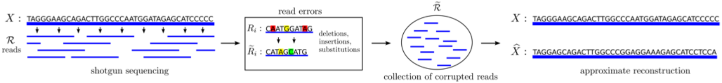
Sequence assembly from corrupted shotgun reads

Shirshendu Ganguly — U. Washington

Elchanan Mossel — UC Berkeley & U. Penn

Miklós Z. Rácz — Microsoft Research

ISIT 2016



X : TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC

$\mathcal{R}$ reads

shotgun sequencing

read errors

R_i : CAATGGATAG

$\widetilde{R}_i$: CATAGCATG

deletions, insertions, substitutions

$\widetilde{\mathcal{R}}$

collection of corrupted reads

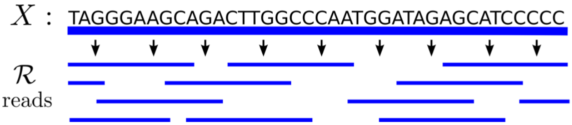
X : TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC

$\widehat{X}$: TAGGAGCAGACTTGGCCCGGAGGAAAGAGCATCCTCCA

approximate reconstruction

DNA sequencing:

- Prevalent technique: shotgun sequencing

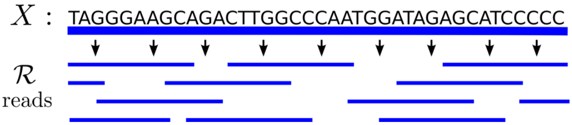


- Goal of de novo assembly:

  reconstruct X from reads $\mathcal{R}$

DNA sequencing: robust algorithms?

- Prevalent technique: shotgun sequencing



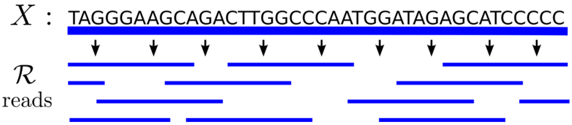
- Goal of de novo assembly:
  reconstruct X from reads $\mathcal{R}$

- Many sequencing technologies
  w/ different error profiles

Q: Are there robust assembly algorithms?

DNA sequencing: robust algorithms

- Prevalent technique: shotgun sequencing



- Goal of de novo assembly:

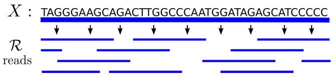
  reconstruct X from reads $\mathcal{R}$

- Many sequencing technologies w/ different error profiles

Q: Are there robust assembly algorithms?

A: Yes, and a simple sequential algorithm works well.

Sequencing technologies



- **Sanger sequencing**
  - 800-1000 bp reads, <1% error
  - very expensive

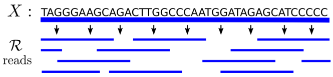
- **Next gen (2nd gen) sequencing**
  - high throughput, cheap
  - short reads (100-200 bp)
  - low error rate (1-3%)

Example:
Illumina

- **Emerging (3rd gen) technologies**
  - long reads (> 10 000 bp)
  - high error rate (10-22%)

Examples:
- PacBio's SMRT
- Oxford Nanopore

Sequencing technologies



- **Sanger sequencing**
  - 800-1000 bp reads, <1% error
  - very expensive

- **Next gen (2nd gen) sequencing**
  - high throughput, cheap
  - short reads (100-200 bp)
  - low error rate (1-3%)

Example:
Illumina

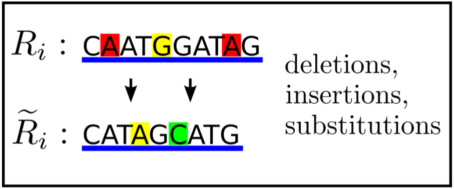
- **Emerging (3rd gen) technologies**
  - long reads (>10 000 bp)
  - high error rate (10-22%)

Examples:
- PacBio's SMRT
- Oxford Nanopore

Q: how robust are reconstruction algorithms w.r.t. different sequencing technologies?

Adversarial corruption/error model

- Instead of getting true reads $\mathcal{R}$,
  get corrupted reads $\widetilde{\mathcal{R}}$



R_i : CAATGGATAG

$\widetilde{R}_i$: CATAGCATG

deletions,
insertions,
substitutions

- Assume only that

$$\boxed{ed(R_i, \widetilde{R}_i) \leqslant \varepsilon L}$$

where ed = edit distance,
and L = length of R_i.

