Large Scale Sequencing By Hybridization

Ron Shamir Dekel Tsur

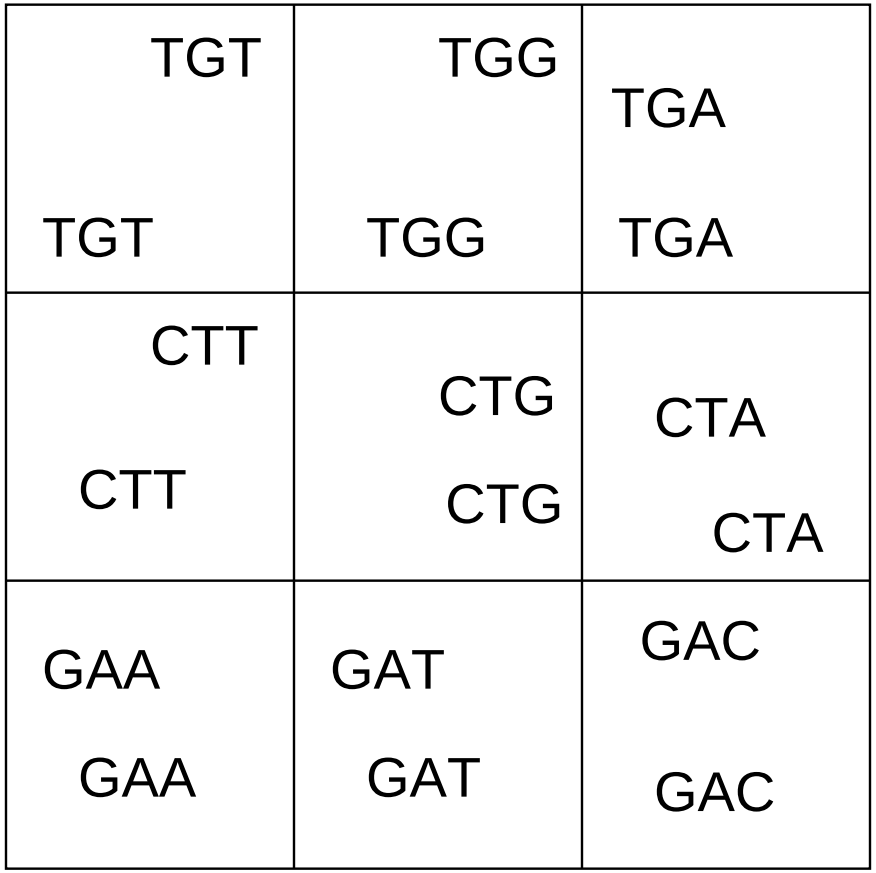
Tel Aviv University

Outline

- Background: SBH
- Shotgun SBH
- Analysis of the errorless case
- Analysis of error-prone

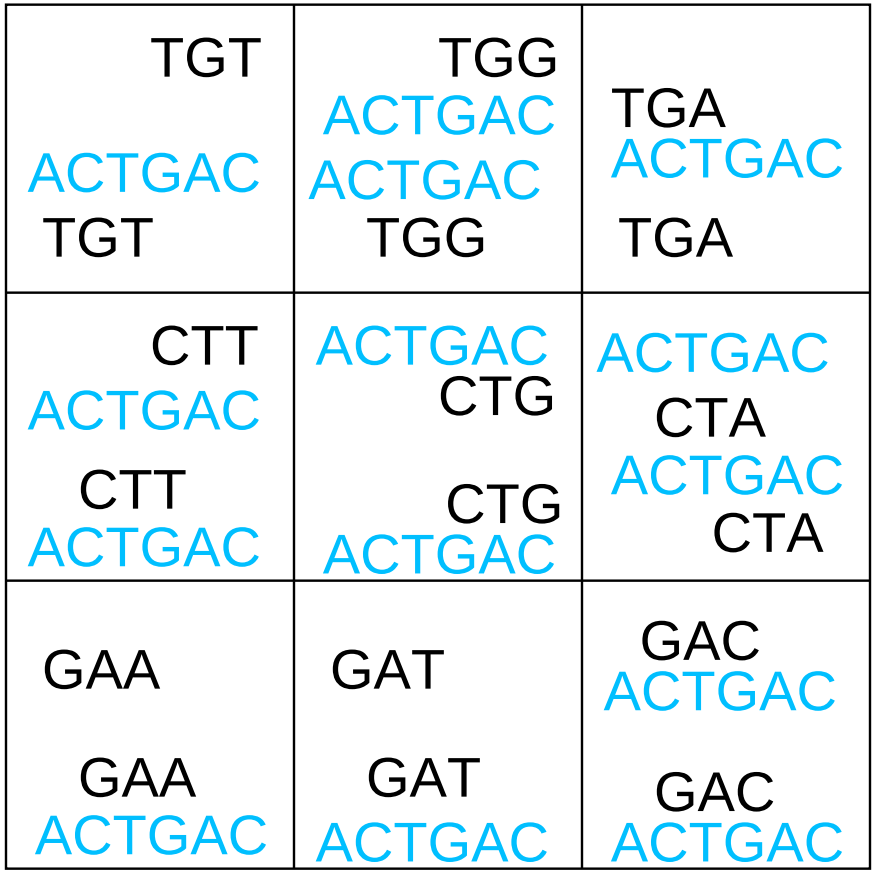
Sequencing By Hybridization (SBH)

Hybridize target to array containing a spot for each possible k-mer.



Sequencing By Hybridization (SBH)

Hybridize target to array containing a spot for each possible k-mer.



Sequencing By Hybridization (SBH)

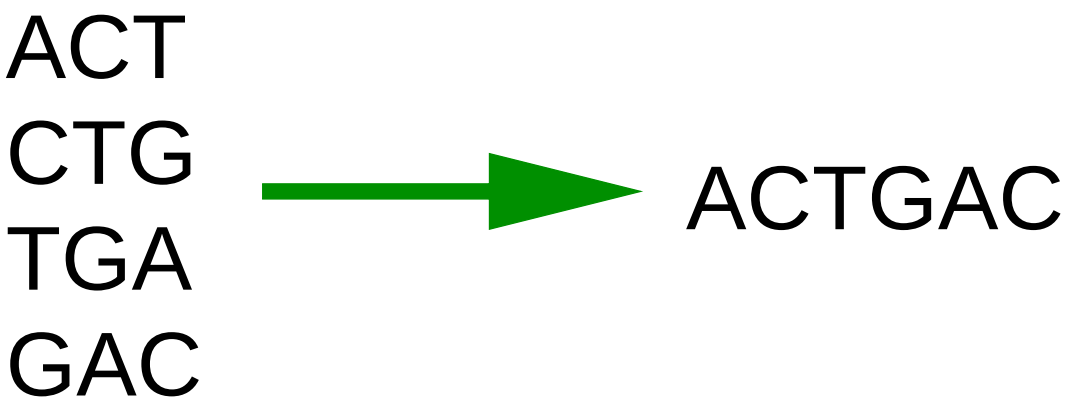
Hybridize target to array containing a spot for each possible k-mer.

TGT ACTGAC TGT	TGG ACTGAC ACTGAC TGG	TGA ACTGAC TGA
CTT ACTGAC CTT ACTGAC	ACTGAC CTG CTG ACTGAC	ACTGAC CTA ACTGAC CTA
GAA GAA ACTGAC	GAT GAT ACTGAC	GAC ACTGAC GAC ACTGAC

Sequencing By Hybridization

The spectrum of a sequence: multi-set of all its k-long substrings (k-mers).

Goal: reconstruct the sequence from its spectrum.

