Repeat Resolution

Nicolas Balcazar Corinna Blasse An Duc Dang
Hannes Hauswedell Sebastian Thieme

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

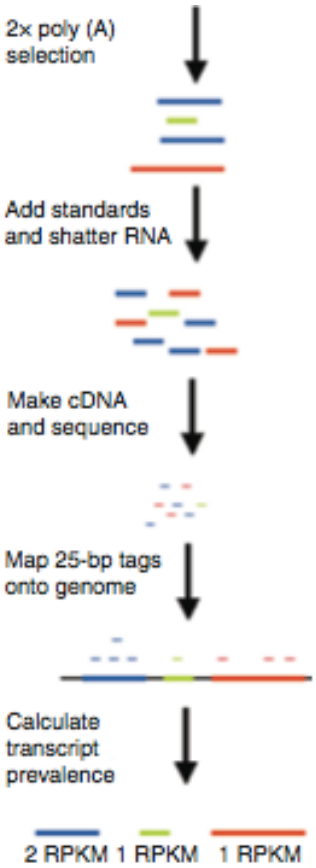
SoSe 2010
FU Berlin

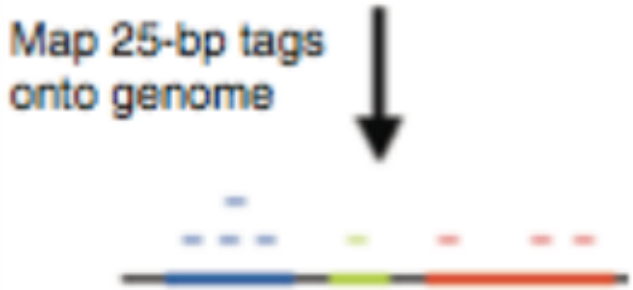
June 16, 2010

Outline

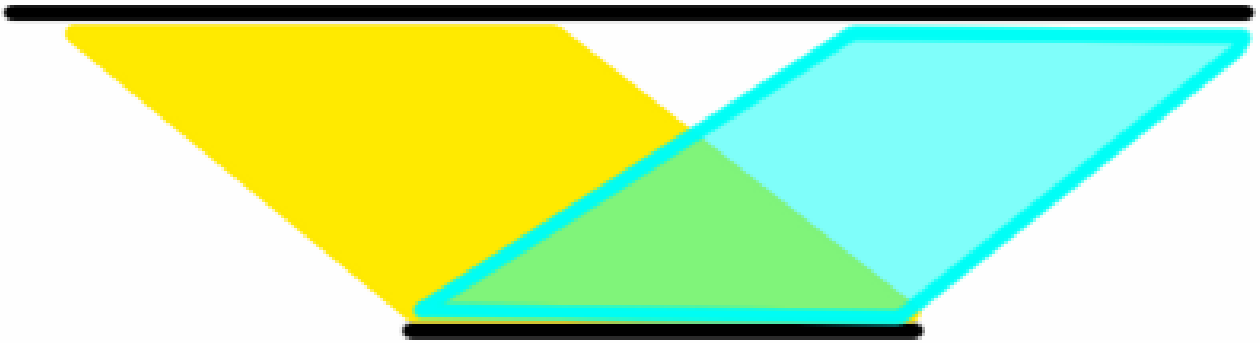
Outline

Read-Mapping



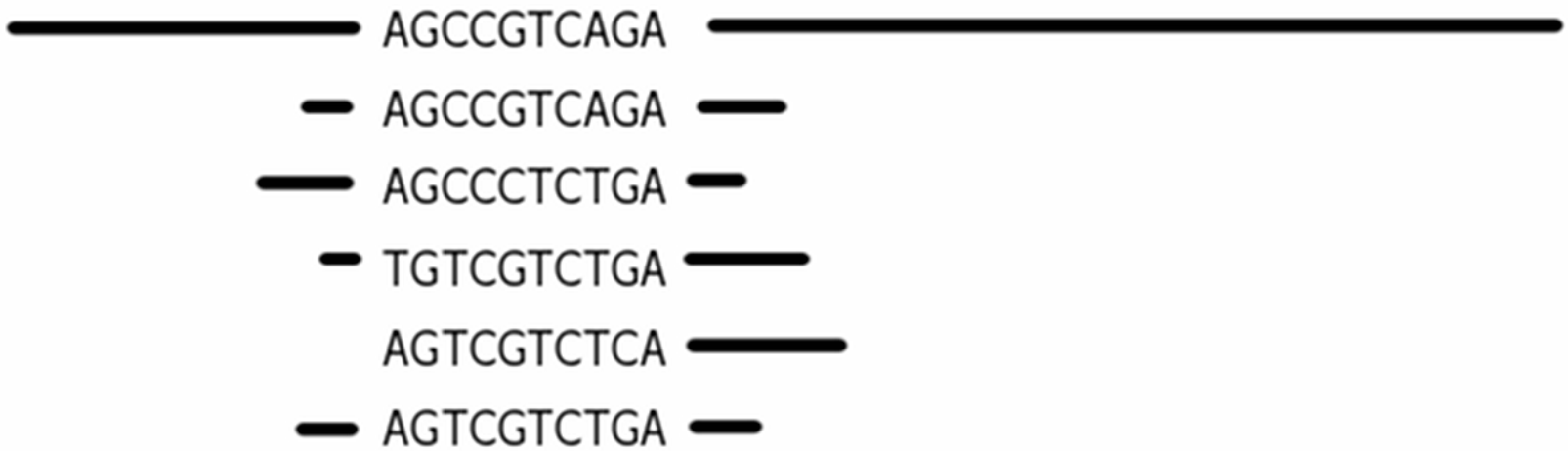


Multi-Reads



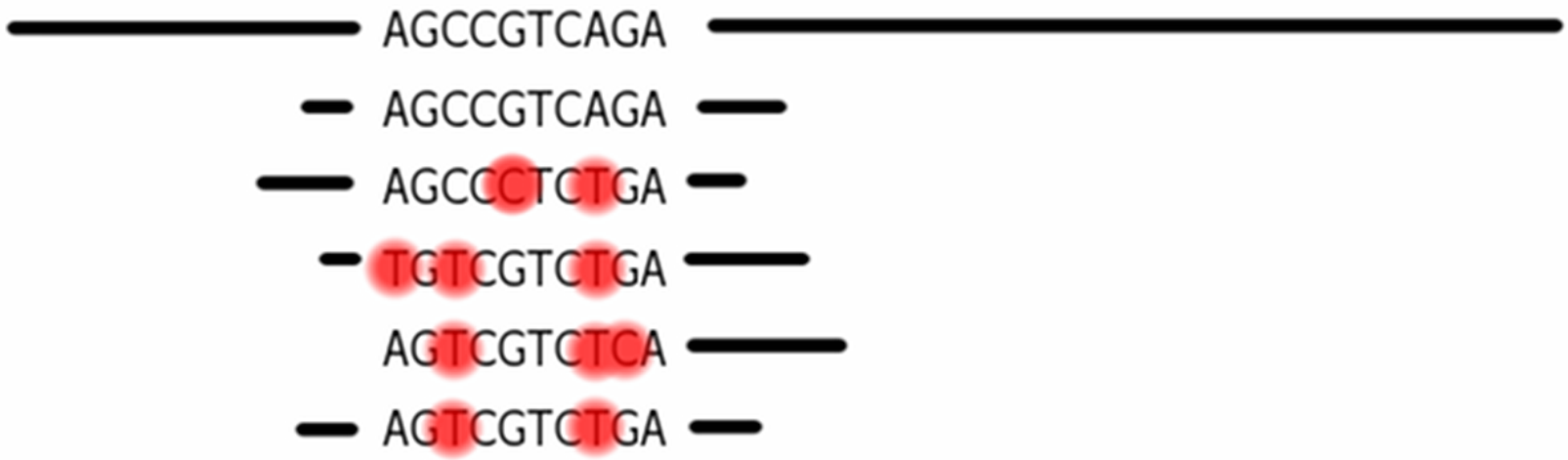
