Repeat Resolution

Nicolas Balcazar    Corinna Blasse    An Duc Dang
Hannes Hauswedell   Sebastian Thieme

Advanced Algorithms for Bioinformatics (P4)
K. Reinert and S. Andreotti

SoSe 2010
FU Berlin

June 16, 2010
Outline

Introduction

Locating

Bounding

Identifying DNPs
  Tammi’s algorithm
  Approach with iid deviations
  Approach with non-iid deviations

Separating repeat copies
  Algorithm of Kececioglu
  ILP
Outline

Introduction
Locating
Bounding
Identifying DNPs
- Tammi’s algorithm
- Approach with iid deviations
- Approach with non-iid deviations
Separating repeat copies
- Algorithm of Kececioglu
- ILP

N. Balcazar, C. Blasse, D. Dang, H. Hauswedell, S. Thieme
RNA-Seq
Read-Mapping

N. Balcazar, C. Blasse, D. Dang, H. Hauswedell, S. Thieme

RNA-Seq
Multi-Reads

- different copies of same/similar gene
- similar sub-sequence due to similar domain/motif
- ...
Multi-Reads

- Is there a pattern?
- certain eye-catching columns: "Defined Nucleotide Position"
- possible to separate reads
Multi-Reads

▶ Is there a pattern?
▶ certain eye-catching columns: “Defined Nucleotide Position”
▶ possible to separate reads

N. Balcazar, C. Blasse, D. Dang, H. Hauswedell, S. Thieme

RNA-Seq
Multi-Reads

- Is there a pattern?
- certain eye-catching columns: “Defined Nucleotide Position”
- possible to separate reads
Multi-Reads

- Is there a pattern?
- certain eye-catching columns: “Defined Nucleotide Position”
- possible to separate reads
Tasks

Algorithmic tasks to resolve the aforementioned problem include:

1. **Locating** positions with ambiguous maps
2. **Bounding** the region for the analysis of DNPs.
3. **Identifying** the DNPs.
4. **Separating** the set of reads into subsets belonging to different instances of the repeated sequence
Outline

Introduction

Locating

Bounding

Identifying DNPs
  Tammi’s algorithm
  Approach with iid deviations
  Approach with non-iid deviations

Separating repeat copies
  Algorithm of Kececioglu
  ILP
The idea

Find locations of possible repeats
Why look for repeat regions

Why not analyze whole genome?
→ runtime reduction!
Why look for repeat regions

Why not analyze whole genome?
→ runtime reduction!
The idea

Unexpectedly high coverage → Repeats
How does it look like?
How to find locations of unexpectedly high coverage?

Comparison to a binomial distribution.
Outline

Introduction

Locating

Bounding

Identifying DNPs
  Tammi’s algorithm
  Approach with iid deviations
  Approach with non-iid deviations

Separating repeat copies
  Algorithm of Kececioglu
  ILP

N. Balcazar, C. Blasse, D. Dang, H. Hauswedell, S. Thieme

RNA-Seq
The idea

separate repeat regions by multiple alignment

...AGCGGTCAGA...
...AGCGGTCAGA...
...AGCCCTCTGA...
...TGTCGTCTGAGA...
...AGTCGTCTCA...
...AGTCGTCTGAGA...
Problem

Narrow window $\rightarrow$ not enough distinguishing base sites (DNPs)
Wide window $\rightarrow$ too few fragments

N. Balcazar, C. Blasse, D. Dang, H. Hauswedell, S. Thieme

RNA-Seq
Problem

Narrow window $\rightarrow$ not enough distinguishing base sites (DNPs)
Wide window $\rightarrow$ too few fragments
Finding an optimal window

What is the best window that both separates many fragments and has sufficiently many distinguishing base sites?

i chose one read out of a repeat region (starting read)  
ii multialign with 1st- and 2nd-order overlapping reads  
iii analyze (step 3 and 4)  
iv mark starting read and 1st-order overlaps as analyzed  
v pick new starting read from unanalyzed reads  
vi repeat until all reads are marked as analyzed
Outline

Introduction

Locating

Bounding

Identifying DNPs
  Tammi’s algorithm
  Approach with iid deviations
  Approach with non-iid deviations