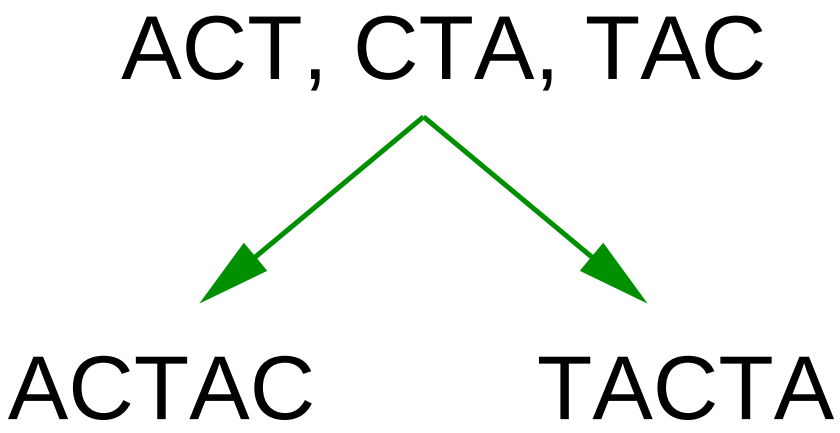


Pevzner 89: reconstruction is polynomial.

But...

Reconstruction May Be Non-unique

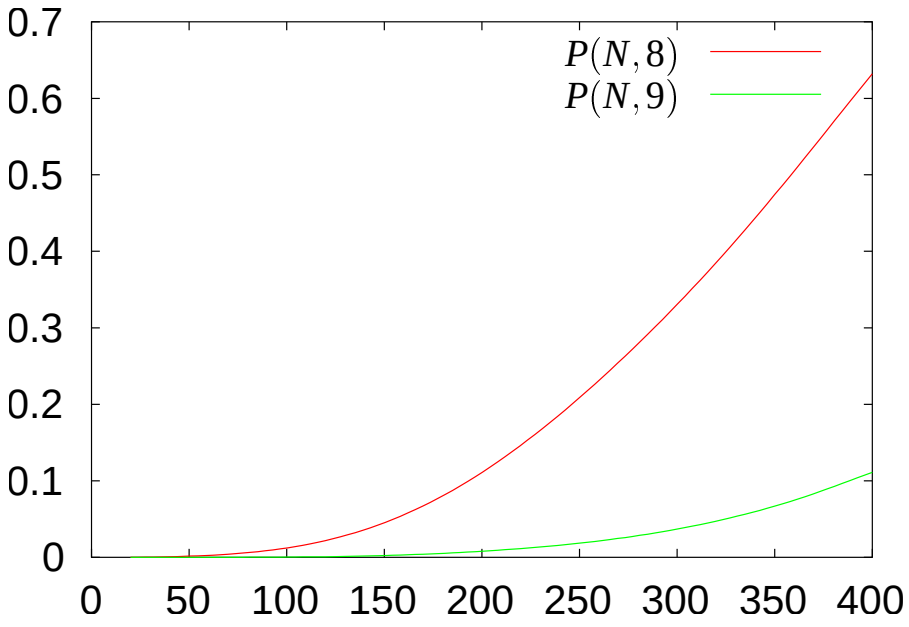
Different sequences can have the same spectrum:



Non-uniqueness Probability

$P(N,k)$: prob. that for a random sequence of length N, $\exists$ another sequence with same k-spectrum (failure probability).

Arratia et al (97): asymptotically tight bounds for $P(N,k)$.



Resuscitating SBH

$\Rightarrow$ SBH is currently not competitive for sequencing.

How can one make it competitive?

Shotgun SBH

(Drmanac, Labat, Brukner, Crkvenjakov 89)

1. Fragment target S into overlapping clones; obtain the spectrum of each clone.

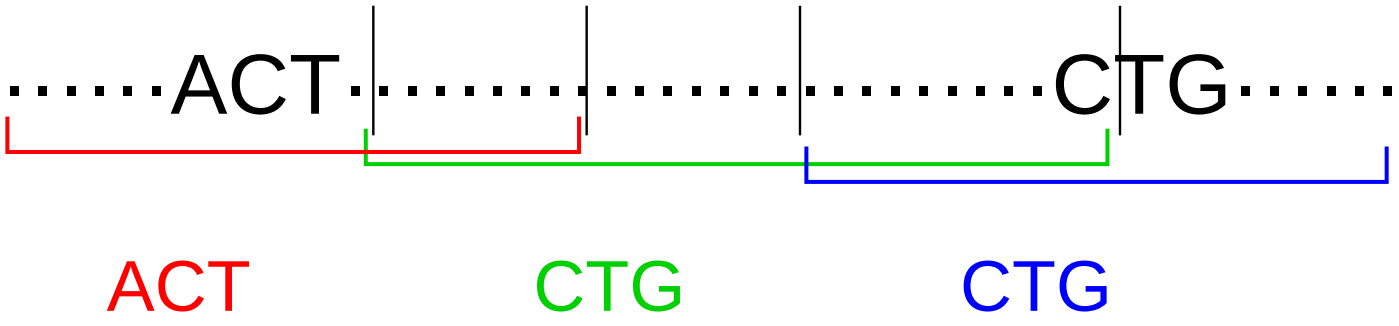
ACT	TAG	TTA
CTA	AGT	TAC
TAG	GTT	ACT
GTT	TTA	CTC
TTA	ACT	TCT
	CTC	CTG

Shotgun SBH

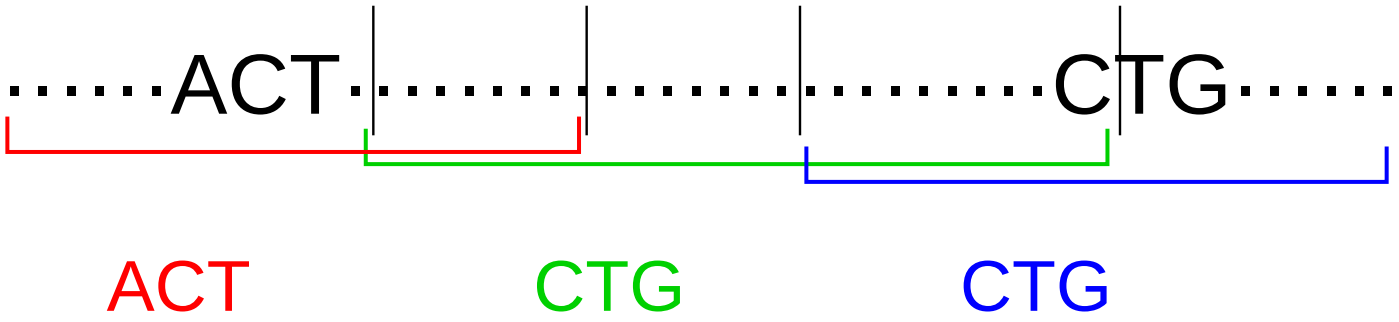
2. Find the correct clone map (e.g., Mayraz and Shamir, 98).

Shotgun SBH

2. Find the correct clone map (e.g., Mayraz and Shamir, 98).
3. The clones endpoints form a partition of the sequence S into subsequences called information fragments (IF). For each IF, compute its spectrum.