- different copies of same/similar gene
- similar sub-sequence due to similar domain/motif
- ...

Multi-Reads



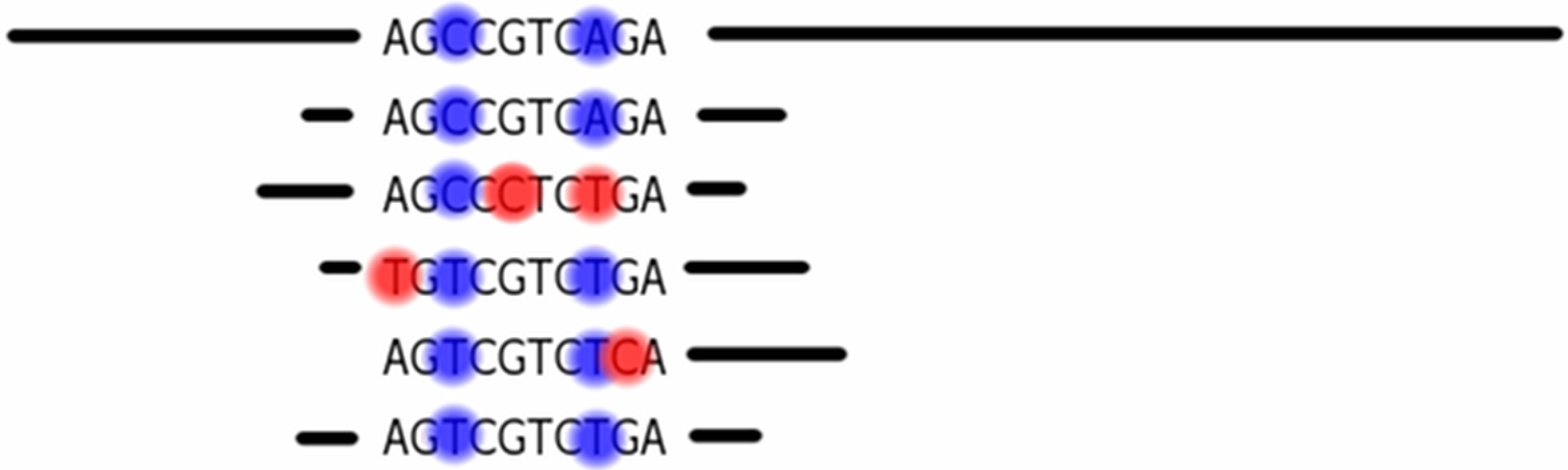
- Is there a pattern?
- certain eye-catching columns: “**D**efined **N**ucleotide **P**osition”
- possible to separate reads

Multi-Reads



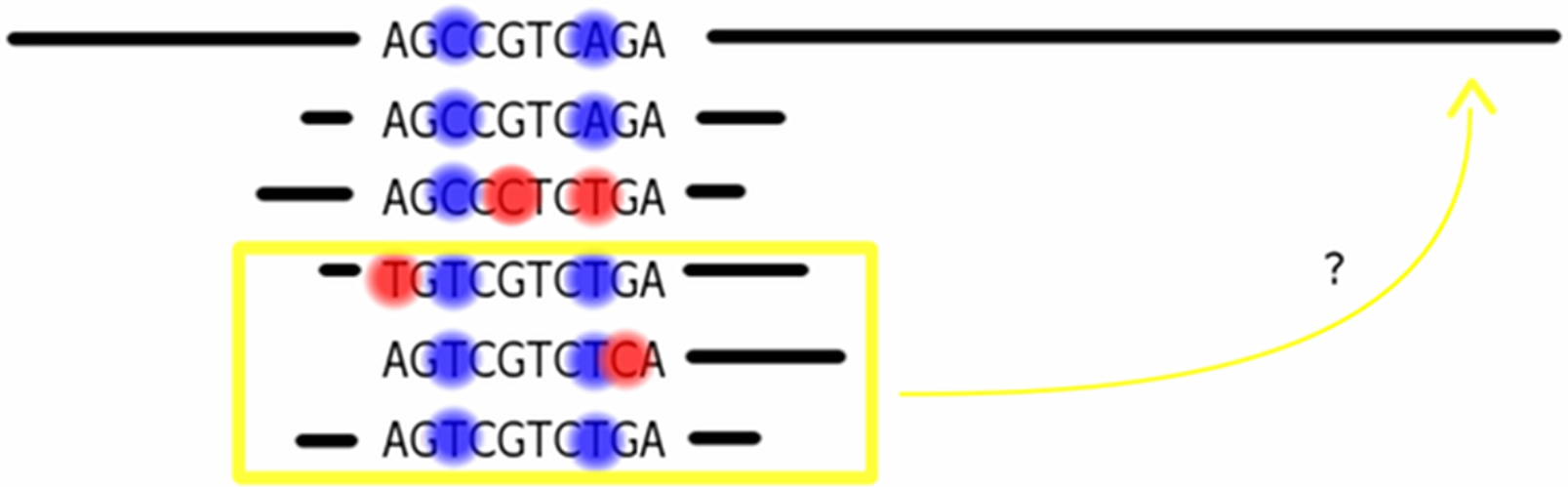
- Is there a pattern?
- certain eye-catching columns: “**D**efined **N**ucleotide **P**osition”
- possible to separate reads