Approximate reconstruction problem

1. Choose $X \in \Sigma^n$ uniformly at random, $\Sigma = \{A, C, G, T\}$

2. Draw reads

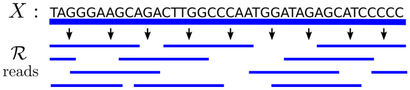
   $\mathcal{R} = \{R_1, R_2, \ldots, R_N\}$ of length L

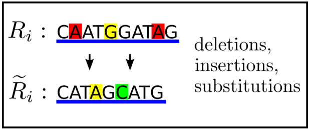
   from uniformly random positions

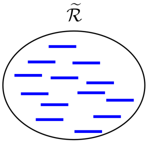
3. Get corrupted reads

   $\tilde{\mathcal{R}} = \{\tilde{R}_1, \tilde{R}_2, \ldots, \tilde{R}_N\}$

   satisfying $\boxed{ed(R_i, \tilde{R}_i) \leqslant \varepsilon L}$







collection of corrupted reads

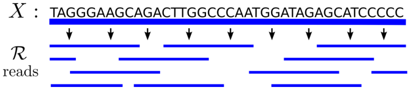
Approximate reconstruction problem

1. **Choose $X \in \Sigma^n$** uniformly at random, $\Sigma = \{A, C, G, T\}$

2. **Draw reads**

   $\mathcal{R} = \{R_1, R_2, \dots, R_N\}$ of length L

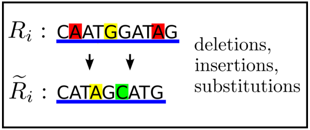
   from uniformly random positions

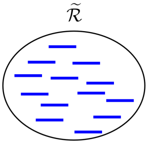


3. **Get corrupted reads**

   $\tilde{\mathcal{R}} = \{\tilde{R}_1, \tilde{R}_2, \dots, \tilde{R}_N\}$

   satisfying $\boxed{ed(R_i, \tilde{R}_i) \leq \varepsilon L}$





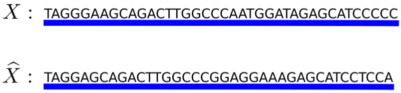
collection of corrupted reads

$$\boxed{\text{Goal: approximate reconstruction}}$$

Output: $\hat{X} = \hat{X}(\tilde{\mathcal{R}}) \in \Sigma^*$ s.t.

$$ed(\hat{X}, X) \leq C\varepsilon n$$

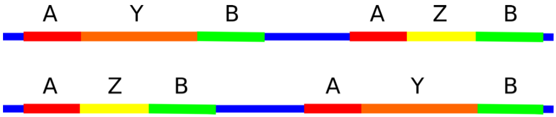
w/prob. $\geqslant 1 - \delta$.


X : TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC
$\widehat{X}$: TAGGAGCAGACTTGGCCCGGAGGAAAGAGCATCCTCCA

approximate reconstruction

Main obstructions to reconstruction

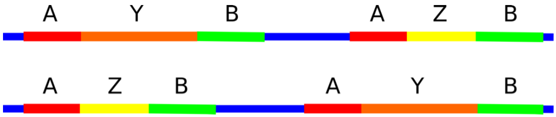
1. Short reads lead to repeats (Ukkonen '92)



repeat-limited regime

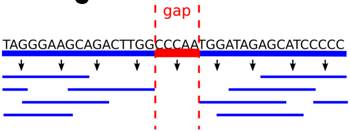
Main obstructions to reconstruction

1. Short reads lead to repeats (Ukkonen '92)



repeat-limited regime

2. Need enough reads to cover X



coverage-limited regime

Lander, Waterman (1988):

$$N_{cov} = N_{cov}(n, L, \delta) \approx \frac{n}{L} \ln\left(\frac{n}{L\delta}\right)$$

Exact reconstruction ($\varepsilon=0$)

Thm (A. Motahari, G. Bresler, D. Tse, 2013)

These are the only obstructions.

Exact reconstruction ($\varepsilon = 0$)

Thm (A. Motahari, G. Bresler, D. Tse, 2013)

These are the only obstructions.

More precisely: let X be random, $L = \bar{L} \ln(n)$, $\delta < 1/2$.