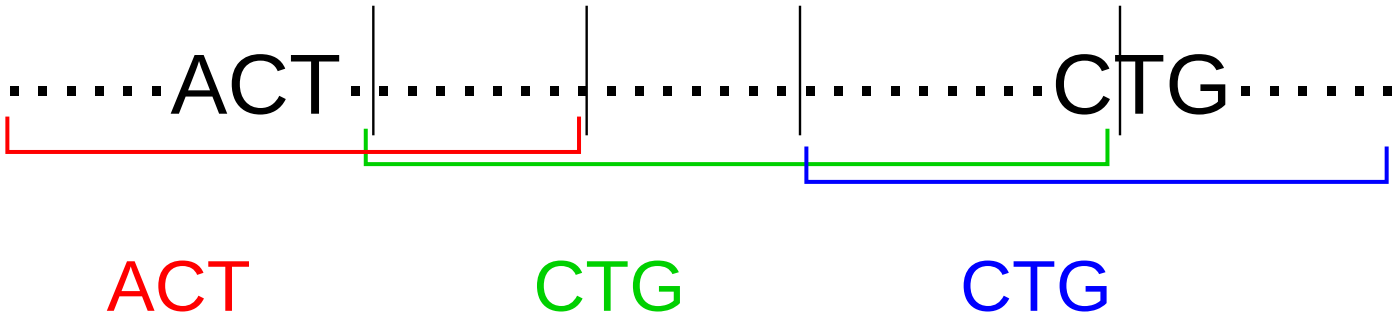


Shotgun SBH



4. Reconstruct the sequence of each IF.

Shotgun SBH

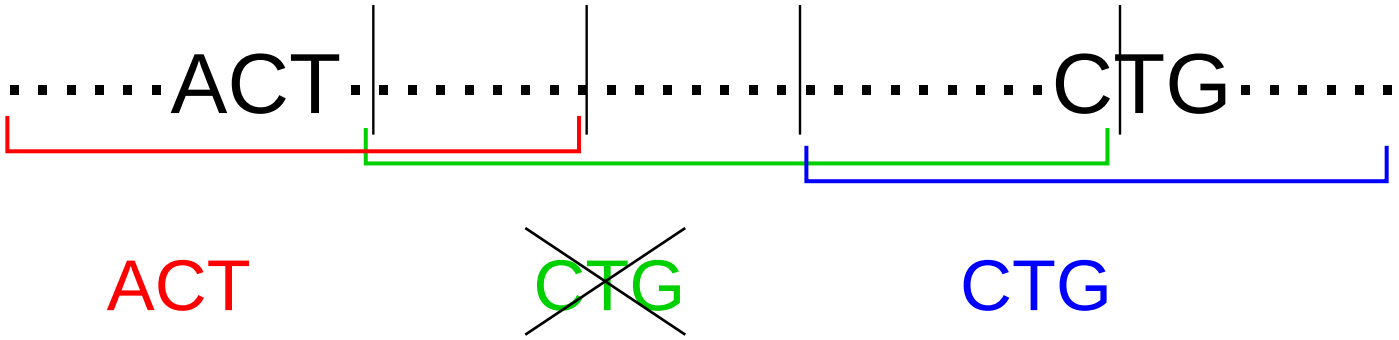


4. Reconstruct the sequence of each IF.
5. Combine the sequences of the IFs.

Hybridization Errors

Hybridization experiments are error prone.

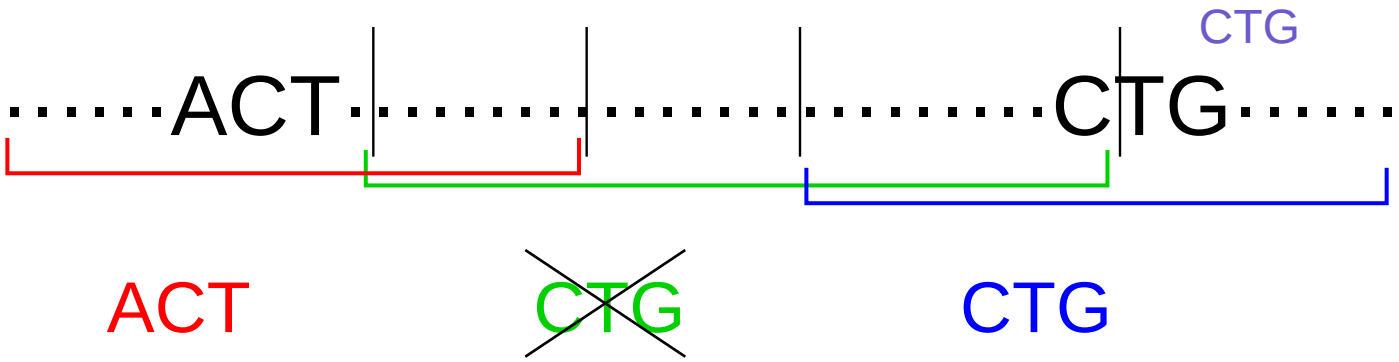
A false negative error: k-mer appears in a clone but does not appear in its measured spectrum.



Hybridization Errors

Hybridization experiments are error prone.

A false negative error: k-mer appears in a clone but does not appear in its measured spectrum.



Goal

Dramanac et al.: simulation evidence that shotgun SBH works in the absence of errors.

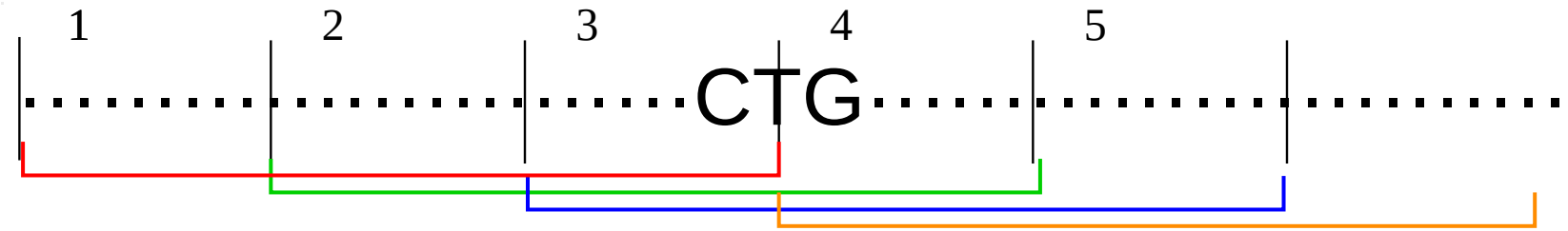
Our Goal: Rigorous analysis, also considering the impact of errors.

Assumptions

- Clones positions are known.
- Equal size IFs ($= d$).
- Each k-mer of target appears in at least one clone spectrum.
- Random sequence: equiprobable bases, independent positions.
- False negative probability p independently for each k-mer and for each clone.

Hybridization Errors (2)

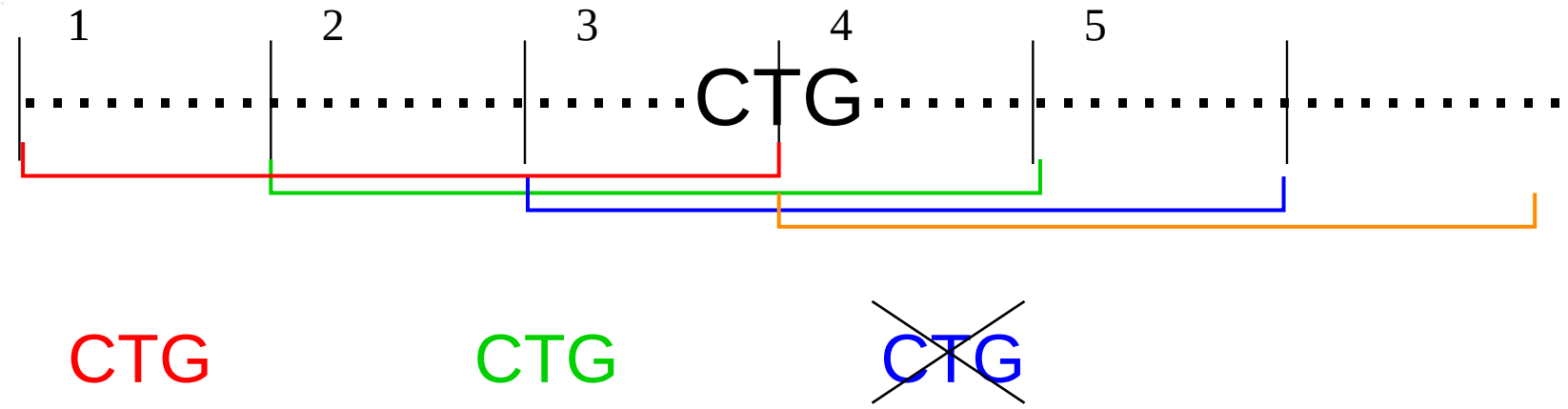
For each k-tuple P in the spectrum, we attribute P to the i-th IF where i is the maximum index of a clone in which P appears.



