IGA
ISTITUTO DI GENOMICA APPLICATA

National Research Council of Italy

IGA
TECHNOLOGY SERVICES

Next-Generation Sequencing and Methylation Profiling.

Alignment to a reference of BS-treated sequences

Alberto Policriti
(joint work with C. Del Fabbro, E. De Paoli, N. Prezza, and F. Vezzi)

Napoli — September 26, 2012

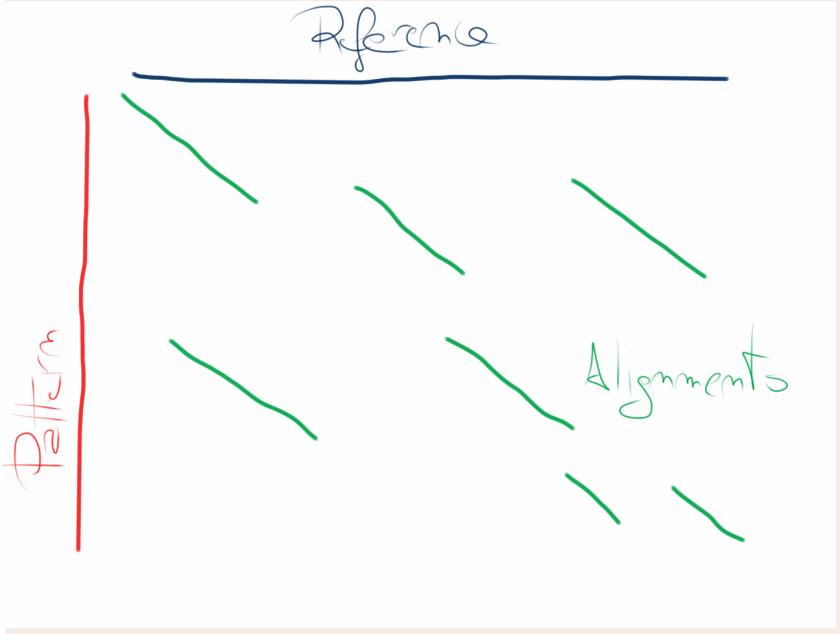
Outline

1. Introduction
2. Alignement
3. Alignment of BS-treated sequences

Dot-plot

Dot-plot

A familiar alignment “program”



- extends to strings a valuation given on the alphabet by a cost *matrix*
- gaps are allowed but *it is not clear how much to charge for them*

Dot-plot

A familiar alignment “program”

- extends to strings a valuation given on the alphabet by a cost *matrix*
- gaps are allowed but *it is not clear how much to charge for them*

Aligning

Computing a *sliding* dot-plot

Distances and Scores

A “negative” or “positive” point of view

Distance $\Downarrow$ Score $\Uparrow$

Distances and Scores

A "negative" or "positive" point of view

Distance $\Downarrow$ Score $\Uparrow$

α and β strings

Definition (Levensthein (edit) distance)

$d_L(\alpha, \beta)$: min number of Insertion Deletion Substitution

to convert a α into β

Distances and Scores

A "negative" or "positive" point of view

Distance $\Downarrow$ Score $\Uparrow$

α and β strings

Definition (Levensthein (edit) distance)

$d_L(\alpha, \beta)$: min number of Insertion Deletion Substitution

to convert a α into β

What are we searching for?

Alignment $\equiv$ Program

Let us simplify our lives

Definition (Hamming distance)

$d_H(\alpha, \beta)$: min number of Substitution

to convert a α into β

Let us simplify our lives

Definition (Hamming distance)

