

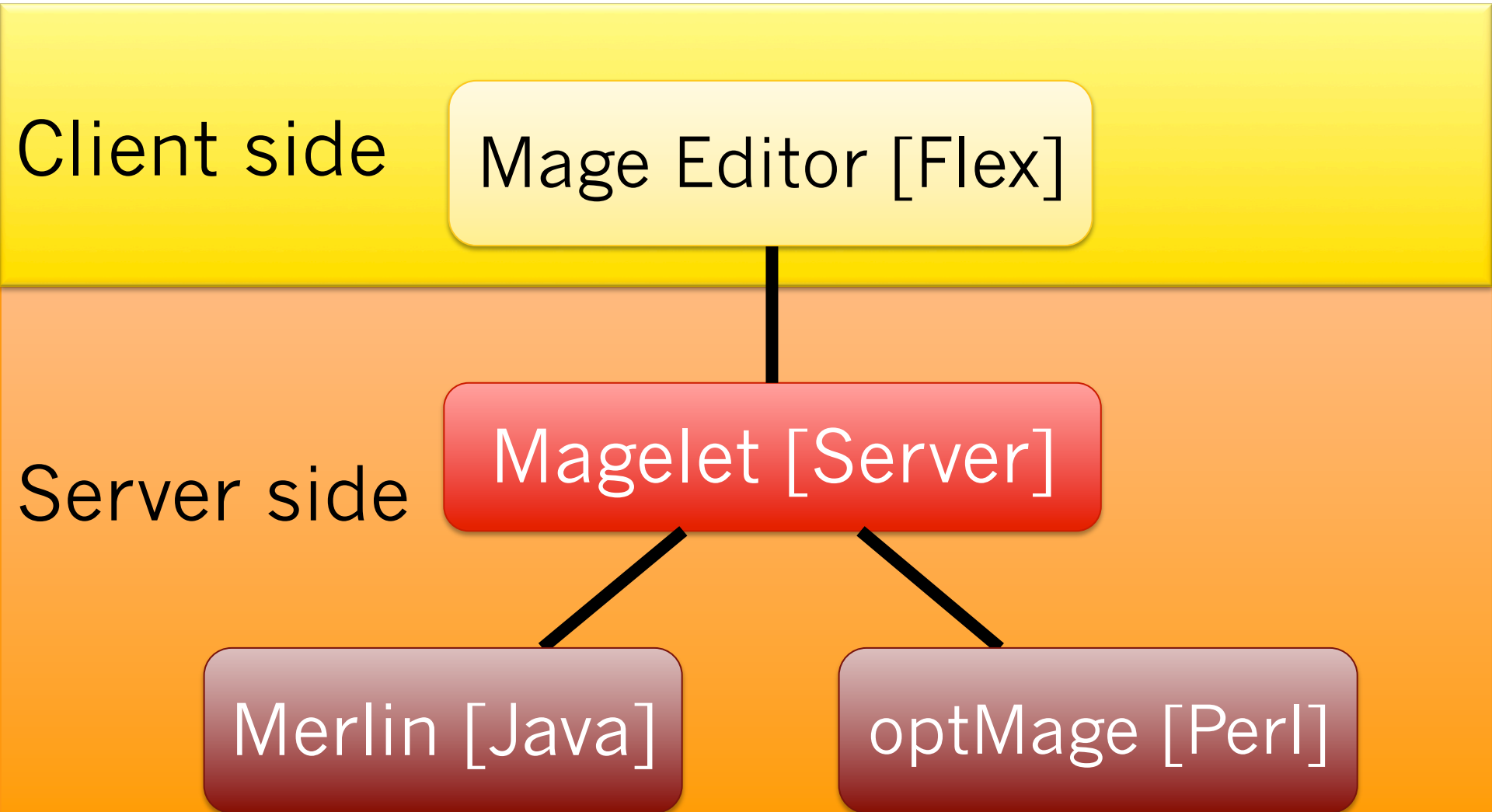
# Merlin

**Overview of the Merlin Tool and Process**

**Samir Ahmed, Douglas Densmore, Swapnil Bhatia**

**Tuesday March 6th**

# Overview of Tools



# Why Merlin?

## Overview

- Merlin attempts to optimized the Oligo selection process for MAGE.
- Oligo Homology with genome
- Oligos Homology with other oligos