Then:

repeat-limited

- if $\bar{L} < \frac{2}{\ln|\Sigma|}$ then exact reconstruction is impossible;

coverage-limited

- if $\bar{L} > \frac{2}{\ln|\Sigma|}$ then $\lim\limits_{n \to \infty} \frac{N_{min}}{N_{cov}} = 1$.

Exact reconstruction ($\varepsilon=0$)

Thm (A. Motahari, G. Bresler, D. Tse, 2013)

These are the only obstructions.

More precisely: let X be random, $L = \bar{L}\ln(n)$, $\delta < 1/2$.

Then:

repeat-limited

- if $\bar{L} < \frac{2}{\ln|\Sigma|}$ then exact reconstruction is impossible;

coverage-limited

- if $\bar{L} > \frac{2}{\ln|\Sigma|}$ then $\lim\limits_{n\to\infty} \frac{N_{min}}{N_{cov}} = 1$.

For arbitrary sequences:

G. Bresler, M. Bresler, D. Tse (2013)

thresholds based on repeat statistics of genome

Approximate reconstruction

Thm Approximate reconstruction is possible, if L and N are large enough.

Approximate reconstruction

Thm Approximate reconstruction is possible, if L and N are large enough.

More precisely: Let X be random, and $L = \bar{L} \ln(n)$.

For every $C > 3$ there exist constants $\bar{C} = \bar{C}(\Sigma)$, $\varepsilon_0 = \varepsilon_0(\Sigma, C)$, $C' = C'(\Sigma, C)$ s.t. for every $\varepsilon \in (0, \varepsilon_0)$ if $\bar{L} \geq \bar{C}/\varepsilon$, $N \geq C' N_{cov}/\varepsilon$ then there exists an approximate reconstruction algorithm for error rate ε with approximation factor C.

Approximate reconstruction

Thm Approximate reconstruction is possible, if L and N are large enough.

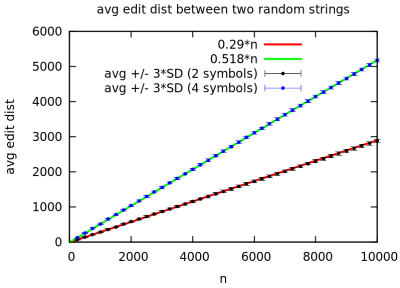
More precisely: Let X be random, and $L = \bar{L}\ln(n)$.
For every $C > 3$ there exist constants $\bar{C} = \bar{C}(\Sigma)$, $\varepsilon_0 = \varepsilon_0(\Sigma, C)$, $C' = C'(\Sigma, C)$ s.t. for every $\varepsilon \in (0, \varepsilon_0)$ if $\bar{L} \geq \bar{C}/\varepsilon$, $N \geq C' N_{cov}/\varepsilon$ then there exists an approximate reconstruction algorithm for error rate ε with approximation factor C.

Comments

- simple sequential algorithm works
- dependence of $\bar{L}$ and N on ε not necessary (but get worse C)
- best achievable C might depend on $\bar{L}$ and N
- related work:

Motahari, Ramchandran, Tse, Ma (2013); Shomorony, Courtade, Tse (2015)

Edit distance between random strings



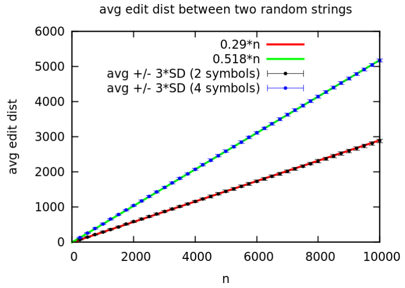
<u>Lemma</u> $X_m, Y_m \in \Sigma^m$ independent, uniformly random. Then

$$\lim \frac{1}{m} ed(X_m, Y_m) = c_{ind} > 0.$$

For $|\Sigma| = 4$:

- empirically $c_{ind} \approx 0.51$.
- volume argument: $c_{ind} > 0.33$.