$d_H(\alpha, \beta)$: min number of Substitution

to convert a α into β

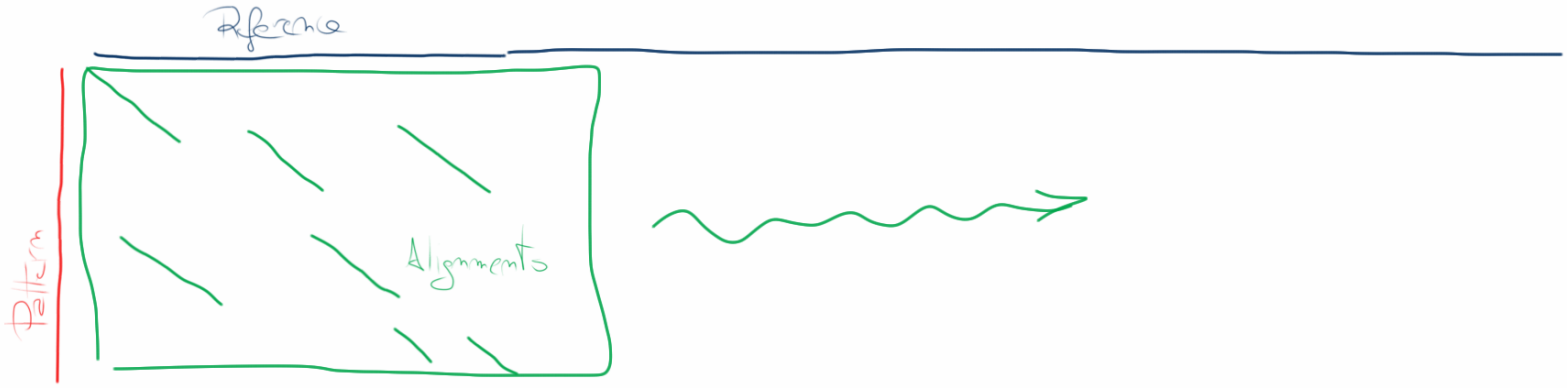
Complexity and gain

Hamming $\Rightarrow$ $|\alpha|$ (the length of α) must be equal to β
Levensthein $\Rightarrow$ $|\alpha|$ and $|\beta|$ may be different

Matrix-based alignment and *fast* alignments

Matrix-based alignments (Levensthein)

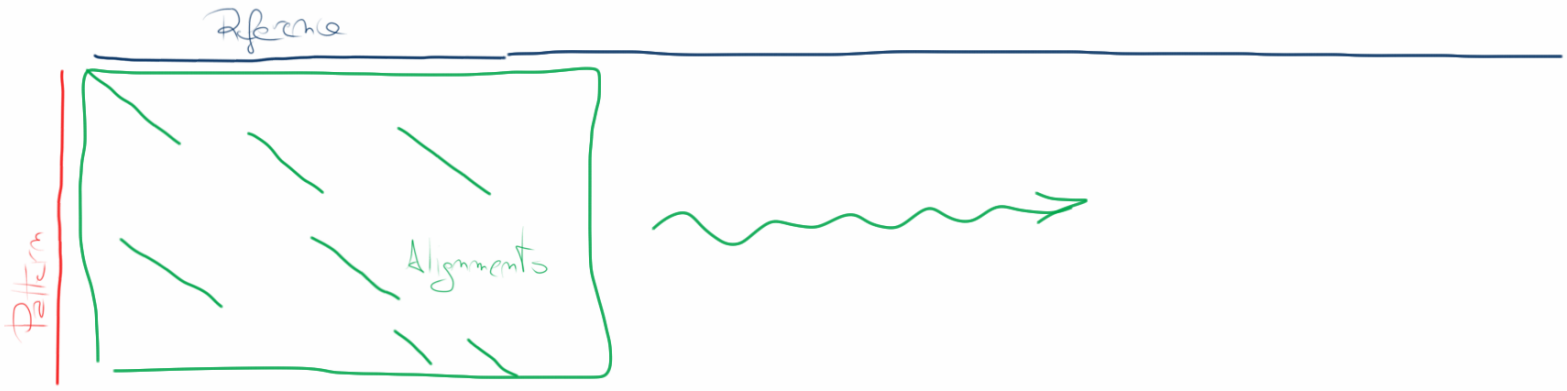
A weighted dot-plot computation of the *Pattern* against the *Reference* along the entire *R*



Matrix-based alignment and *fast* alignments

Matrix-based alignments (Levensthein)

A weighted dot-plot computation of the *Pattern* against the *Reference* along the entire *R*



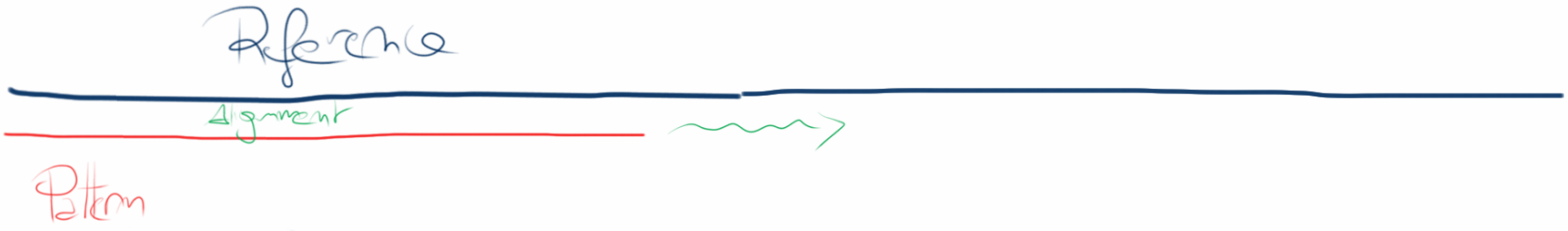
Cost

3G $\times$ 100 cells to scan (for Human)!