Separating repeat copies
  Algorithm of Kececioglu
  ILP

N. Balcazar, C. Blasse, D. Dang, H. Hauswedell, S. Thieme

RNA-Seq
Idea of Tammi’s algorithm

- after computation of multialignment determine DNP
- at each column assume most frequent base \(b\) as consensus base
- DNP: column with \(\geq D_{\text{min}}\) non consensus bases of a same type \(b_1\)
Idea of Tammi’s algorithm

- one DNP not enough to separate $\rightarrow$ support column of a DNP
- two methods:
  - basic: find column pairs having $\geq D_{min}$ non consensus bases in same rows of alignment
  - extended: compute probability of observing coinciding deviations by chance
Idea of Tammi’s algorithm

- e.g.: blue columns are DNPs with $D_{\text{min}} = 2$
- 1st blue col has 3 deviating bases of type $C$
- 2nd blue col is supporting col with 2 deviating bases of type $A$

```plaintext
...AGCCGTTCAGA...
...AGCCGTTCAGA...
...AGCCCTCTTGA...
...TGTCGTCTTGA...
...AGTCGTCTCA...
...AGTCGTCTTGA...
```
Mathematical definitions

- fixed positions in alignment: $u, v$
- base at $u$ in $j$th sequence: $a_{u,j}$
- indicator that $a_{u,j}$ deviates from consensus: $l_{u,j}$
- if deviation, then $l_{u,j} = 1$ and $l_{u,j} = 0$ otherwise
- probability of deviation at $u$ in $j$: $p_{u,j} = P(l_{u,j} = 1)$
Mathematical definitions

- total number of deviations at $u$ in $j$: $N_u = \sum_{j=1}^{k} I_{u,j}$
- all definitions analogue for $v$
- indicator for coincidence in $j$: $l_j = I_{u,j} I_{v,j}$
- $l_j = 1$ if deviation at both positions in $j$ and $l_j = 0$ otherwise
- total number of coincidences: $C = \sum_{j=1}^{k} l_j = \sum_{j=1}^{k} I_{u,j} I_{v,j}$
But ...

The quality values of the base recall is neglected!
Approach with iid deviations

Recap of hypergeometric distribution

- assume deviations are iid (not always true but approximates quite well):
  \[ p_{u,1} = \ldots = p_{u,k}, \quad p_{v,1} = \ldots = p_{v,k} \]
  → use hypergeometric distribution

- prob. for number of successes in \( n \) draws from a finite population \textit{without} replacement
Recap of hypergeometric distribution

toy example:

- $N$ balls in bucket: $M$ white, $N - M$ black
- draw $n$ balls without returning, then prob. drawing $k$ white balls is:

$$P_k(N, M, n) = \binom{M}{k} \binom{N-M}{n-k} \binom{N}{n}$$
Approach with iid deviations

Application of hypergeometric distribution

- analogously derive distribution $C$ given $N_u = n_u$, $N_v = n_v$:
  
  $$P(C = x) = \binom{n_v}{x} \binom{k-n_v}{n_u-x} \binom{k}{n_u},$$

  \[0 \leq x \leq n_v, \ 0 \leq n_u-x \leq k-n_v\]

- in reality deviations do not occur iid $\rightarrow$ non-iid approach
Approach with non-iid deviations

- For different $p$s each deviation has separate prob. and expectation of $C$ using hypergeometric distr.
- Assume dev. prob. different, but very small -> use Poisson distr.
- Since conditioning on $N_u$ and $N_v$ introduces weak dependencies Poisson distr. still approximates well
- Goal: compute mean of Poisson $E(C|N_u = n_u, N_v = n_v)$
Poisson approximation

- from definition $C = \sum_{j=1}^{k} I_{u,j} I_{v,j}$ follows:
  
  $E(C|N_u = n_u, N_v = n_v) = \sum_{j=1}^{k} E(I_{u,j} = 1|N_u = n_u)E(I_{v,j} = 1|N_v = n_v)$