Multi-Reads



- Is there a pattern?
- certain eye-catching columns: “**D**efined **N**ucleotide **P**osition”
- possible to separate reads

Multi-Reads



- Is there a pattern?
- certain eye-catching columns: “**D**efined **N**ucleotide **P**osition”
- 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

The idea

Find locations of possible repeats

Why look for repeat regions

Why not analyze whole genome?

→ runtime reduction!

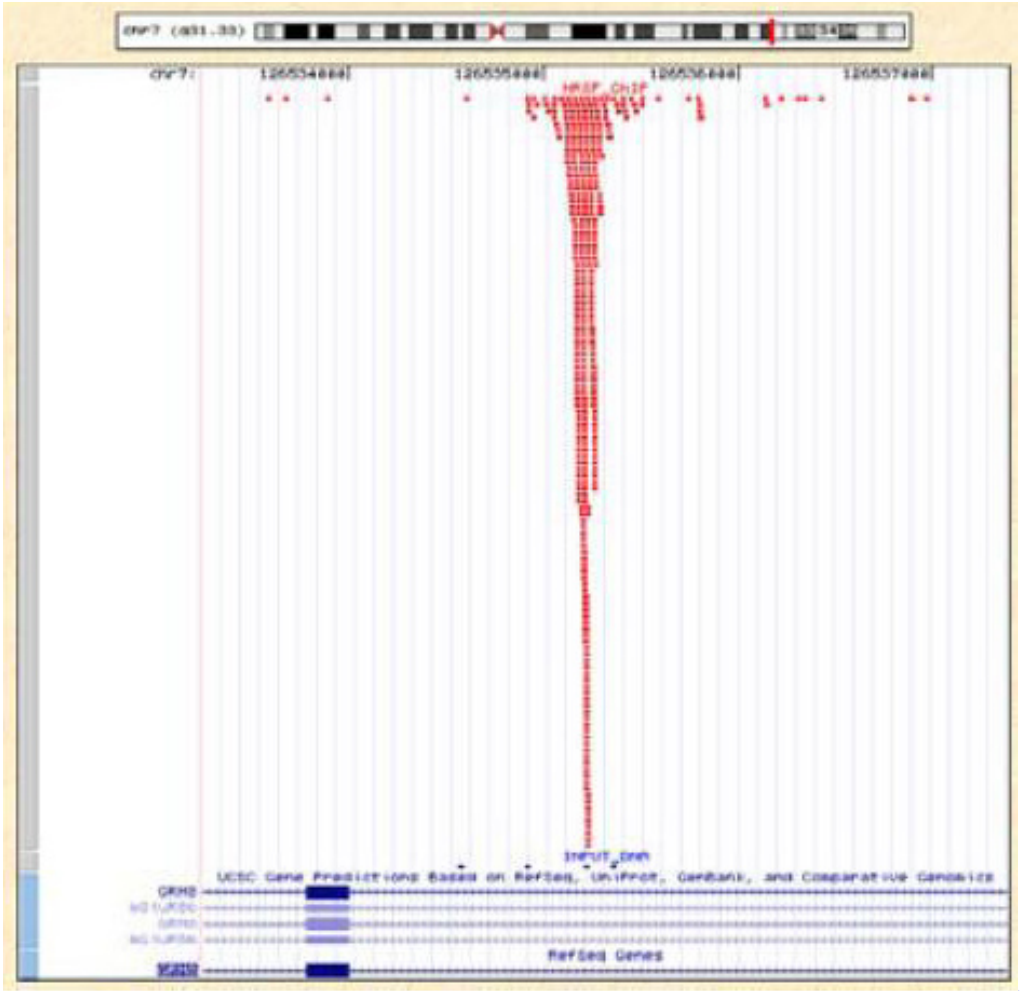
Why look for repeat regions

Why not analyze whole genome?
$\rightarrow$ runtime reduction!