CTG CTG CTG

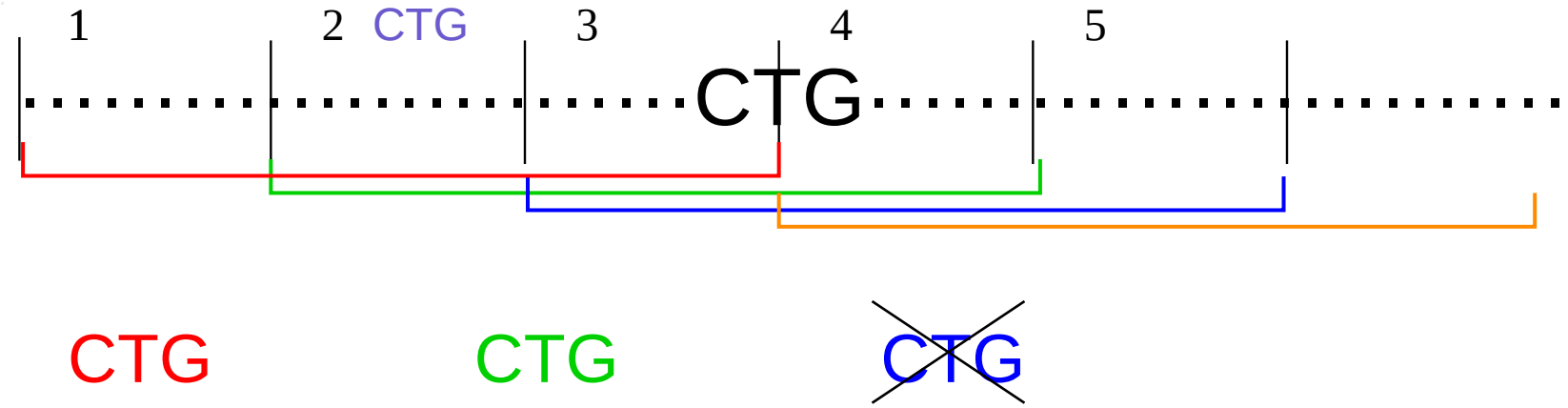
Hybridization Errors (2)

For each k-tuple P in the spectrum, we attribute P to the i-th IF where i is the maximum index of a clone in which P appears.



Hybridization Errors (2)

For each k-tuple P in the spectrum, we attribute P to the i-th IF where i is the maximum index of a clone in which P appears.



The computed index is always $\leq$ the true index.

Main Result

$N =$ sequence length

$k =$ probe length

$d =$ length of IFs

$p =$ false negative probability

$P(N, k, d, p)$: failure probability

Theorem $P(N, k, d, p) \leq \left(1 + \frac{c_p}{d}\right) P(N, k, d, 0).$

Overview of the Proof

Will show:

- $P(N, k, d, 0) = \Omega(\frac{d^3 N}{4^{2k}})$.
- $P(N, k, d, p) - P(N, k, d, 0) = O(\frac{d^2 N}{4^{2k}})$.

The de-Bruijn Graph (Pevzner 89)

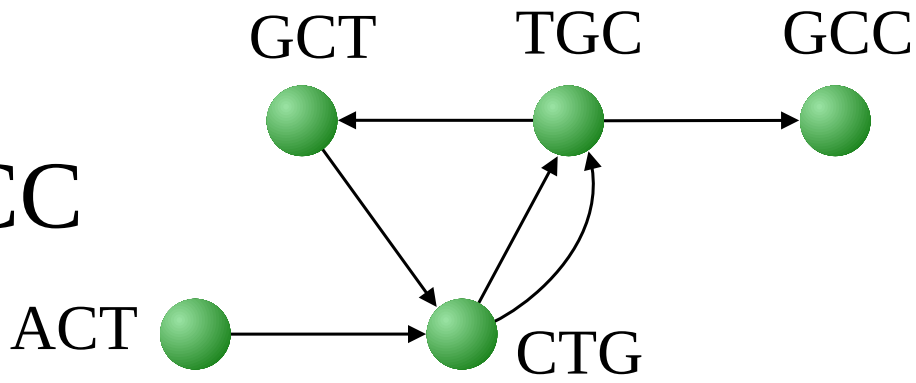
$A = a_1 \cdots a_{n+k-1}$: the sequence.

A_i : the $(k-1)$-mer $a_i a_{i+1} \cdots a_{i+k-2}$.

The de-Bruijn graph of A : $G_A = (V, E)$ where

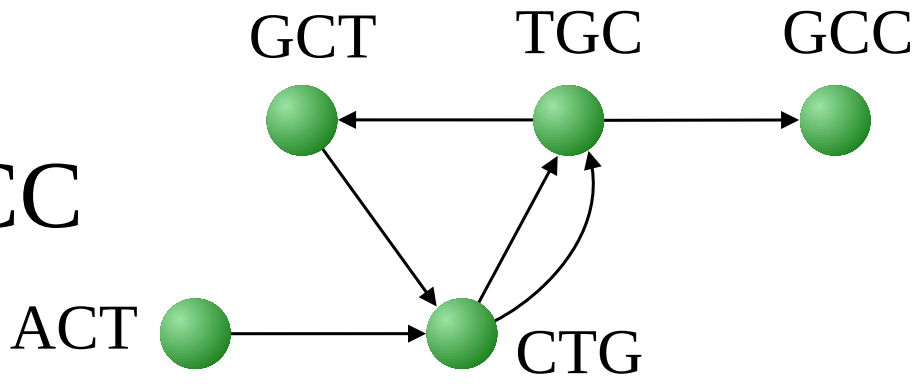
- $V = \{A_i : i = 1, \ldots, n+1\}$
- $E = \{e_i : i = 1, \ldots, n\}$, $e_i = (A_i, A_{i+1})$

ACTGCTGCC



The de-Bruijn Graph

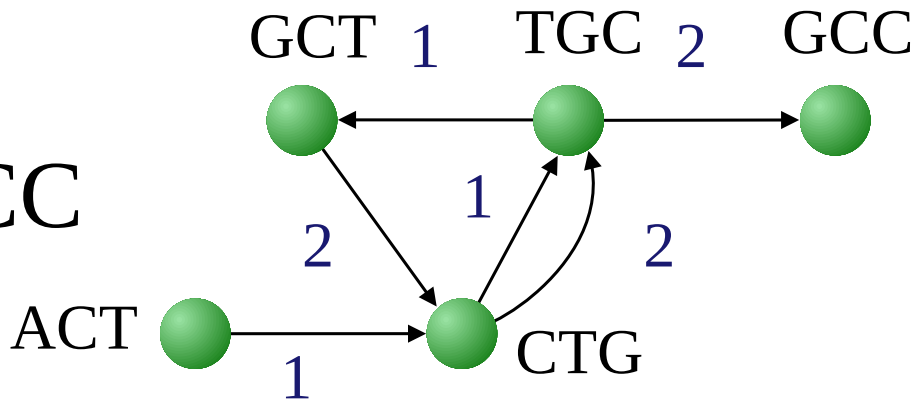
ACTGCTGCC



Classical SBH: Any solution corresponds to an Euler path in G_A.

The de-Bruijn Graph

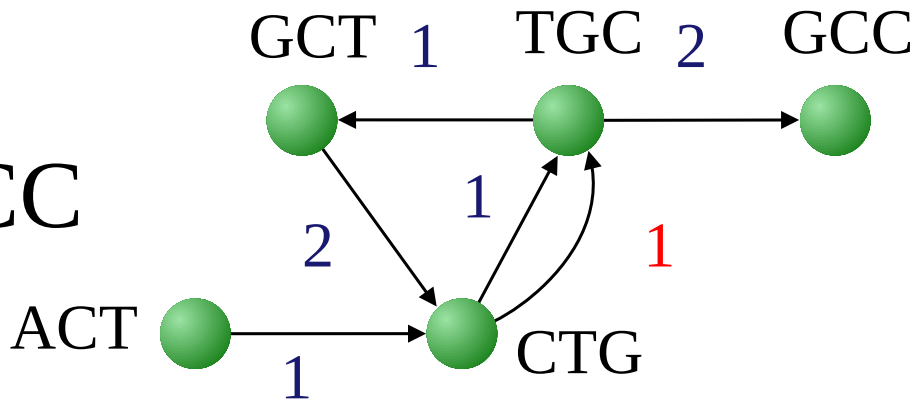
ACTGCTGCC



Shotgun SBH w/o errors: Each edge e_i has a label $l_i = \lceil \frac{i}{d} \rceil$ = the number of IF containing e_i. A solution corresponds to an Euler path in which each e_i is in the l_i-th IF (i.e. in $[(l_i - 1)d + 1, l_i d]$).