- from definition of conditional prob. follows:

  $$P(I_{u,j} = 1|N_u = n_u) = \frac{P(I_{u,j}=1,N_u=n_u)}{P(N_u=n_u)}$$

  $$= \frac{P(I_{u,j}=1,N_{u}^{(j)}=n_u-1)}{P(I_{u,j}=1,N_{u}^{(j)}=n_u-1)+P(I_{u,j}=0,N_{u}^{(j)}=n_u)}$$

  where $N_{u}^{(j)} = N_u - I_{u,j}$: number of deviations at $u$ without $j$
Approach with non-iid deviations

Application of Poisson distribution

\( I_{u,j}, N_{u}^{(j)} \) independent, \( N_u, N_{u}^{(j)} \) Poisson distr. and \( \lambda_u = \sum_{i=1} p_{u,i} \) and \( \lambda_u^{(j)} = \lambda_u - p_{u,j} \), respectively

\[
\Rightarrow \quad \frac{P(I_{u,j}=1,N_{u}^{(j)}=n_u-1)}{P(I_{u,j}=1,N_{u}^{(j)}=n_u-1) + P(I_{u,j}=0,N_{u}^{(j)}=n_u)} \\
\approx \quad \frac{p_{u,j}e^{-\lambda_u^{(j)}} \lambda_u^{(j)n_u-1} / (n_u-1)!}{p_{u,j}e^{-\lambda_u^{(j)}} \lambda_u^{(j)n_u-1} / (n_u-1)! + (1-p_{u,j})e^{-\lambda_u^{(j)}} \lambda_u^{(j)n_u-1} / n_u!}
\]
Application of Poisson distribution

this becomes

\[
\frac{P(I_{u,j}=1,N_u^{(j)}=n_u-1)}{P(I_{u,j}=1,N_u^{(j)}=n_u-1)+P(I_{u,j}=0,N_u^{(j)}=n_u)} \approx \frac{n_u p_{u,j}}{n_u p_{u,j}+\lambda_u^{(j)}(1-p_{u,j})}
\]

applied analogously for \( P(I_{v,j} = 1|N_v = n_v) \) we get

\[
E(C|N_u = n_u, N_v = n_v) \approx \sum_{j=1}^{k} \left( \frac{n_u p_{u,j}}{n_u p_{u,j}+\lambda_u^{(j)}(1-p_{u,j})} \times \frac{n_v p_{v,j}}{n_v p_{v,j}+\lambda_v^{(j)}(1-p_{v,j})} \right)
\]
Testing of Poisson distribution

- with Poisson distr. approximated distr. of $C$ given $N_u = n_u, N_v = n_v$ and described mean test hypothesis that coincidences occur by chance
- compare observed value of $c_{obs}$ with expected value
Testing of Poisson distribution

\[ p^{corr} = 1 - \sum_{i=0}^{c_{obs}^{-1}} Po(i), \]
where \( Po(i) \): prob. for Poisson variable with mean \( E(C|N_u = n_u, N_v = n_v) \), \( p^{corr} \): prob. observing \( c_{obs} \) or more coincidences between \( u \) and \( v \)

- hypothesis accepted, if \( p^{corr} > p^{corr}_{max} \)
- for columns with lots of expected sequencing error, Tammi et. al. introduce correction strategy
Outline

Introduction

Locating

Bounding

Identifying DNPs

Tammi’s algorithm
Approach with iid deviations
Approach with non-iid deviations

Separating repeat copies

Algorithm of Kececioglu
ILP
## Where are we?

1. Locating the positions
2. Bounding the region
3. Identifying the DNPs
4. **Separating the set of reads into subsets**
Algorithm of Kececioglu

- Assume there are $k$ copies of a repeat
- Idea:
  - Partition the reads into $k$ classes $P_1, P_2, \ldots, P_k$
  - Each partition has a consensus string $S_1, S_2, \ldots, S_k$
  - Minimization of the sum of errors between each $S_i$ and all reads of $P_i$

\[
\sum_{1 \leq i \leq k} \sum_{F \in P_i} D(F, S_i)
\]
Constructing a graph theoretical problem

▶ Complete, edge weighted graph $K_n$:
  ▶ *Nodes*: Reads
  ▶ *Edges*: Hamming distance between 2 reads