# Why Merlin?

## Oligo Selection Process

- A target mutation is used to define an oligo for generation.
- For a given target mutation are several ( $10^1$  to  $10^2$ ) possibilities.
- For a set of target mutations there are millions of permutations

# Why Merlin?

## Problem Definition

*Given a set of target mutations, Construct a set of oligos that minimize genome homology and oligo pool homology without compromising recombination efficiency.*

- optMage used as the benchmark
- Free Energy is not over threshold

# Terminology

## Scoring Terms

- **$\Delta G$**  : Free Energy Score
  - Value to quantify the energy associated with secondary structure formation
- **$B_G$**  : Blast Genome Score
  - Value to quantify an individual oligo's homology to the genome, excluding the desired region to be mutated
- **$B_o$**  : Blast Oligo Score
  - Value to quantify an individual oligo's homology to the remainder of the oligos in the selection pool

# Terminology

## Oligo Terms

- **Target Mutation** : Desired insertion, deletion or mismatch
- **LP, RP** : Left and Right start position of the genome
- **$b_{5'}$  ,  $b_{3'}$** : Buffer regions between 5' and 3' ends [bp]
- **t**: Target region length [bp]
- **L** : Desired Length of Oligo
- **Span** : Entire region of interest for a given mutation to the genome
- **Margin** : Margin of flexibility for selecting oligos

# Terminology

## Oligo Visualized

Number of Possible Oligos = Margin Length

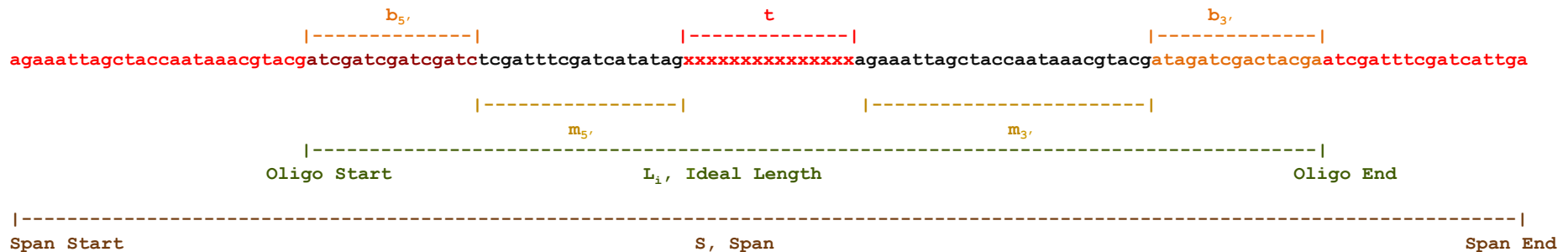
ARBITRARY OLIGO

$b_5, = 15, b_3, = 15, t = 15$

$m_5, = 18, m_3, = 27, m = 45$

$L_i = b_5, + b_3, + m = 90$

$S = L_i + m = 135$





# Terminology

## Mistarget Terms

- **Mistarget** : Regions that could be aligned with undesired regions of the genome or other oligos
- Mistarget scores are calculated from BLAST results (in different ways)
- Mistarget scores are used to calculate  $B_G$  and  $B_o$

# High Level Overview

## optMage

- Creates an Oligo at b<sub>5</sub>'
- Shifts target towards 3'
- Stops when  $\Delta G$  below threshold