The de-Bruijn Graph

ACTGCTGCC



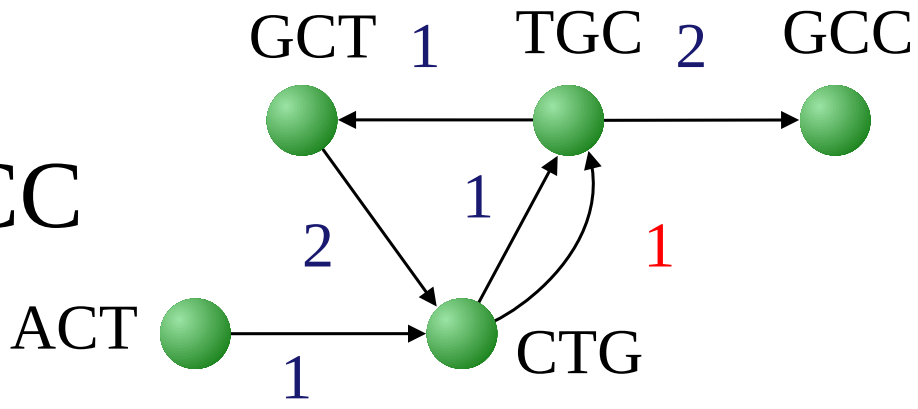
Shotgun SBH with errors:

l_i = number of IF containing e_i's sequence.

l'_i = max clone containing e_i's sequence. $l'_i \leq l_i$.

The de-Bruijn Graph

ACTGCTGCC



Shotgun SBH with errors:

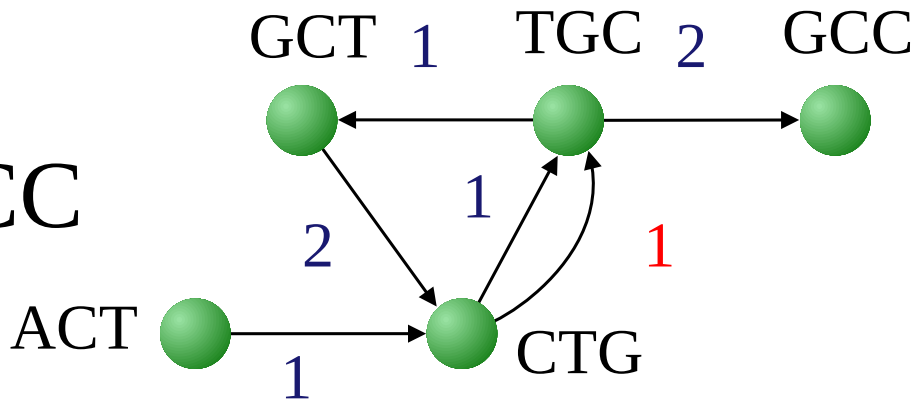
l_i = number of IF containing e_i's sequence.

l'_i = max clone containing e_i's sequence. $l'_i \leq l_i$.

- The distribution of $l_i - l'_i$ is geometric with parameter p.

The de-Bruijn Graph

ACTGCTGCC



Shotgun SBH with errors:

l_i = number of IF containing e_i's sequence.

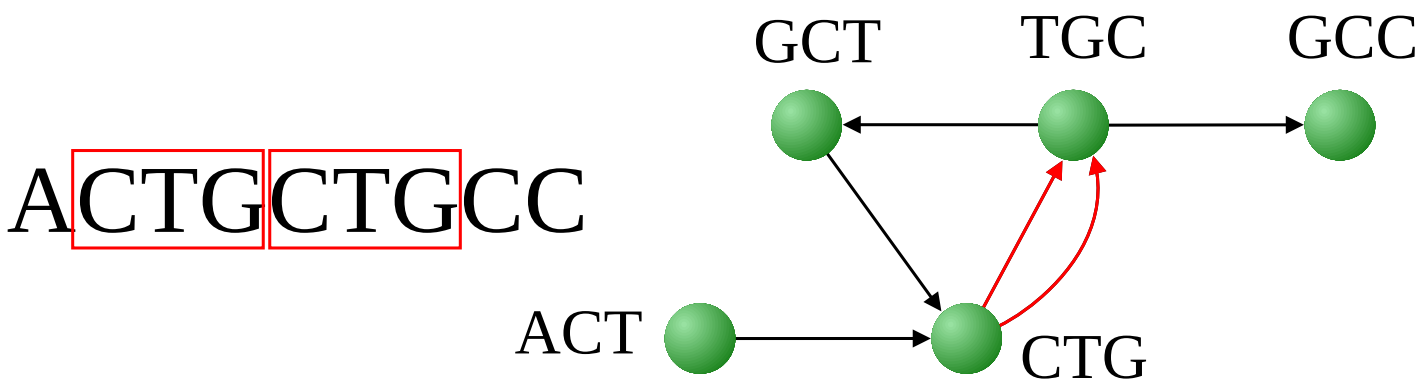
l'_i = max clone containing e_i's sequence. $l'_i \leq l_i$.

- The distribution of $l_i - l'_i$ is geometric with parameter p.
- A solution corresponds to an Euler path in which each e_i is in an IF with index $\geq l'_i$.

Definitions

Recall: A_i - the $(k-1)$-mer $a_i a_{i+1} \cdots a_{i+k-2}$.

A pair (i, j) is a repeat if $A_i = A_j$.



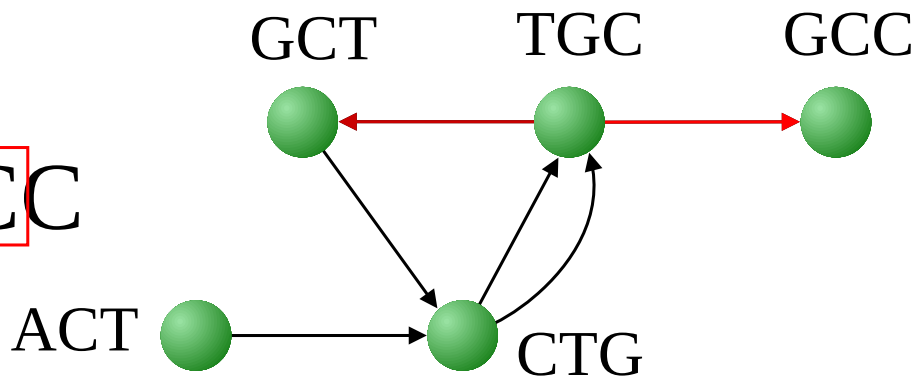
Definitions

Recall: A_i - the $(k-1)$-mer $a_i a_{i+1} \cdots a_{i+k-2}$.

A pair (i, j) is a repeat if $A_i = A_j$.

(i, j) is rightmost repeat if $(i+1, j+1)$ is not a repeat.

ACTGCTGCC

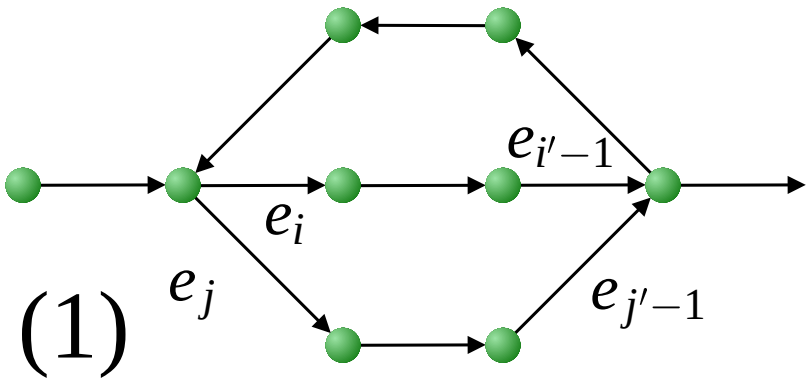


Failure Conditions

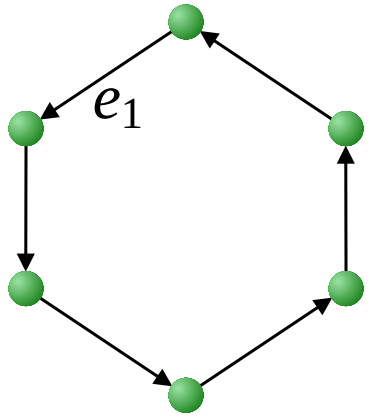
Interleaved pair of repeats: a rightmost repeat (i, j) and a repeat (i', j') with $i \leq i' < j < j'$.