Matrix-based alignment and *fast* alignments

Fast alignments (Hamming)

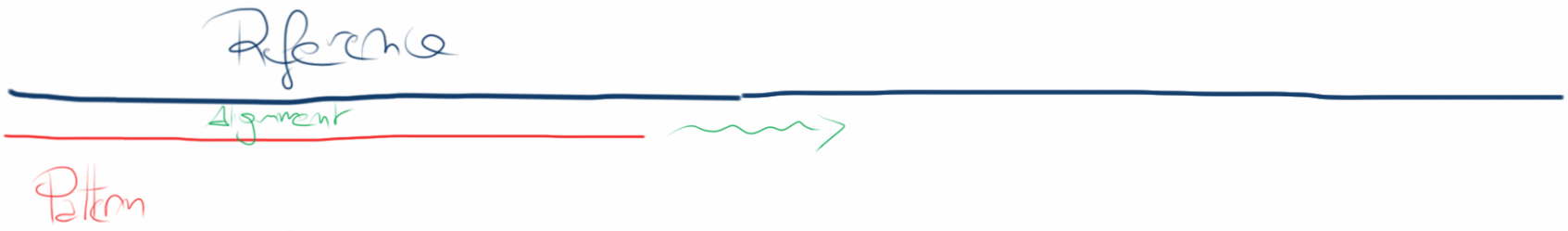
A scan of R with P stored in memory



Matrix-based alignment and *fast* alignments

Fast alignments (Hamming)

A scan of R with P stored in memory



Cost

time necessary to load the reference

Data structures

Observation

We can structure up R while loading it

Advantages: example

Ordering a set of 1G ($\approx 2^{30}$) numbers allows to search *any* element in no more than 30 steps

Data structures

Observation

We can structure up R while loading it

Advantages: example

Ordering a set of 1G ($\approx 2^{30}$) numbers allows to search *any* element in no more than 30 steps

Question

Can we order a text (like *R*)?