Entire Span



Selected Oligo



Target Region



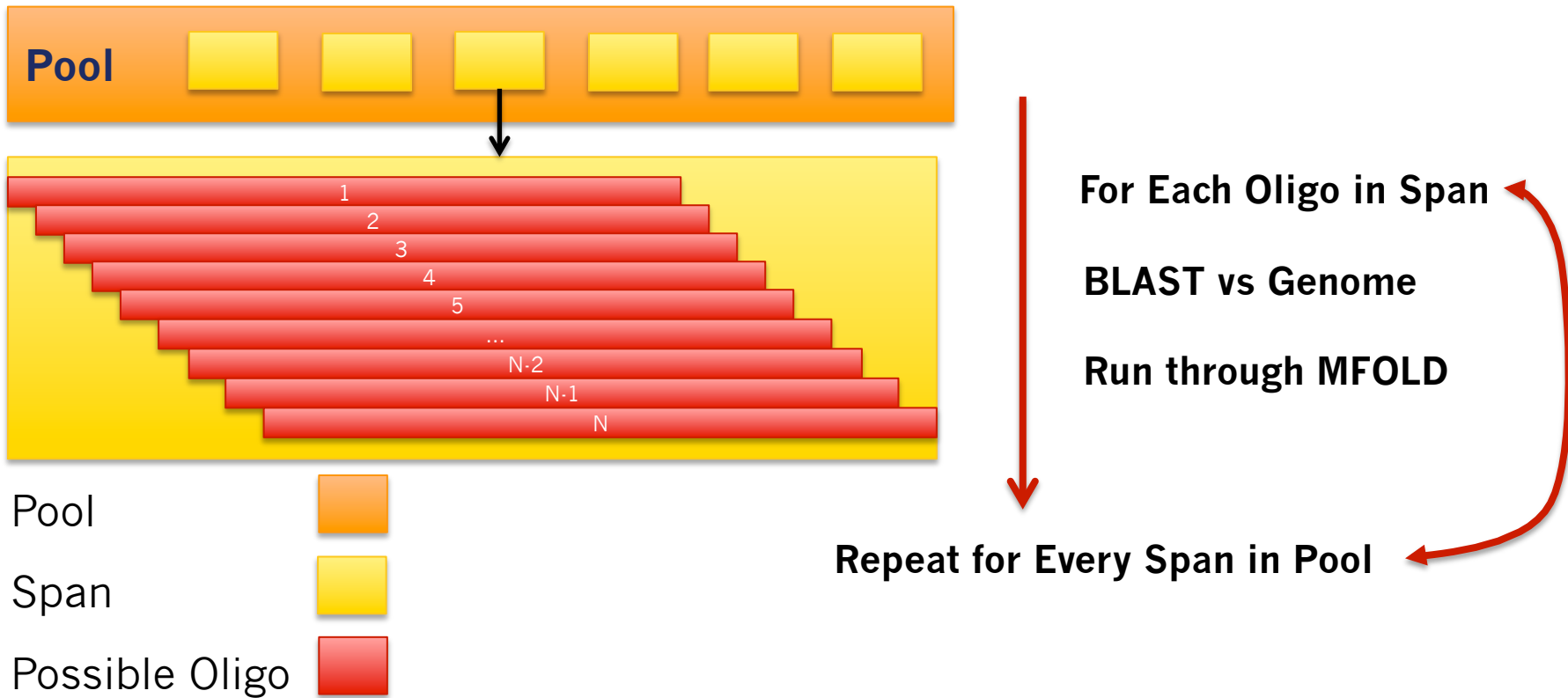
# High Level Overview

## Merlin

- Oligo Score ( $\Delta G$  ,  $B_G$  ,  $B_o$  )
- Every possible  $\Delta G$  ,  $B_G$  score is calculated
- $B_o$  is calculated for entire spans
- Heuristic algorithm to minimize Oligo Score
- Uses lexicographical ordering to compare scores