The idea

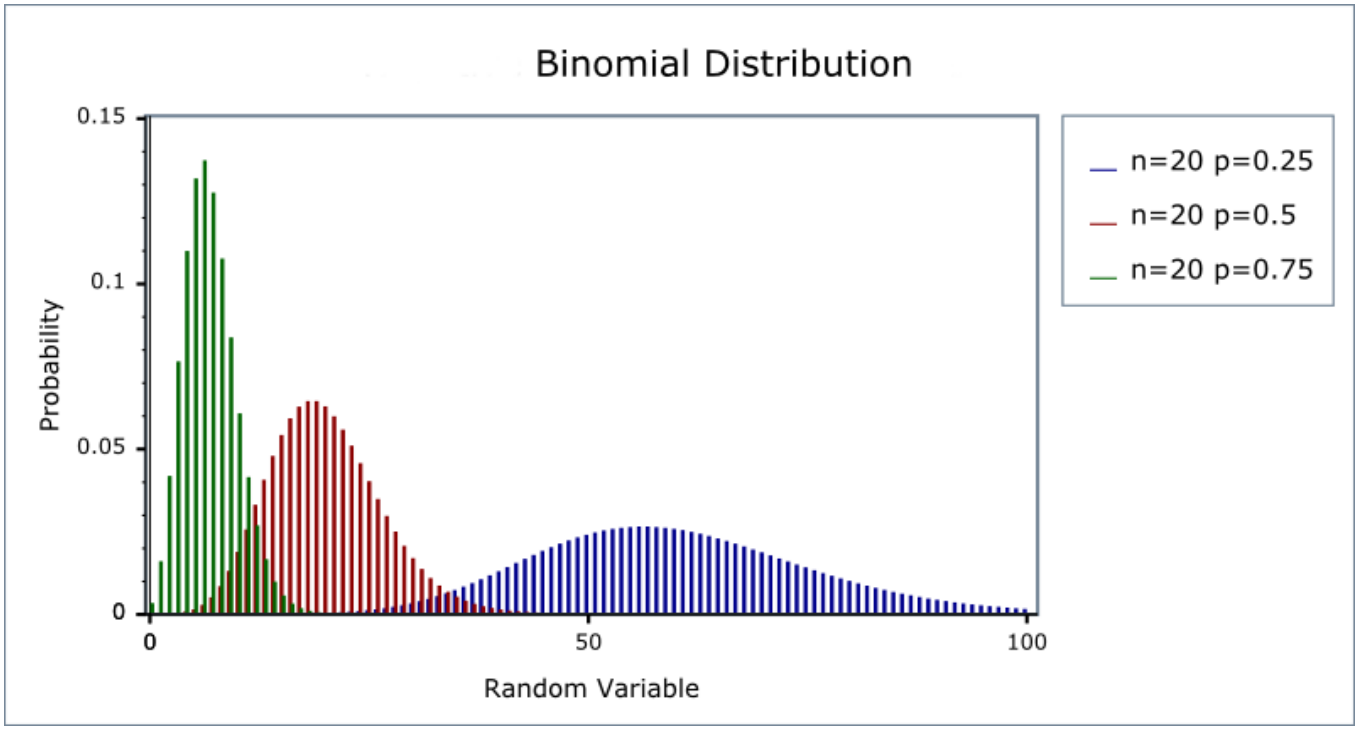
Unexpectedly high coverage $\rightarrow$ Repeats

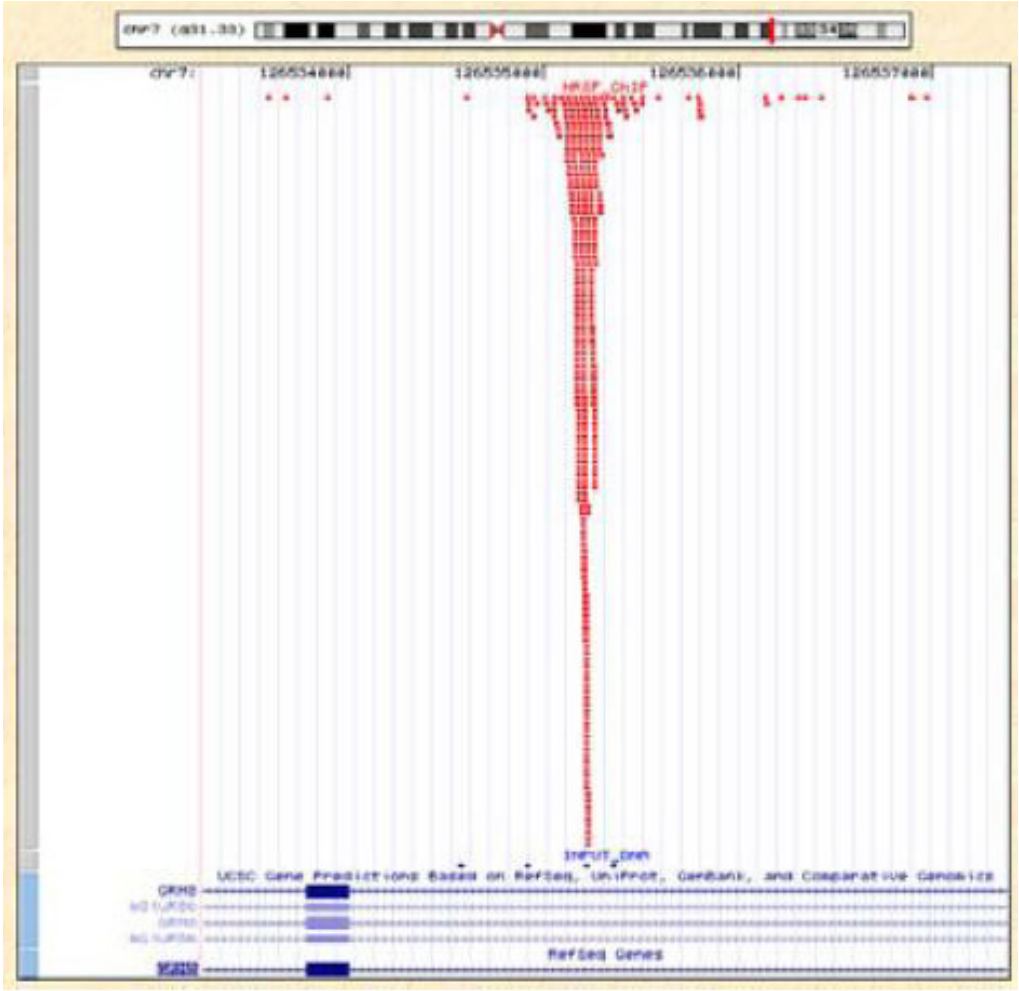
How does it look like?