Theorem (Pevzner 95) A sequence A is not uniquely recoverable iff either

1. A contains an interleaved pair of repeats, or
2. $A_1 = A_{n+1}$.



(1)



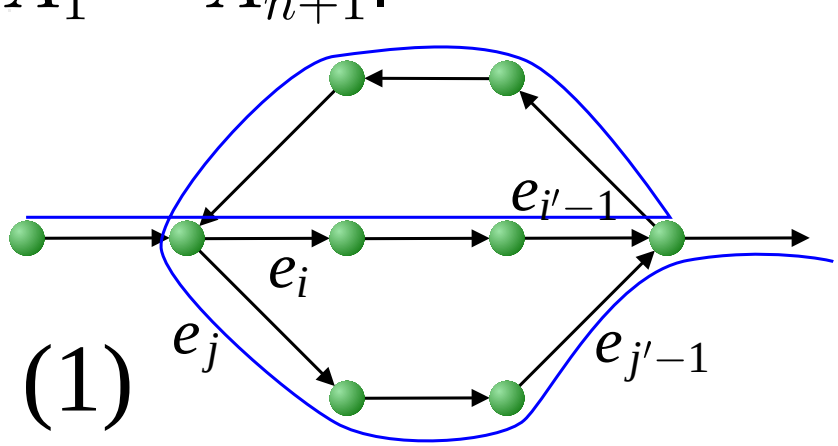
(2)

Failure Conditions

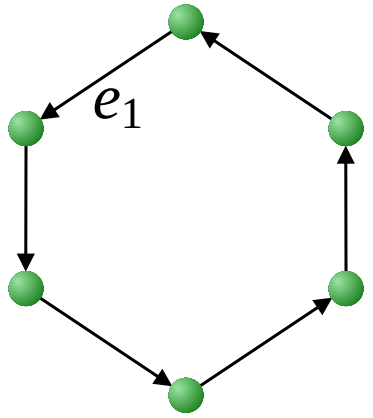
Interleaved pair of repeats: a rightmost repeat (i, j) and a repeat (i', j') with $i \leq i' < j < j'$.

Theorem (Pevzner 95) A sequence A is not uniquely recoverable iff either

1. A contains an interleaved pair of repeats, or
2. $A_1 = A_{n+1}$.



(1)



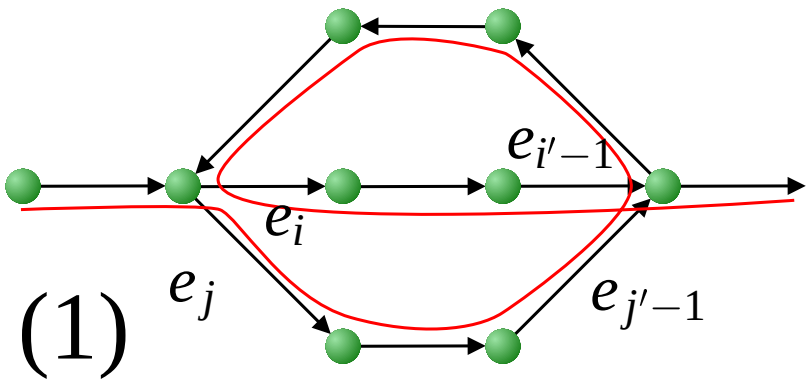
(2)

Failure Conditions

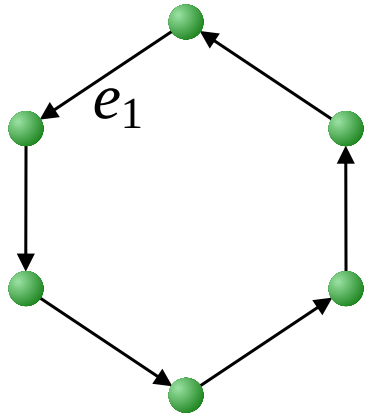
Interleaved pair of repeats: a rightmost repeat (i, j) and a repeat (i', j') with $i \leq i' < j < j'$.

Theorem (Pevzner 95) A sequence A is not uniquely recoverable iff either

1. A contains an interleaved pair of repeats, or
2. $A_1 = A_{n+1}$.



(1)

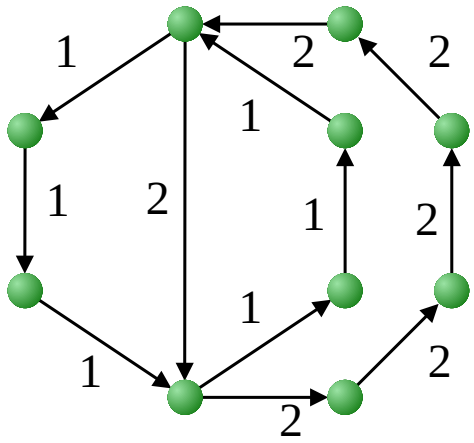


(2)

Failure Conditions - Shotgun SBH

Theorem A sequence A is not uniquely recoverable iff either

1. A contains an interleaved pair of repeats $(i, j)(i', j')$ with $l_i = l_{j'-1}$, or
2. $A_1 = A_{d+1} = \cdots = A_{cd+1}$ and $A_{i_1} = A_{i_2} = \cdots = A_{i_c} \neq A_1$ for indices $i_1, i_2, \ldots i_c$ with $l_{i_j} = j$ and $i_j \neq (j-1)d + 1$.



Failure Probability: The Errorless Case

Using the theorem we show that

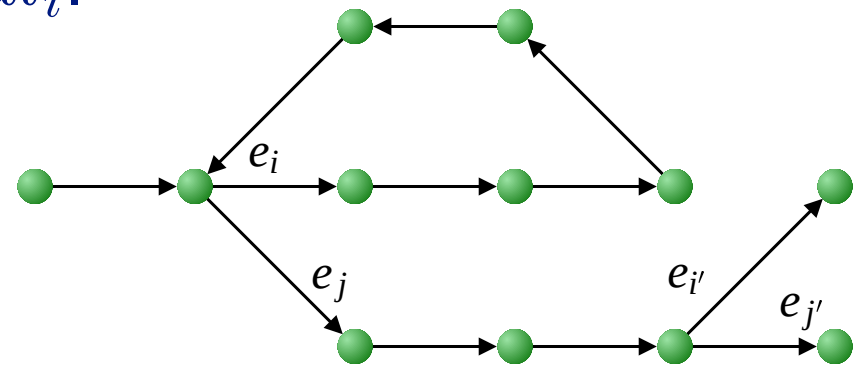
$$P(N,k,d,0) = \Theta(\frac{n}{d} \cdot \binom{d}{4} \cdot \frac{1}{4^{2k-2}}) = \Theta(\frac{d^3 n}{4^{2k}})$$

		Arratia et al.		Our bounds		
n	k	lower	upper	lower	upper	Simulation
193	8	0	0.5923	0.0051	0.1233	0.0907
791	10	0	0.2648	0.0083	0.1341	0.0996
3175	12	0.0502	0.1500	0.0094	0.1356	0.1009
12195	14	0.0742	0.1000	0.0084	0.1152	0.0875

Error-prone Spectra

Define event X: the solution is not unique when there are errors, but is unique in the errorless case.