Edit distance between random strings

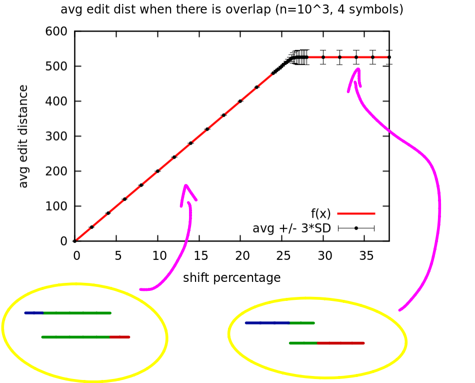


Lemma $X_m, Y_m \in \Sigma^m$ independent, uniformly random. Then

$$\lim \frac{1}{m} ed(X_m, Y_m) = c_{ind} > 0.$$

For $|\Sigma| = 4$:

- empirically $c_{ind} \approx 0.51$.
- volume argument: $c_{ind} > 0.33$.

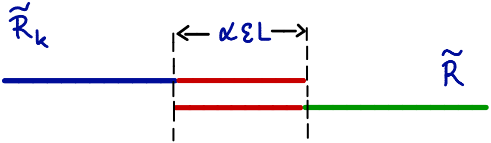


Lemma $X \in \Sigma^{2m}$ uniformly random. Then:

$$ed\big(X[1,m], X[1+k, m+k]\big) = 2k$$

for all $k \leq cm$ with prob. $\geq 1 - e^{-c'm}$.

Sequential reconstruction algorithm



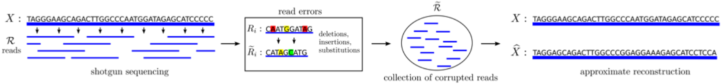
- Fix α appropriately.
- Given $\tilde{R}_k$, find $\tilde{R} \in \tilde{\mathcal{R}}$ s.t.

$$ed\left(\tilde{R}_k^{\text{suffix}}, \tilde{R}^{\text{prefix}}\right) \leqslant (2+2/c')\varepsilon L$$

- concatenate $\tilde{R}_k$ and $\tilde{R}^{\text{suffix}}$.
- at each step, gain $\approx (1-\alpha\varepsilon)L$, make error $\lesssim 3\varepsilon L$

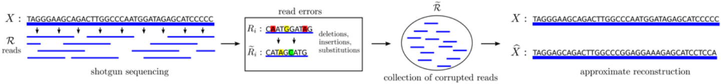
Summary

- introduced adversarial read error model
- approximate reconstruction is possible
- simple sequential algorithm works
- edit distance results key to analysis


X : TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC
$\mathcal{R}$
reads
shotgun sequencing
read errors
R_i : CAATGGATAG
$\widetilde{R}_i$: CATAGCATG
deletions, insertions, substitutions
$\widetilde{\mathcal{R}}$
collection of corrupted reads
X : TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC
$\widehat{X}$: TAGGAGCAGACTTGGCCCGGAGGAAAGAGCATCCTCCA
approximate reconstruction

Summary

- introduced adversarial read error model
- approximate reconstruction is possible
- simple sequential algorithm works
- edit distance results key to analysis



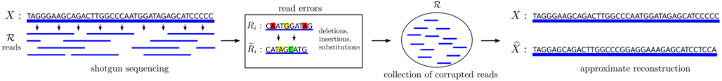
X : TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC
$\mathcal{R}$
reads
shotgun sequencing
read errors
R_i : CAATGGATAG
$\widetilde{R}_i$: CATAGCATG
deletions, insertions, substitutions
$\widetilde{\mathcal{R}}$
collection of corrupted reads
X : TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC
$\widehat{X}$: TAGGAGCAGACTTGGCCCGGAGGAAAGAGCATCCTCCA
approximate reconstruction

Challenges

- determine fundamental limits of approximate reconstruction
- results for arbitrary sequences
- bridge gap between models

Summary

- introduced adversarial read error model
- approximate reconstruction is possible
- simple sequential algorithm works
- edit distance results key to analysis



X : TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC
$\mathcal{R}$ reads
shotgun sequencing
read errors
R_i : CAATGGATAG
$\widetilde{R}_i$: CATAGCATG
deletions, insertions, substitutions
$\widetilde{\mathcal{R}}$
collection of corrupted reads
X : TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC
$\widehat{X}$: TAGGAGCAGACTTGGCCCGGAGGAAAGAGCATCCTCCA
approximate reconstruction

Challenges

- determine fundamental limits of approximate reconstruction
- results for arbitrary sequences
- bridge gap between models

Thank you!