▶ Task:
  ▶ Find k star centers
  ▶ Find for the other nodes the best star center, such that the weight of edges is minimal
  → *K-star problem*
Formulating an ILP

Two variables

- $x_{ij}$:
  - Edge between two nodes $i$ and $j$
  - Pair $(i,j)$ is ordered
  - Encodes the edges of the $k$-star
  - Number of $x_{ij}$: $n^2$

- $y_i$:
  - Vertex in $K_n$
  - Encodes the centers of the stars
  - Number of $y_i$: $n$
Objective function

Find a partition into k groups that minimize:

\[
\sum_{1 \leq i \leq k} \min_{F^* \in P_i} \left\{ \sum_{F \in P_i} H(F, F^*) \right\}
\]

\(H(., .)\) Hamming distance

Formulation for the ILP:

\[
\min \sum_{i \neq j} w_{ij} x_{ij}
\]
Constraints

Each edge can be part of a k-star or not
\[ \forall i \ \forall j, \quad x_{ij} \geq 0 \]
Each node can be a center of a k-star or not
\[ \forall i, \quad y_i \geq 0 \]
Each node has at least one incoming edge
\[ \forall j, \quad \sum_{1 \leq i \leq n} x_{ij} \geq 1 \]
Just the star centers have outgoing edges
\[ \forall i \ \forall j, \quad y_i \geq x_{ij} \]
Just k \( y_i \)'s have the value 1
\[ \sum_{1 \leq i \leq n} y_i = k \]
Objective function:
\[
\min x_{12} + 2x_{13} + x_{21} + 2x_{23} + 2x_{31} + 2x_{32}
\]
## Example

<table>
<thead>
<tr>
<th>Inequality</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\forall i \forall j, \quad x_{ij} \geq 0$</td>
<td>$x_{11} = 0, x_{12} = 0, x_{13} = 0, x_{21} = 1, x_{22} = 1, x_{23} = 1, x_{31} = 0, x_{32} = 0, x_{33} = 0$</td>
</tr>
<tr>
<td>$\forall i, \quad y_i \geq 0$</td>
<td>$y_1 = 0, y_2 = 1, y_3 = 0$</td>
</tr>
<tr>
<td>$\forall j, \quad \sum_{1\leq i\leq n} x_{ij} \geq 1$</td>
<td>$x_{11} + x_{21} + x_{31} = 0 + 1 + 0 = 1$, $x_{12} + x_{22} + x_{32} = 0 + 1 + 0 = 1$, $x_{13} + x_{23} + x_{33} = 0 + 1 + 0 = 1$</td>
</tr>
<tr>
<td>$\forall i \forall j, \quad y_i \geq x_{ij}$</td>
<td>$y_1 = 0 = x_{11}, y_1 = 0 = x_{12}, y_1 = 0 = x_{13}$, $y_2 = 1 = x_{21}, y_2 = 1 = x_{22}, y_2 = 1 = x_{23}$, $y_3 = 0 = x_{31}, y_3 = 0 = x_{32}, y_3 = 0 = x_{33}$</td>
</tr>
<tr>
<td>$\sum_{1\leq i\leq n} y_i = k$</td>
<td>$y_1 + y_2 + y_3 = 0 + 1 + 0 = 1 = k$</td>
</tr>
</tbody>
</table>
Solving the ILP

- Using:
  - LP relaxation
  - Branch-and-bound

- What happens if the solution is not integral?
  - Use rounding
Solving the ILP

- Using:
  - LP relaxation
  - Branch-and-bound

- What happens if the solution is not integral?
  - Use rounding

N. Balcazar, C. Blasse, D. Dang, H. Hauswedell, S. Thieme

RNA-Seq
Solving the ILP

- Using:
  - LP relaxation
  - Branch-and-bound

- What happens if the solution is not integral?
  - Use rounding
Sources

Knut Reinert.  
*Repeat resolution script.*  
Lecture: Advanced algorithms.

Tammi et al.  
*Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions.*  
Bioinformatics, 2002.

John Kececioglu & Jun Yu.  
*Separating repeats in DNA sequence assembly.*  

Knut Reinert.  
*Integer Linear Programming script.*  
Lecture: Discrete Mathematics.