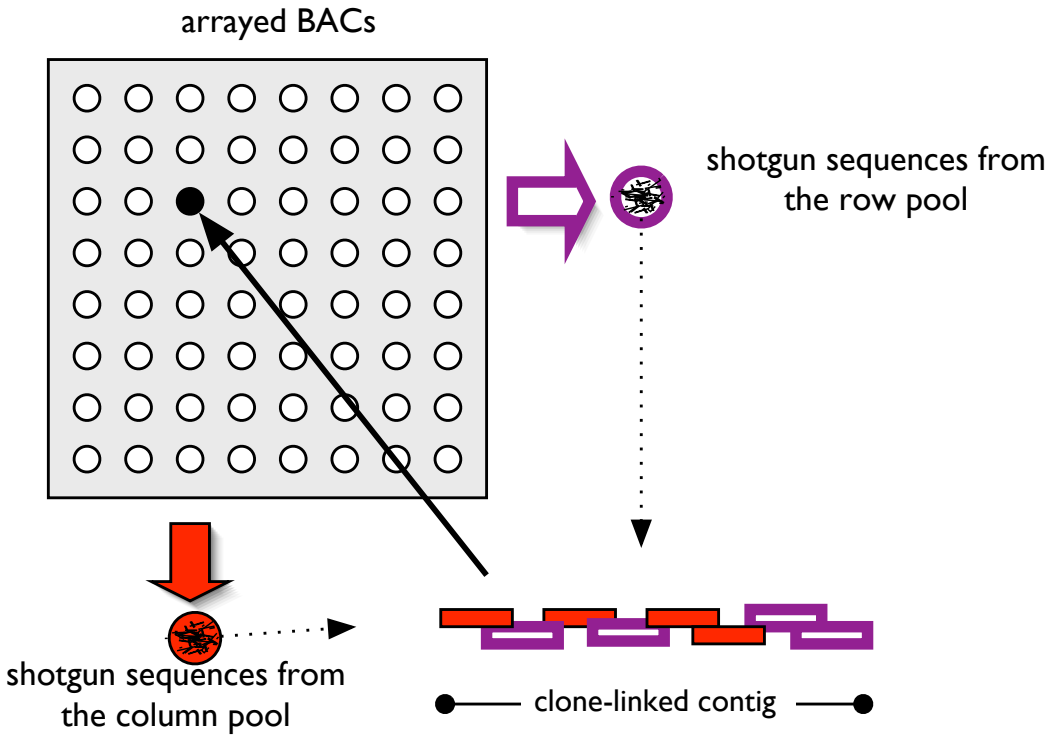
Edge bundling : collapse edges between the same BACs (length and stdev calculated as weighted sums) ; weights assigned to bundled edges.

Path bundling : similar but eliminates longer edges.

Happy edge : when its constraints are satisfied.

Problem : maximize the weight of happy edges — NP-hard.

Clone-array pooled shotgun sequencing



Pooling advantage

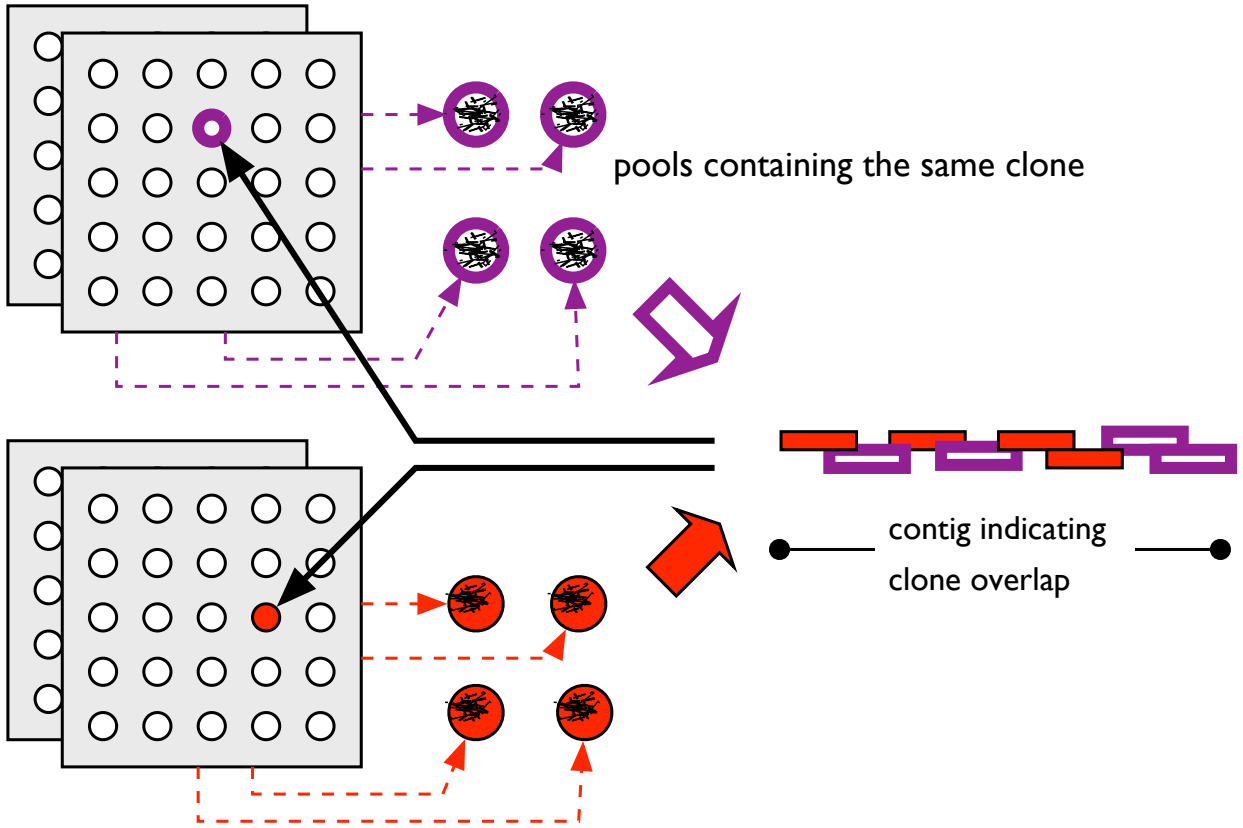
$\Rightarrow$ much fewer shotgun libraries