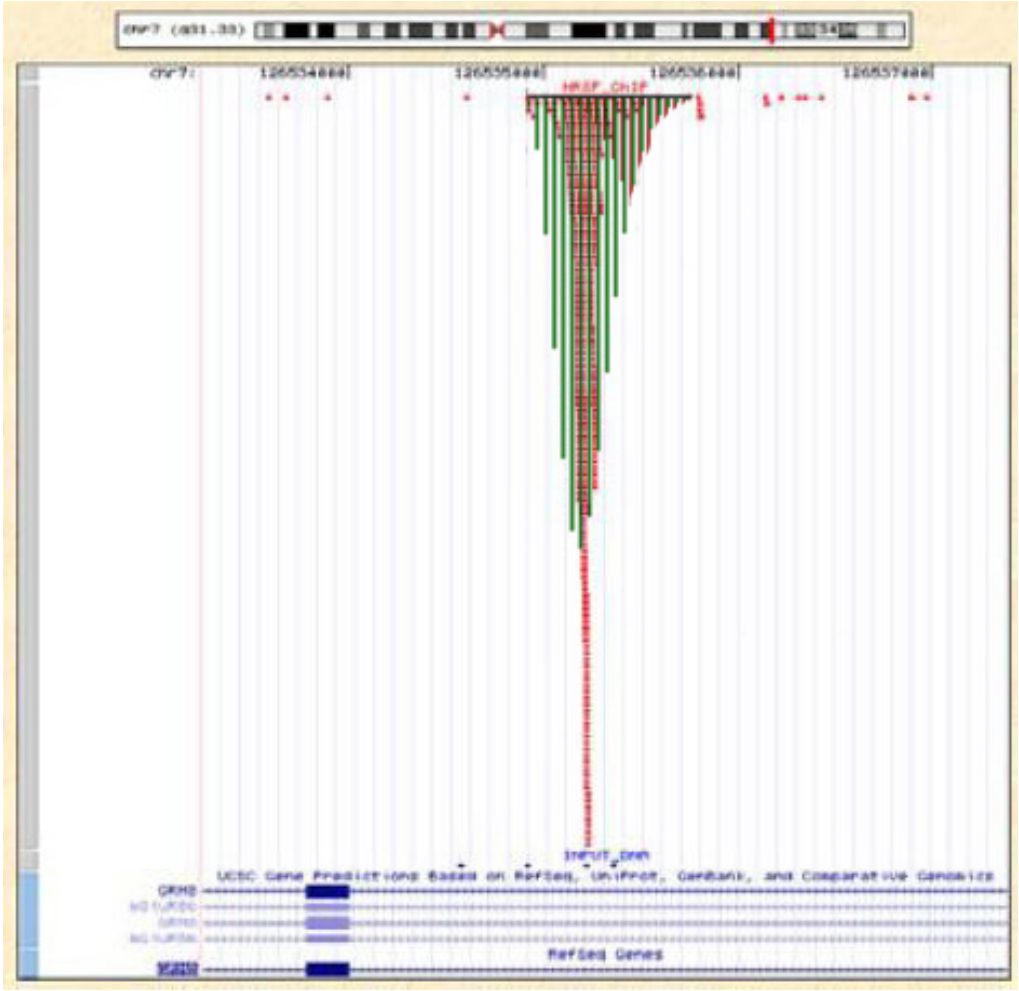


How to find locations of unexpectedly high coverage?

Comparison to a binomial distribution.




chr7:
126554444
126555444
126556444
126557444
HNRNPA2B1 CHIP
Input_DNA
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
CREM
RefSeq Genes


chr7:
RefSeq Genes

Outline

The idea

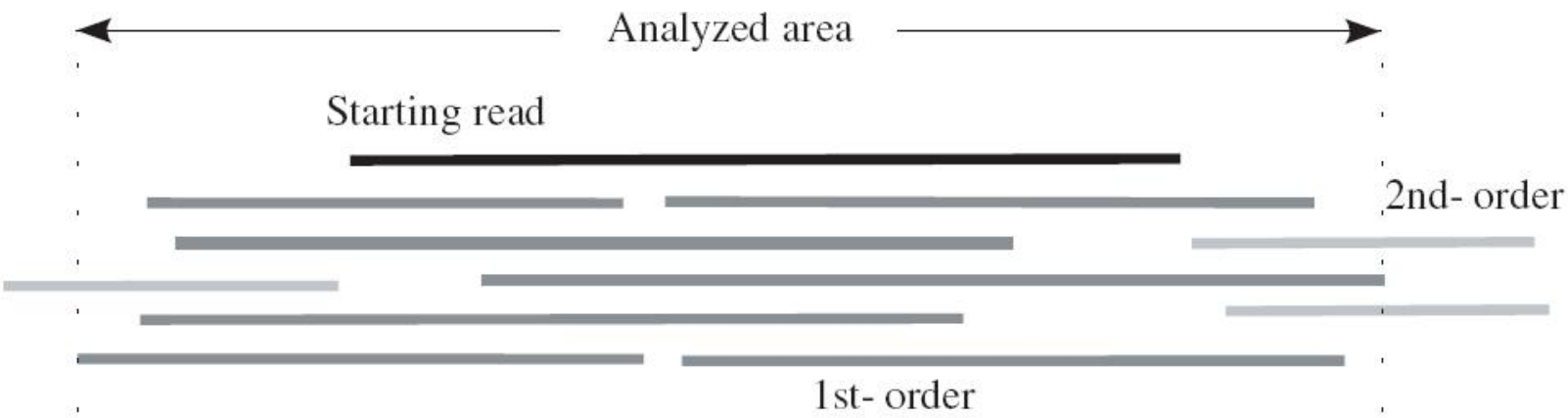
separate repeat regions by multiple alignment

```
...AGCCGTCAGA...
...AGCCGTCAGA...
...AGCCCTCTGA...
...TGTCGTCTGA...
...AGTCGTCTCA...
...AGTCGTCTGA...
```

