Sequence assembly from corrupted shotgun reads

Shirshendu Ganguly  Elchanan Mossel  Miklós Z. Rácz
U. Washington  UC Berkeley & U. Penn  Microsoft Research

$X: TAGGAAGGAGACGTTGGGCTCCACATGAGAAGCCTCCCC$
$
\hat{R}_c$ reads  $\hat{R}_v: CACGATAG$  $\hat{R}_l: CACGAGNT$

shotgun sequencing  read errors  collection of corrupted reads

$X: TAGGAAGGAGACGTTGGGCTCCACATGAGAAGCCTCCCC$
$\hat{R}: CACGAGNT$
$\hat{X}: TAGGAAGGACGTTCGGCTCCACATGAGAAGCCTCCCC$

approximate reconstruction
DNA Sequencing

- Prevalent technique: shotgun sequencing

\[ X : \text{TAGGGAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC} \]

reads

- Goal of de novo assembly: reconstruct \( X \) from reads \( R \)
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%)

• **Emerging (3rd gen) technologies**
  - Long reads (>10,000 bp)
  - High error rate (10-22%)

Example: Illumina

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%)

• Emerging (3rd gen) technologies
  - Long reads (>10,000 bp)
  - High error rate (10-22%)

Q: How robust are reconstruction algorithms w.r.t. different sequencing technologies?
Adversarial corruption/error model

- Instead of getting true reads $R_i$, get corrupted reads $\tilde{R}_i$

\[
\begin{align*}
R_i & : \textcolor{blue}{\text{CAATGGATAG}} \\
\tilde{R}_i & : \textcolor{red}{\text{CATAGATG}}
\end{align*}
\]

- Assume only that

\[\text{ed}(R_i, \tilde{R}_i) \leq EL\]

where $\text{ed} = \text{edit distance}$, and $L = \text{length of } R_i$. 
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 $\text{ed}(R_i, \tilde{R}_i) \leq \epsilon L$

$X :$ TAGGAAAGCAGACTTGGCCCAATGGATAGAGCATCCCCC

$\mathcal{R}$ reads

$R_i :$ CAATGGATAG

$\tilde{R}_i :$ CATAGCATG

deletions, insertions, substitutions

collection of corrupted reads
Approximate reconstruction problem

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

2. Draw reads
\( \mathcal{R} = \{R_1, R_2, \ldots, R_N\} \) 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 \( \text{ed}(R_i, \tilde{R}_i) \leq 3L \)

Goal: approximate reconstruction

Output: \( \hat{X} = \hat{X}(\tilde{\mathcal{R}}) \in \Sigma^* \) s.t.
\( \text{ed}(\hat{X}, X) \leq C \varepsilon n \)

w/ prob. \( \geq 1 - \delta \).
Main obstructions to reconstruction

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

Diagram:

A Y B A Z B

A Z B A Y B

repeat-limited regime
Main obstructions to reconstruction

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

2. Need enough reads to cover X

Lander, Waterman (1988):

\[ N_{cov} = N_{cov}(n, L, \delta) \approx \frac{n}{L} \ln \left( \frac{n}{L \delta} \right) \]
Exact reconstruction ($\varepsilon = 0$)

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

These are the only obstructions.
Exact reconstruction ($\varepsilon=0$)

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

These are the only obstructions.

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

- if $\bar{L} < \frac{2}{\ln|\Sigma|}$ then exact reconstruction is impossible;
- if $\bar{L} > \frac{2}{\ln|\Sigma|}$ then $\lim_{n \to \infty} \frac{N_{\text{min}}}{N_{\text{cov}}} = 1$. 
Exact reconstruction ($\varepsilon = 0$)

*Thm.* (A. Motahai, G. Bresler, D. Tse, 2013)

These are the only obstructions.

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

- if $\bar{L} < \frac{2}{\ln|\Xi|}$ then exact reconstruction is impossible;
- if $\bar{L} > \frac{2}{\ln|\Xi|}$ then $\lim_{n \to \infty} \frac{N_{\min}}{N_{\cov}} = 1$.

For arbitrary sequences:

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

thresholds based on repeat statistics of genome
Approximate reconstruction is possible, if $L$ and $N$ are large enough.
Approximate reconstruction

Thus, approximate reconstruction is possible, if \( L \) and \( N \) are large enough.

More precisely: Let \( X \) be random, and \( L = I \ln(n) \).

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 \( L \geq \bar{C}/\varepsilon \), \( N \geq C'N_{\text{cov}}/\varepsilon \)

then there exists an approximate reconstruction algorithm for error rate \( \varepsilon \) with approximation factor \( C \).
Approximate reconstruction

Thus, approximate reconstruction is possible, if $L$ and $N$ are large enough.

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

For every $C > 3$ there exist constants $\overline{C} = \overline{C}(\Sigma)$, $\epsilon_0 = \epsilon_0(\Sigma, C)$, $C' = C'(\Sigma, C)$ s.t. for every $\epsilon \in (0, \epsilon_0)$ if $\overline{L} \geq \overline{C}/\epsilon$, $N \geq C' N_{conv}/\epsilon$

then there exists an approximate reconstruction algorithm for error rate $\epsilon$ with approximation factor $C$.

Comments

- Simple sequential algorithm works
- dependence of $\overline{L}$ and $N$ on $\epsilon$ not necessary
  (but get worse $C$)
- best achievable $C$ might depend on $\overline{L}$ and $N$
- related work:
  Motahari, Ramchandran, Tse, Ma (2013); Shomorony, Courtade, Tse (2015)
**Lemma** \( X_m, Y_m \in \Sigma^m \) independent, uniformly random. Then
\[
\lim_{m \to \infty} \frac{1}{m} \text{ed} (X_m, Y_m) = c_{\text{ind}} > 0.
\]

For \(|\Sigma| = 4\):
- empirically \( c_{\text{ind}} \approx 0.51\).
- volume argument: \( c_{\text{ind}} > 0.33\).
**Lemma** \( X_m, Y_m \in \Sigma^m \) independent, uniformly random. Then
\[
\lim \frac{1}{m} \text{ed}(X_m, Y_m) = c_{\text{ind}} > 0.
\]

For \(|\Sigma| = 4\):
\[
\begin{align*}
&\text{empirically } c_{\text{ind}} \approx 0.51. \\
&\text{volume argument: } c_{\text{ind}} > 0.33.
\end{align*}
\]

**Lemma** \( X \in \Sigma^{2m} \) uniformly random. Then:
\[
\text{ed}(X[1,m], X[1+k, m+k]) = 2k
\]
for all \( k \leq cm \) with prob. \( \geq 1 - e^{-c'm} \).
Sequential reconstruction algorithm

- Fix $\alpha$ appropriately.
- Given $\tilde{R}_k$, find $\bar{R} \in \bar{R}$ s.t.
  \[
ed(\tilde{R}_k^{\text{suffix}}, \tilde{R}_k^{\text{prefix}}) \leq (2+2/c')3L\]
- Concatenate $\tilde{R}_k$ and $\tilde{R}^{\text{suffix}}$.
- At each step, gain $\approx (1-\delta)L$, make error $\leq 3\delta L$.
Summary

- introduced adversarial read error model
- approximate reconstruction is possible
- simple sequential algorithm works
Summary

• introduced adversarial read error model
• approximate reconstruction is possible
• simple sequential algorithm works

Challenges

• determine fundamental limits of approximate reconstruction
• results for arbitrary sequences
• bridge gap between models
• heterogeneous error rates
Summary

- introduced adversarial read error model
- approximate reconstruction is possible
- simple sequential algorithm works

Challenges

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

Thank you!