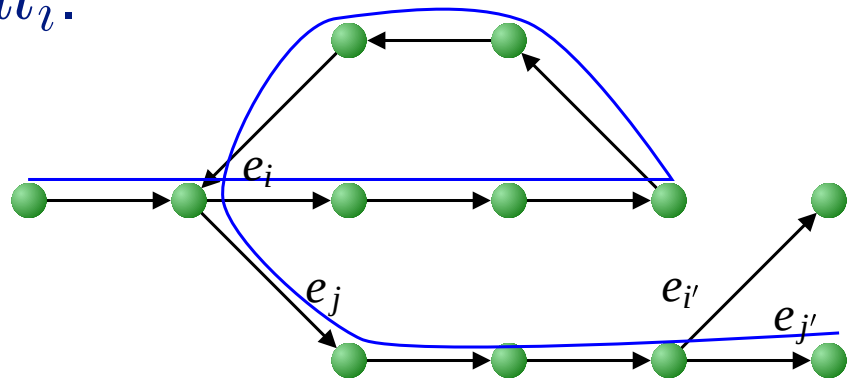
If event X happens, then A contains a rightmost repeat (i, j) and a repeat (i', j') with $i < j < j'$, $i' \notin [j, j']$, and $l_{i'} \geq l_i$ and either $l_i < l_{j'-1}$, or $j' - 1 = dl_i$.



Error-prone Spectra

Define event X: the solution is not unique when there are errors, but is unique in the errorless case.

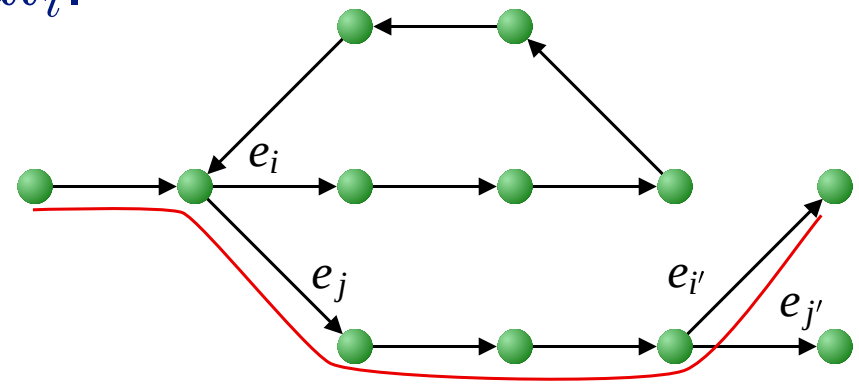
If event X happens, then A contains a rightmost repeat (i, j) and a repeat (i', j') with $i < j < j'$, $i' \notin [j, j']$, and $l_{i'} \geq l_i$ and either $l_i < l_{j'-1}$, or $j' - 1 = dl_i$.



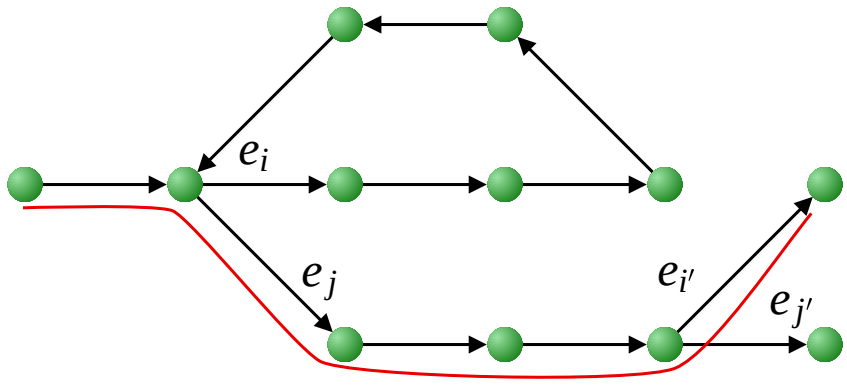
Error-prone Spectra

Define event X: the solution is not unique when there are errors, but is unique in the errorless case.

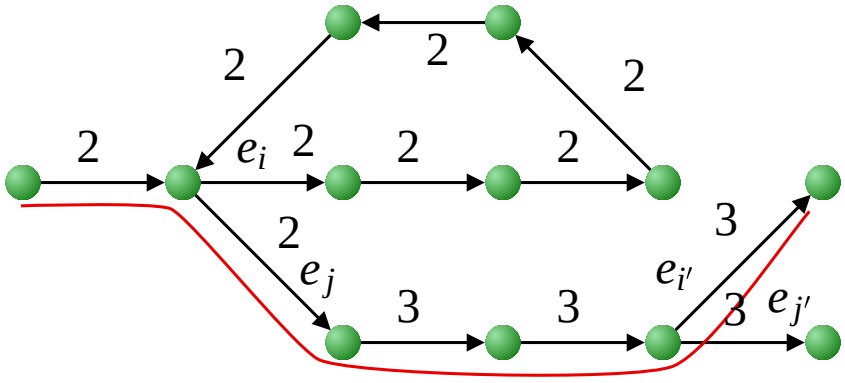
If event X happens, then A contains a rightmost repeat (i, j) and a repeat (i', j') with $i < j < j'$, $i' \notin [j, j']$, and $l_{i'} \geq l_i$ and either $l_i < l_{j'-1}$, or $j' - 1 = dl_i$.