Problem

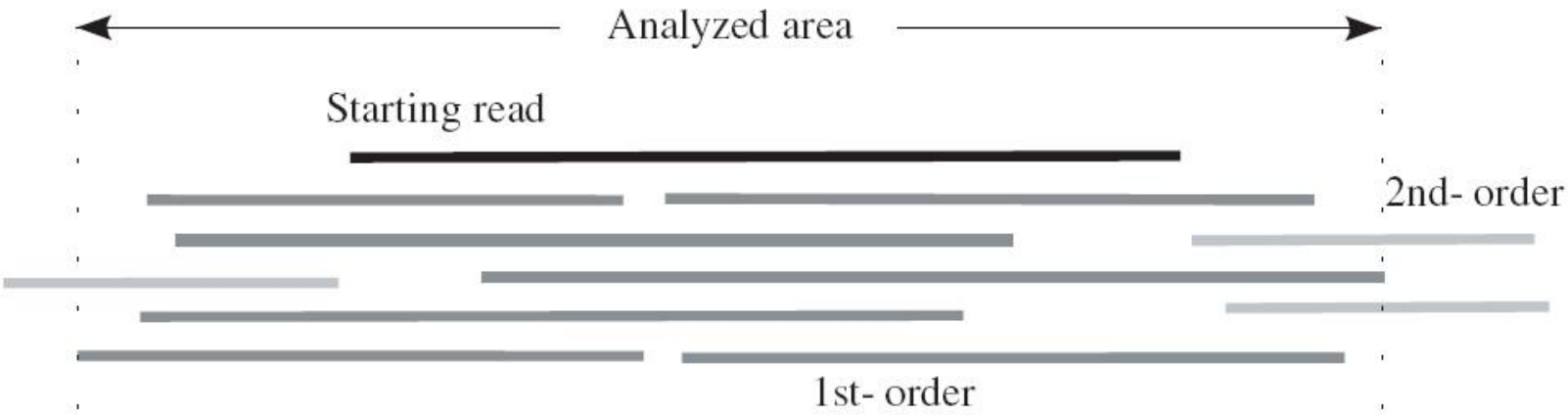
Narrow window → not enough distinguishing base sites (DNPs)

Wide window → too few fragments



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

Idea of Tammi's algorithm

- after computation of multialignment determine DNPs
- at each column assume most frequent base b as consensus base
- DNP: column with $\geq D_{min}$ non consensus bases of a same type b_1

```
...AGCCGTCAGA...
...AGCCGTCAGA...
...AGCCCTCTGA...
...TGTCGTCTGA...
...AGTCGTCTCA...
...AGTCGTCTGA...
```

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 $D_{min} = 2$
- 1st blue col has 3 deviating bases of type C
- 2nd blue col is supporting col with 2 deviating bases of type A

```
...AGCCGTCAGA...
...AGCCGTCAGA...
...AGCCCTCTGA...
...TGTCGTCTGA...
...AGTCGTCTCA...
...AGTCGTCTGA...
```

Mathematical definitions

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

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: $I_j = I_{u,j} I_{v,j}$
- $I_j = 1$ if deviation at both positions in j and $I_j = 0$ otherwise
- total number of coincidences: $C = \sum_{j=1}^{k} I_j = \sum_{j=1}^{k} I_{u,j} I_{v,j}$

But ...

The quality values of the base recall is neglected!

Recap of hypergeometic distribution

- assume deviations are iid (not always true but approximates quite well):
  $p_{u,1} = \ldots = p_{u,k}, p_{v,1} = \ldots = p_{v,k}$
  $\rightarrow$ use hypergeometric distribution
- prob. for number of successes in n draws from a finite population *without* replacement

Recap of hypergeometic distribution

toy example:

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

$$P_k(N, M, n) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

Application of hypergeometric distribution

- analogously derive distribution C given $N_u = n_u, N_v = n_v$:

$$P(C = x) = \frac{\binom{n_v}{x}\binom{k-n_v}{n_u-x}}{\binom{k}{n_u}},$$

$$0 \leq x \leq n_v, 0 \leq n_u - x \leq k - n_v$$

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

```
...AGCCGTCAGA...
...AGCCGTCAGA...
...AGCCCTCTGA...
...TGTCGTCTGA...
...AGTCGTCTCA...
...AGTCGTCTGA...
```

Approach with non-iid deviations

- for different *p*s each deviation has separate prob. and expectation of C using hypergeometr. 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

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

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

$$\approx \frac{p_{u,j} e^{-\lambda_u^{(j)}} {\lambda_u^{(j)}}^{n_u-1}/(n_u-1)!}{p_{u,j} e^{-\lambda_u^{(j)}} {\lambda_u^{(j)}}^{n_u-1}/(n_u-1)! + (1-p_{u,j}) e^{-\lambda_u^{(j)}} {\lambda_u^{(j)}}^{n_u-1}/n_u!}$$

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

$$E(C|N_u=n_u, N_v=n_v) \approx \sum_{j=1}^{k} \left(\frac{n_u p_{u,j}}{n_u p_{u,j}+\lambda_u^{(j)}(1-p_{u,j})} \times \frac{n_v p_{v,j}}{n_v p_{v,j}+\lambda_u^{(j)}(1-p_{v,j})}\right)$$

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

Where are we?

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

```
...AGCCGTCAGA...
...AGCCGTCAGA...
...AGCCCTCTGA...
...TGTCGTCTGA...
...AGTCGTCTCA...
...AGTCGTCTGA...
```

Algorithm of Kececioglu

- Assume there are k copies of a repeat
- Idea:
  - Partition the reads into k classes $P_1, P_2, ..., P_k$
  - Each partition has a consensus string $S_1, S_2, ..., S_K$
  - 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

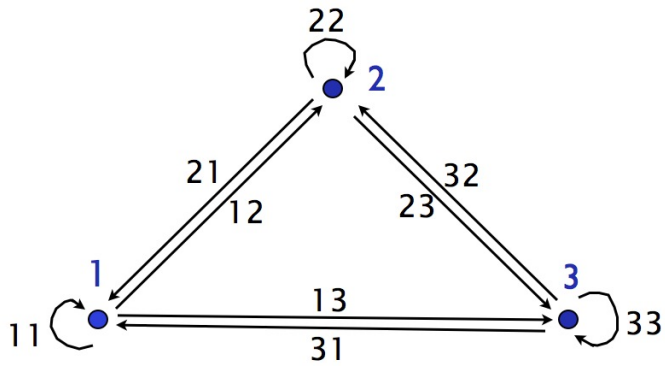
- Task:
  - Find k star centers
  - Find for the other nodes the best star center, such that the weight of edges is minimal