# Merlin Stage 1

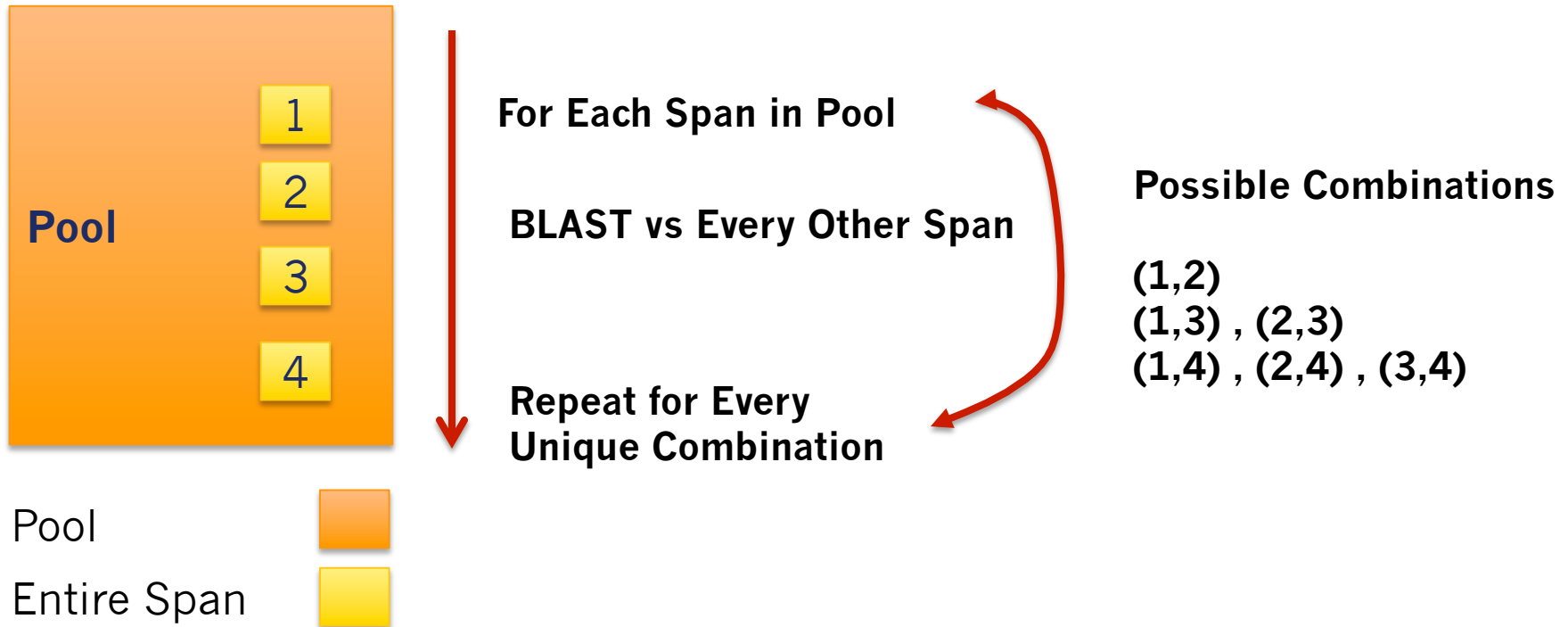
1 : Calculate  $\Delta G$  ,  $B_G$  for every possible oligo



# Merlin

## Stage 2

2 : BLAST Every Span against each other



\* We will use a linear scaling to approximate mistarget score

# Merlin

## Stage 3

### 3 : Oligo Score Minimization Heuristic

Given a pool of  $k$  spans with  $n_k$  possible oligo for a given span,

For all  $k$ , calculate local  $B_o$  values for each oligo\*.