Example : human genome

N clones ($N \approx 20$ thousand)

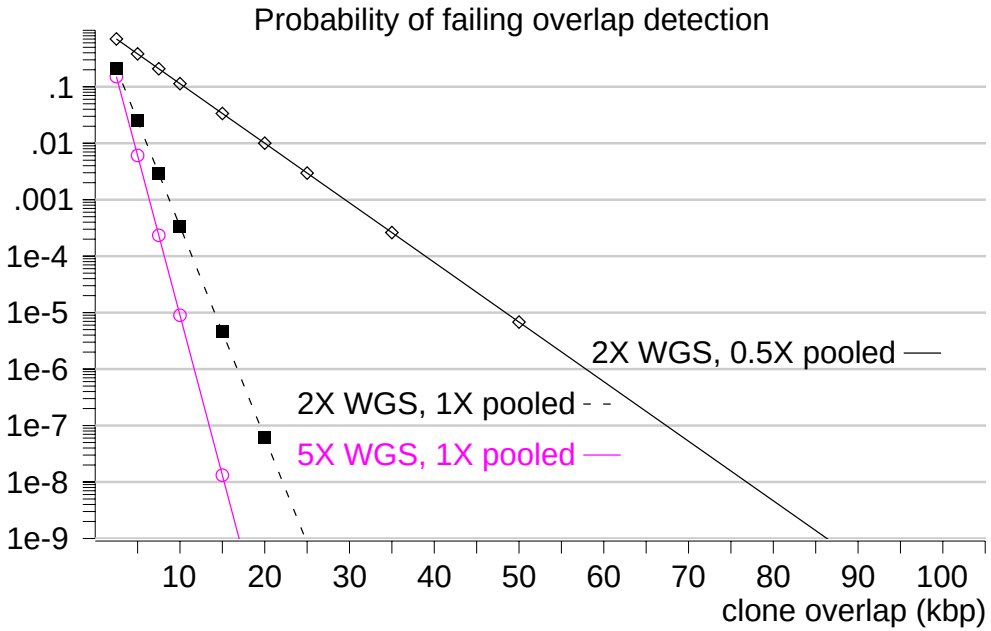
	shotgun libraries
CAPSS	$2\sqrt{N} \approx 300$
BAC-based sequencing	$N \approx 20000$

CAPS-MAP : physical mapping



Physical map of honeybee

135Mb genome :
3312 clones, 135kbp clone length : 3.3X clone coverage
six 24×24 array pairs [transversal design for each pair]
1X pooled reads, 5X whole-genome shotgun reads

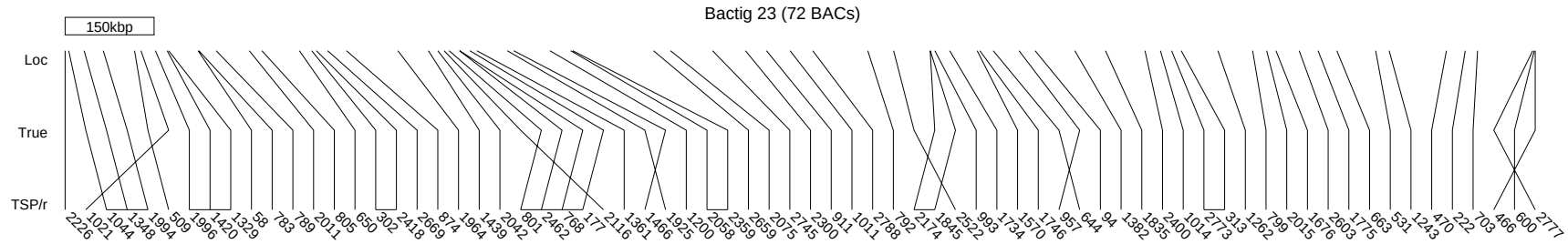


CAPS-MAP simulation

Simulated assembly of Drosophila melanogaster genome.

2880 BACs, 1.4X pooled shotgun, 4X whole-genome shotgun reads.

Example BAC ordering.



Sequencing by hybridization

Set of probes : all length-k sequences (4^k of them). Hybridize target sequence to probes : get spectrum.

Def. Spectrum $\mathcal{F}_k(S)$ of a string S : set of all of its substrings with length k.

SBH Problem :
Given a set of k-length sequences $\mathcal{F} = \{s_1, \cdots, s_n\}$, find a sequence S for which the spectrum equals $\mathcal{F}$.

Euler path formulation : vertices of $(k-1)$-length strings, edges correspond to s_i.

Another look at shotgun assembly

Euler-path formulation for shotgun assembly : cut shotgun sequences into smaller pieces. Use mate pair information to untangle paths.

(Waterman & Idury, Pevzner)

Finishing

Use pooling (Beigel et al.)