  $\rightarrow$ *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

$$\forall i \; \forall j, \qquad x_{ij} \geq 0$$

Each node can be a center of a k-star or not

$$\forall i, \qquad y_i \geq 0$$

Each node has at least one incoming edge

$$\forall j, \qquad \sum_{1 \leq i \leq n} x_{ij} \geq 1$$

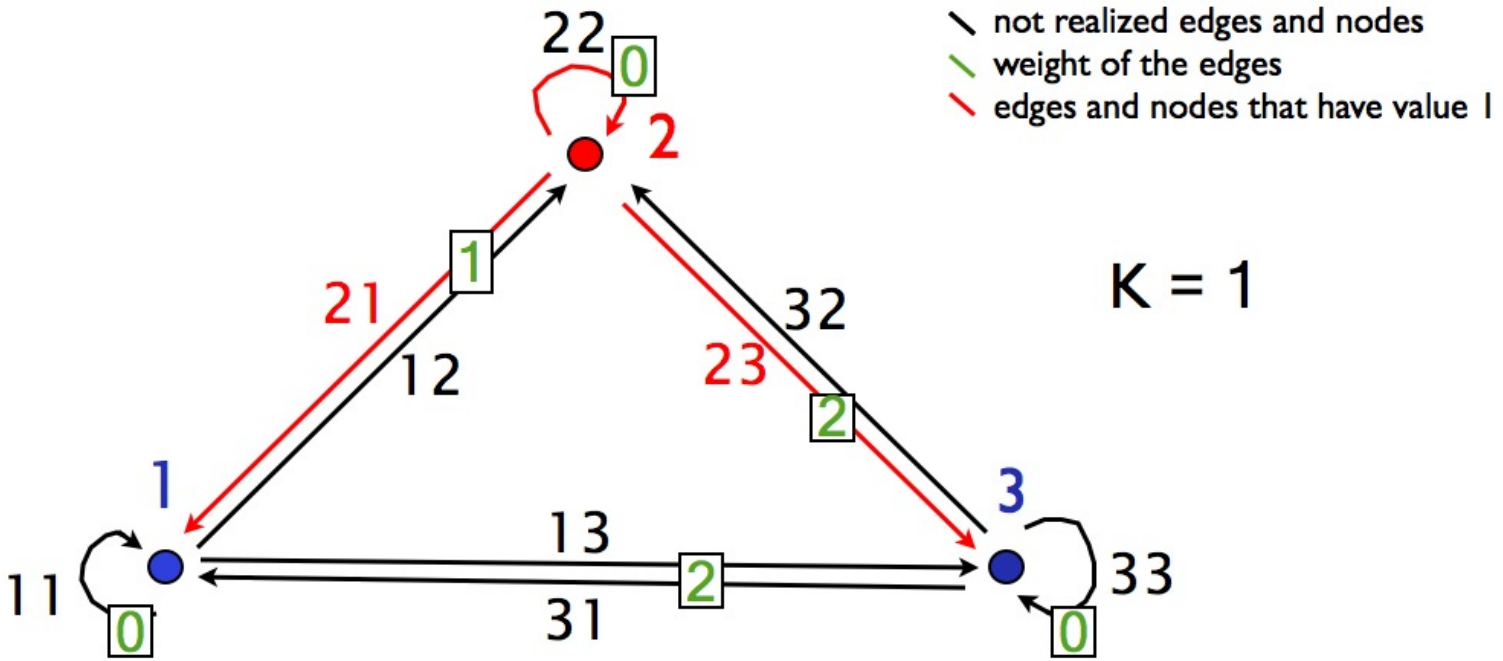
Just the star centers have outgoing edges

$$\forall i \; \forall j, \qquad y_i \geq x_{ij}$$

Just k y_is have the value 1

$$\sum_{1 \leq i \leq n} y_i = k$$

Example



- Objective function:

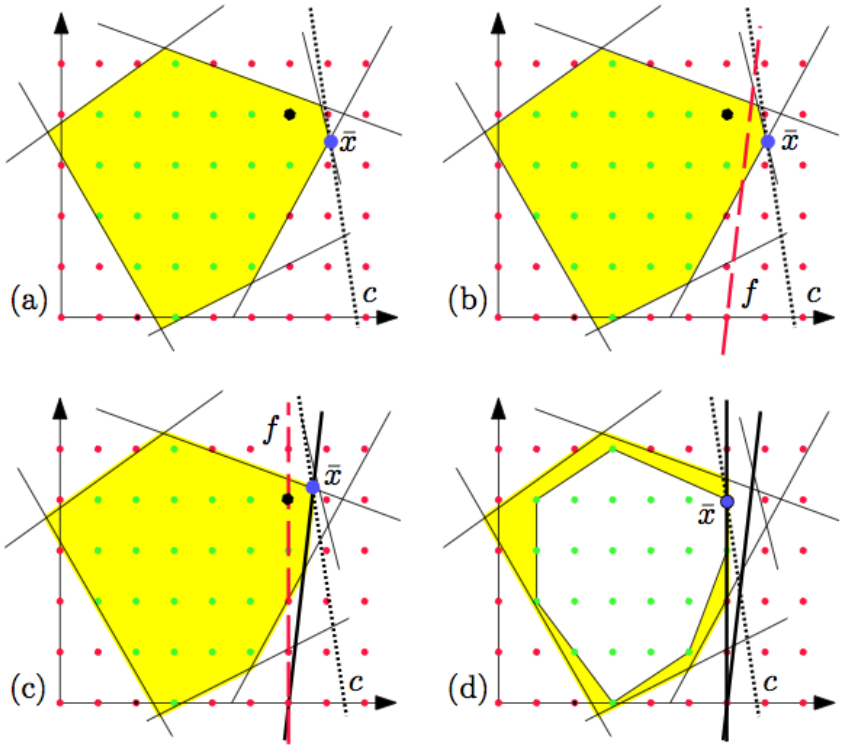
$$\min x_{12} + 2x_{13} + x_{21} + 2x_{23} + 2x_{31} + 2x_{32}$$