Find the span with the lowest cumulative Bo Score  $\sum_0^{n_k} B_o$

Select that oligo with the minimal score ( $\Delta G$  ,  $B_G$  ,  $B_o$  )

Lock the selected oligo

Repeat until all oligos in span have been 'locked'

# Merlin

## Heuristic by Example

Create an Artificial Genome to Show Heuristic Intuition

> No fixed matches

```
CCATACCGTCCTAATTCTTCGGTTATGTTTCCGATGTAGGAGTGAGCCTA
CCTGCCTTTGCGTCTTGATACCAATGAAAAACCTATGCACTTTGTACAGG
GTGCCATCGGGTTTCTGAACTCTCAGATAGTGGGGATCCCGGGTAAAGAC
CTATATCTGCGGTCCAACCTTAGGCATAAACCTCCATGCTACCTAGTCAGA
CCCACCCCGCACGGGGTAAATATGGCACGCGTCCGACCTGGTTCCTGGCG
TTCTACGCTGCCACGTGTTCACTAACTGTTGTTTGGTAGCACAAAAGTAT
TACCATGGTCCTAGAAGTTCGGCACAGTTAGTTCGAGCCTAATGTCACAA
ATGACGCAGAACGCCAATGAGTGCCAGACATTAGGTGGAGTTCAGTTCGG
TAACGGAGAGACTCTGCGGCGTACTTAATTATGCATTTGAAACGCGCCCA
AGTGACGCTAGGCAAGTCAGAGCAGGTTCCCGTGTTAGCTTGAGGGTAAA
CCATACCGTCCTAATTCTTCGGTTATGTTTCCGATGTAGGAGTGAGCCTA
CCTGCCTTTGCGTCTTGATACCAATGAAAAACCTATGCACTTTGTACAGG
GTGCCATCGGGTTTCTGAACTCTCAGATAGTGGGGATCCCGGGTAAAGAC
CTATATCTGCGGTCCAACCTTAGGCATAAACCTCCATGCTACCTAGTCAGA
CCCACCCCGCACGGGGTAAATATGGCACGCGTCCGACCTGGTTCCTGGCG
TTCTACGCTGCCACGTGTTCACTAACTGTTGTTTGGTAGCACAAAAGTAT
TACCATGGTCCTAGAAGTTCGGCACAGTTAGTTCGAGCCTAATGTCACAA
ATGACGCAGAACGCCAATGAGTGCCAGACATTAGGTGGAGTTCAGTTCGG
TAACGGAGAGACTCTGCGGCGTACTTAATTATGCATTTGAAACGCGCCCA
AGTGACGCTAGGCAAGTCAGAGCAGGTTCCCGTGTTAGCTTGAGGGTAAA
```

A & C



A & B



No Match Between B & C

# 3 Oligo Visual Explanation

Sort by B0 Values by Smallest Possible,  
Valid by DG Threshold

Lowest  
Individual B0  
Score



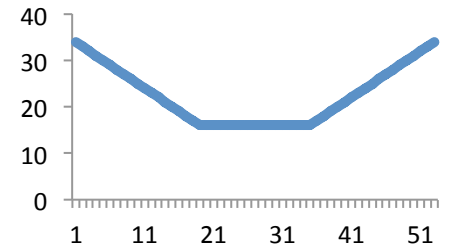
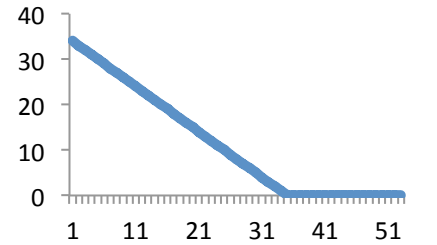
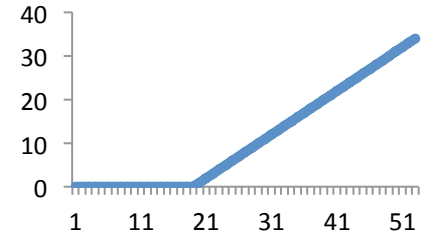
Oligo B