Two important techniques

Suffixes

R-suffixes can be ordered *lexicographically*

Hashing

R-blocks (*k*-mers) can be used as numbers to *index* an array

Two important techniques

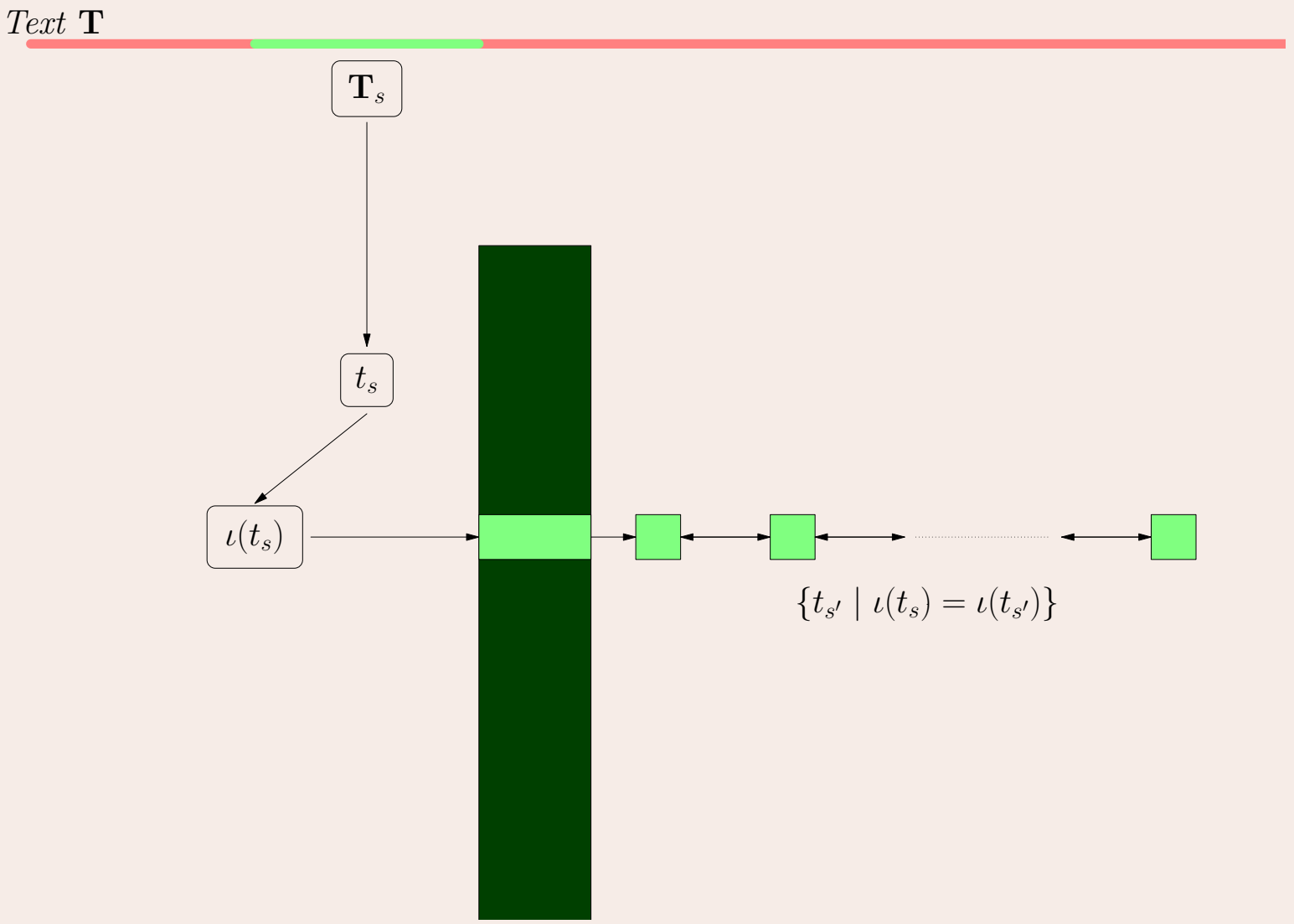
Suffixes

1	AGGTTGCCAGTGT
2	GGTTGCCAGTGT
3	GTTGCCAGTGT
4	TTGCCAGTGT
5	TGCCAGTGT
6	GCCAGTGT
7	CCAGTGT
8	CAGTGT
9	AGTGT
10	GTGT
11	TGT
12	GT
13	T

1	AGGTTGCCAGTGT
9	AGTGT
8	CAGTGT
7	CCAGTGT
6	GCCAGTGT
2	GGTTGCCAGTGT
12	GT
10	GTGT
3	GTTGCCAGTGT
13	T
5	TGCCAGTGT
11	TGT
4	TTGCCAGTGT

Two important techniques

Hashing



Compression

Ordered *R*-suffixes can be *compressed*: Burrows-Wheeler

AGGTTGCCAGTGT
AGTGT
CAGTGT
CCAGTGT
GCCAGTGT
GGTTGCCAGTGT
GT
GTGT
GTTGCCAGTGT
T
TGCCAGTGT
TGT
TTGCCAGTGT

Compression

Ordered *R*-suffixes can be *compressed*: Burrows-Wheeler

$
AGGTTGCCAGTGT$
AGTGT$
CAGTGT$
CCAGTGT$
GCCAGTGT$
GGTTGCCAGTGT$
GT$
GTGT$
GTTGCCAGTGT$
T$
TGCCAGTGT$
TGT$
TTGCCAGTGT$

Compression

Ordered *R*-suffixes can be *compressed*: Burrows-Wheeler

$AGGTTGCCAGTGT
AGGTTGCCAGTGT$
AGTGT$AGGTTGCC
CAGTGT$AGGTTGC
CCAGTGT$AGGTTG
GCCAGTGT$AGGTT
GGTTGCCAGTGT$A
GT$AGGTTGCCAGT
GTGT$AGGTTGCCA
GTTGCCAGTGT$AG
T$AGGTTGCCAGTG
TGCCAGTGT$AGGT
TGT$AGGTTGCCAG
TTGCCAGTGT$AGG