Example

$\forall i\ \forall j,$	$x_{ij} \geq 0$	$x_{11} = 0, x_{12} = 0, x_{13} = 0, x_{21} = 1, x_{22} = 1, x_{23} = 1, x_{31} = 0, x_{32} = 0, x_{33} = 0$
$\forall i,$	$y_i \geq 0$	$y_1 = 0, y_2 = 1, y_3 = 0$
$\forall j,$	$\sum_{1 \leq i \leq n} x_{ij} \geq 1$	$x_{11} + x_{21} + x_{31} = 0 + 1 + 0 = 1,$ $x_{12} + x_{22} + x_{32} = 0 + 1 + 0 = 1,$ $x_{13} + x_{23} + x_{33} = 0 + 1 + 0 = 1$
$\forall i\ \forall j,$	$y_i \geq x_{ij}$	$y_1 = 0 = x_{11}, y_1 = 0 = x_{12}, y_1 = 0 = x_{13},$ $y_2 = 1 = x_{21}, y_2 = 1 = x_{22}, y_2 = 1 = x_{23},$ $y_3 = 0 = x_{31}, y_3 = 0 = x_{32}, y_3 = 0 = x_{33}$
	$\sum_{1 \leq i \leq n} y_i = k$	$y_1 + y_2 + y_3 = 0 + 1 + 0 = 1 = k$

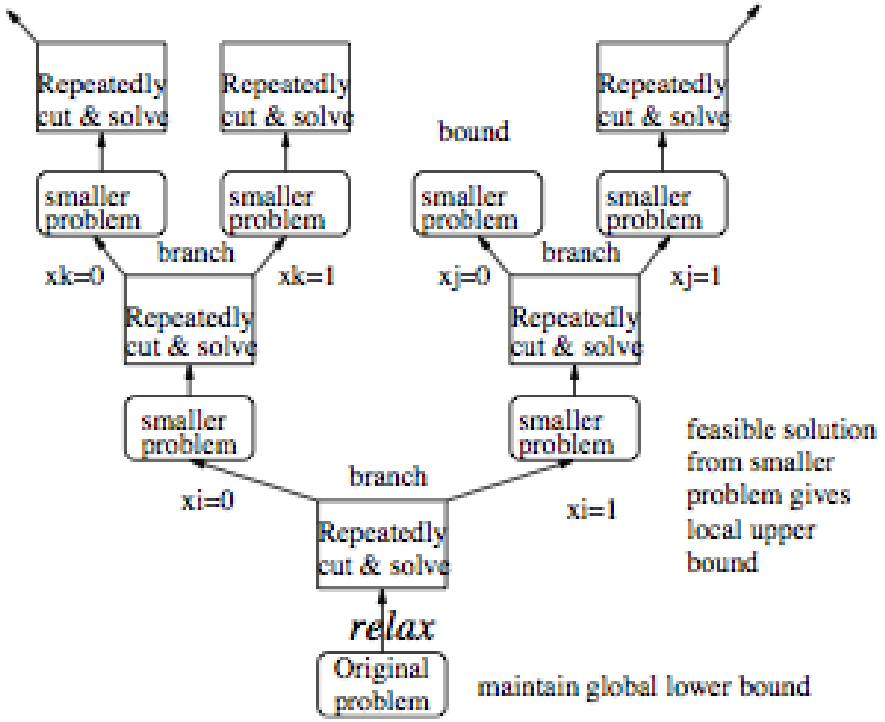
Solving the ILP

- Using:
  - LP relaxation
  - Branch-and-bound
- What happens if the solution is not integral?
  - Use rounding



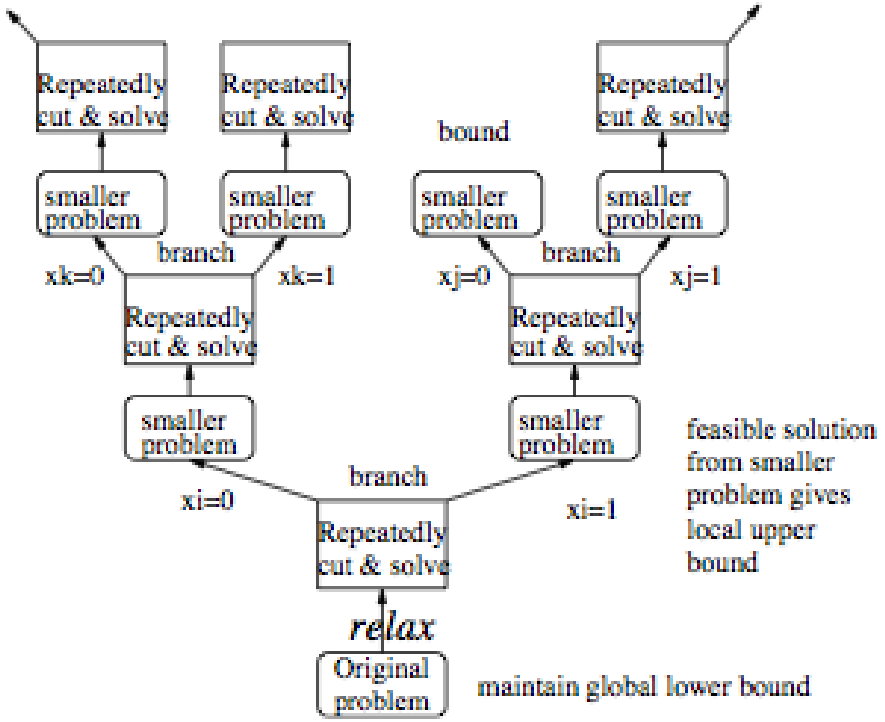
Solving the ILP

- Using:
  - LP relaxation
  - Branch-and-bound
- What happens if the solution is not integral?
  - Use rounding



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 shortgun assemblies using defined nucleotide poitions.
Bioinformatics, 2002.

John Kececioglu & Jun Yu.
Separating repeats in DNA sequence assembly.
RECOMB, 2001.

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