Oligo C



Oligo A



Entire Span



Mistarget Region





# 3 Oligo Visual Explanation

Redefine the span to represent the optimal Oligo

Lowest  
Individual BO  
Score



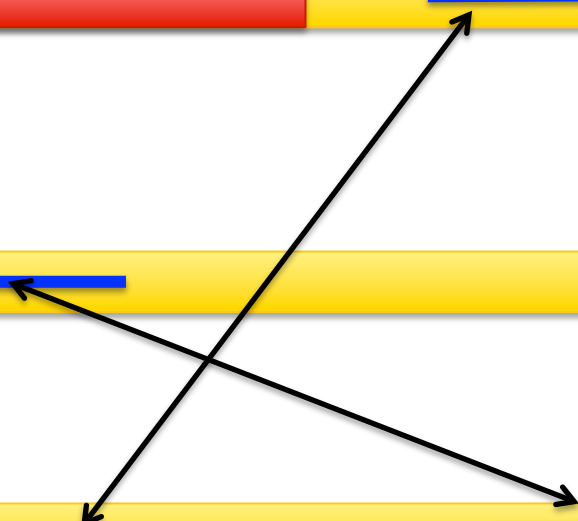
Oligo B



Oligo C



Oligo A



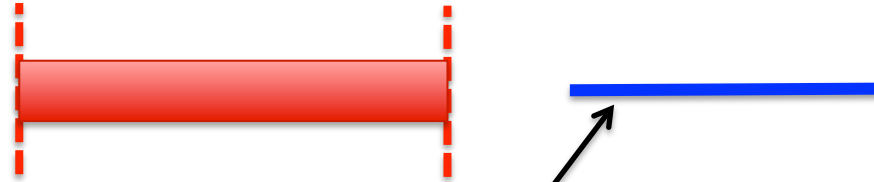
# 3 Oligo Visual Explanation

Recalculate the Blast Oligo Scores for the remainder  
Of the Oligos, to update what has been affected

Lowest  
Individual BO  
Score



Oligo B



Oligo C



Oligo A



# Merlin

## Live Example

- Now using an real MAGE test sample

Gene	Name	Sense	Replicore	LP	RP	Mutation Sequence
pps	-	2	1785091	1785181	M	CATCATcatcatCATCAT
tktA	-	2	3079612	3079702	M	CATCATcatcatCATCAT
aroF	-	2	2739127	2739217	M	CATCATcatcatCATCAT
aroG	+	1	784811	784901	M	CATCATcatcatCATCAT
aroH	+	2	1786414	1786504	M	CATCATcatcatCATCAT
aroD	+	2	1772665	1772755	M	CATCATcatcatCATCAT
aroE	-	2	3428815	3428905	M	CATCATcatcatCATCAT
ydiB	+	2	1771768	1771858	M	CATCATcatcatCATCAT
aroK	-	2	3517041	3517131	M	CATCATcatcatCATCAT
aroL	+	1	405584	405674	M	CATCATcatcatCATCAT
aroA	+	1	957990	958080	M	CATCATcatcatCATCAT
aroC	-	2	2445450	2445540	M	CATCATcatcatCATCAT
trpE	-	1	1320925	1321015	M	CATCATcatcatCATCAT
trpD	-	1	1319363	1319453	M	CATCATcatcatCATCAT
trpC	-	1	1317764	1317854	M	CATCATcatcatCATCAT
trpA	-	1	1315201	1315291	M	CATCATcatcatCATCAT
trpB	-	1	1316394	1316484	M	CATCATcatcatCATCAT
crp	+	2	3484097	3484187	M	CATCATcatcatCATCAT

# Switches

- Blast Scoring Method
  - Sum of raw bitscore
  - Expected value of alignment
- Free Energy Threshold
  - Below threshold is zero
  - Below threshold is negative
- Framework for adding more switches exists

# Comparison

## Results vs Runtime

- Slower than optMage
- BLAST is time consuming
- Can be parallelized for time efficiency
- $B_G$  and  $B_0$  results provide us with more complete analysis

# Comparison

## Why Merlin?

- By design cannot be worse than optMage
- Could drastically reduce  $B_G$  score