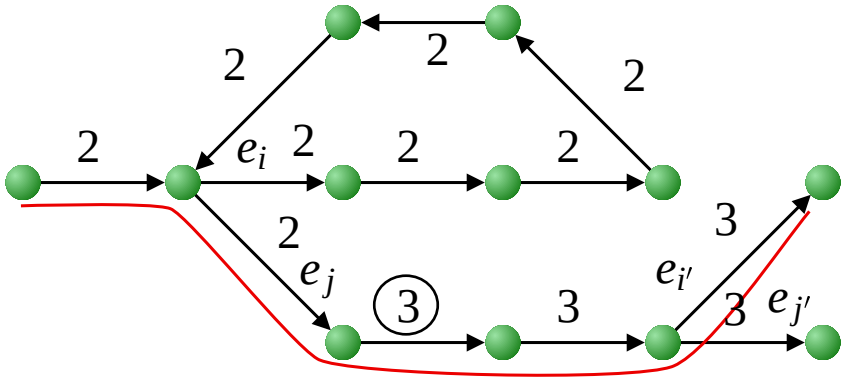
Error-prone Spectra



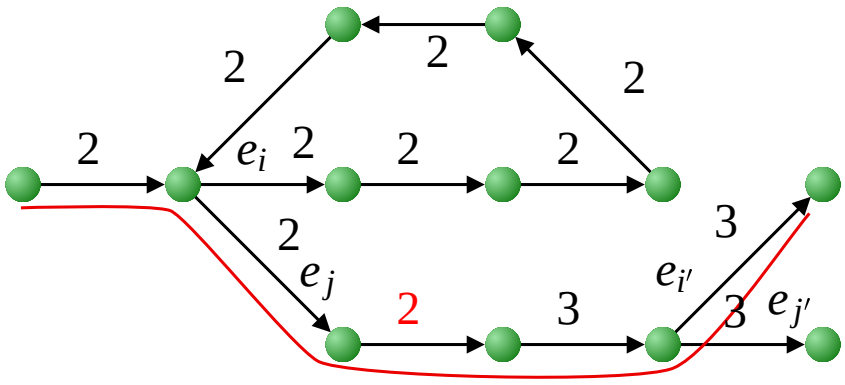
Error-prone Spectra



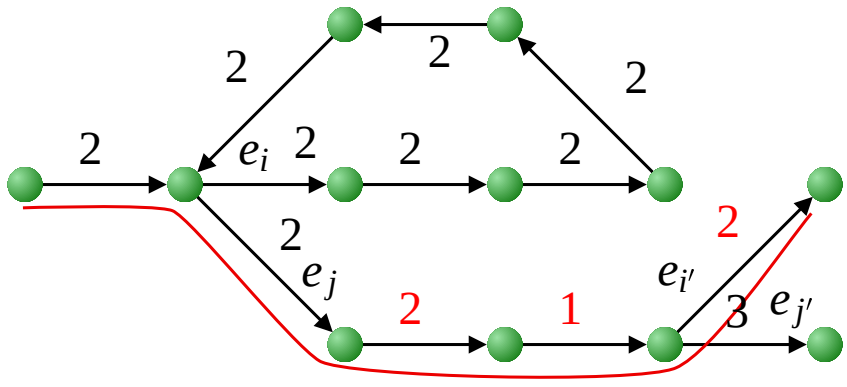
Error-prone Spectra



Error-prone Spectra



Error-prone Spectra



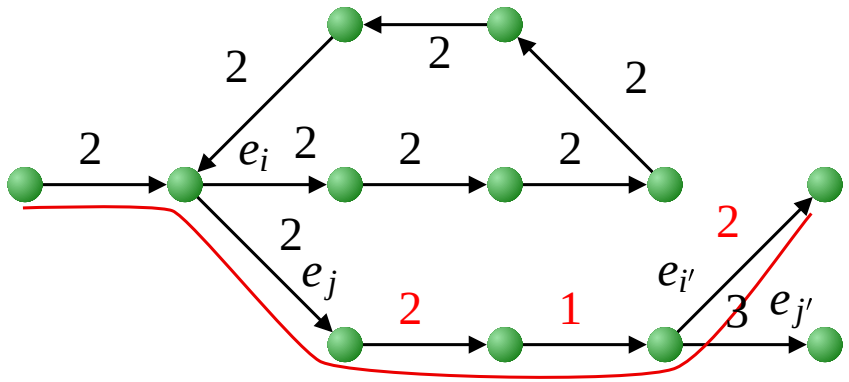
We need $l'_{j+1} \leq 2$, $l'_{j+2} \leq 2$, and $l'_{i'} \leq 2$.

The probability for these events is p^3.

Error-prone Spectra

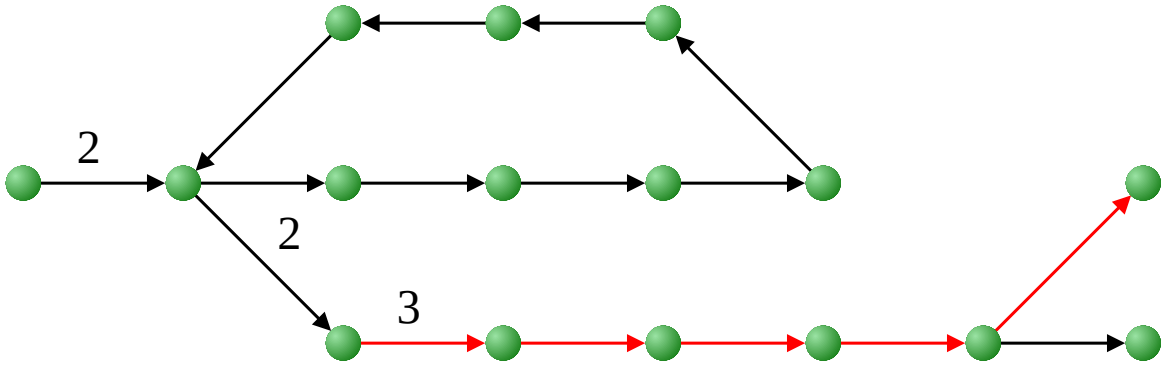
Theorem If event X happens, then A contains a rightmost repeat (i, j) and a repeat (i', j') with $i < j < j'$, $i' \notin [j, j']$, and $l_{i'} \geq l_i$ and either $l_i < l_{j'-1}$, or $j' - 1 = dl_i$.

Furthermore, $l'_r \leq l_{r-(j-i)}$ for all $j \leq r \leq j' - 1$, and $l'_{i'} \leq l_{j'-(j-i)}$.

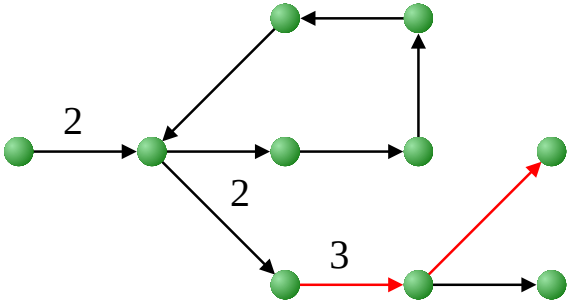


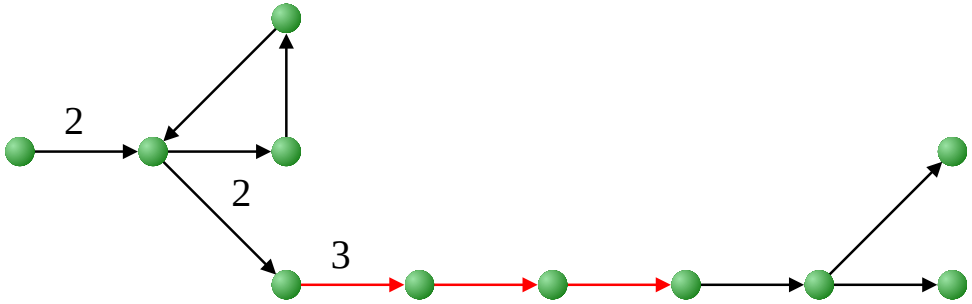
Error-prone Spectra

Low probability cases:



High probability cases:





Error-prone Spectra

Using the previous theorem, we can bound the probability that event X happen:

Theorem $P[X] = O(\frac{p}{(1-p)^4} \cdot \frac{n}{d} \cdot \frac{d^3}{4^{2k}})$.

Simulations

Generated data under the assumptions used for the theoretical analysis.

The Impact of d

n	k	d	$P(n,k,d,0)$ (%)	$P(n,k,d,0.5)$ (%)
7200	8	30	1.61	2.69
7200	8	40	3.67	5.20
7200	8	50	7.86	9.63
7200	8	60	12.85	15.45
7200	8	72	21.28	24.03
7200	8	80	27.08	30.36
7200	8	90	36.27	39.61
7200	8	100	46.12	49.46

The Impact of Errors

n	k	d	$P(n,k,d)$	$P(n,k,d,0.5)$	$\frac{P(n,k,d,0.5)}{P}$
18880	8	40	9.85	13.53	1.374
9550	8	50	10.25	12.60	1.229
5520	8	60	9.94	11.90	1.197
3500	8	70	9.79	11.14	1.138
2320	8	80	9.56	10.74	1.123
1620	8	90	9.03	10.06	1.114
1200	8	100	8.96	9.64	1.076
880	8	110	8.90	9.50	1.067

Variable Size IFs

IF sizes are Poisson distributed with expectation d.

			Prob(fail)		E(# errors)	
n	k	d	p=0	p=0.5	p=0	p=0.5
5000	9	40	3.8	4.6	1.11	1.39
10000	9	40	9.8	10.8	3.38	3.79
20000	9	40	15.8	19.2	5.50	6.93
30000	9	40	22.8	27.7	8.51	10.82
40000	9	40	31.6	36.0	13.67	16.11

Real DNA Sequences

			Prob(error)		Avg. # errors	
n	k	d	p=0	p=0.5	p=0	p=0.5
5000	9	40	40	40	5.7	6.9
10000	9	40	50	50	9.0	9.4
20000	9	40	80	80	13.4	15.0
30000	9	40	80	100	26.2	31.3

$\Rightarrow$ With 9-mer chip, can handle cosmid size target with 99.9% accuracy even with 50% false negative.

Summary

- Full analysis of failure prob. in errorless SBH: improves over Arratia et al. (96) for small k.
- Main result: Analysis of failure probability in Shotgun SBH, in the presence of errors.
- Errors have very little effect on failure probability.
- Simulation show result holds even when some of the assumptions are relaxed.

Open Problems

- Analyze the case of Poisson distribution of clone positions.
- Analyze the expected no. of errors.
- Relax independence assumption on errors.
- In simulation, compute the clones positions from the data.
- Handle false positives.

Open Problems

- Analyze the case of Poisson distribution of clone positions.
- Analyze the expected no. of errors.
- Relax independence assumption on errors.
- In simulation, compute the clones positions from the data.
- Handle false positives.

The End