Compression

Ordered *R*-suffixes can be *compressed*: Burrows-Wheeler

$AGGTTGCCAGTGT
AGGTTGCCAGTGT$
AGTGT$AGGTTGCC
CAGTGT$AGGTTGC
CCAGTGT$AGGTTG
GCCAGTGT$AGGTT
GGTTGCCAGTGT$A
GT$AGGTTGCCAGT
GTGT$AGGTTGCCA
GTTGCCAGTGT$AG
T$AGGTTGCCAGTG
TGCCAGTGT$AGGT
TGT$AGGTTGCCAG
TTGCCAGTGT$AGG

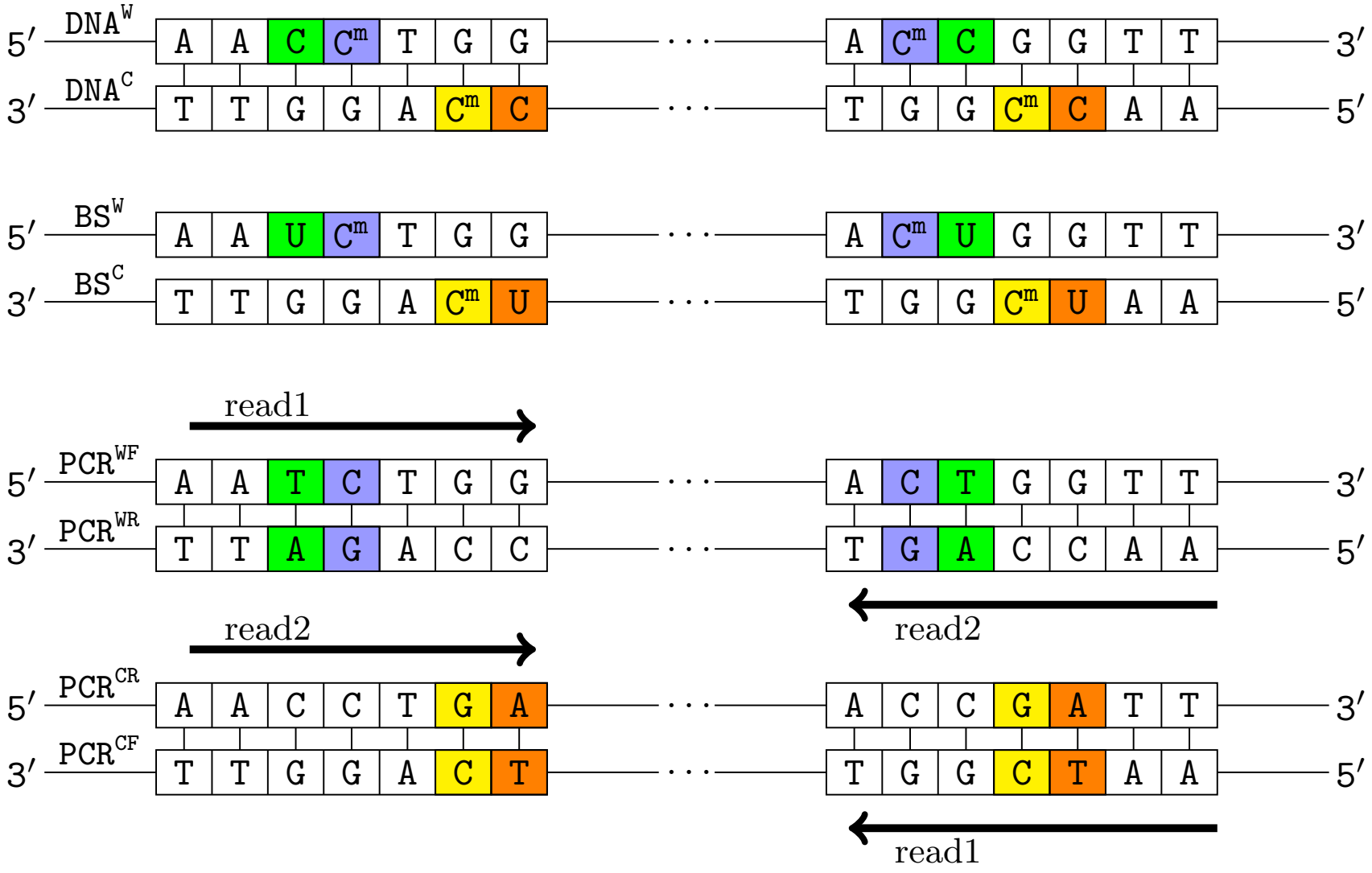
Aligning BS-treated sequences: the problem

Determine Methylated C's (i.e. C^m's) along the reference R

- Sodium Bisulphite converts un-methylated cytosines to uraciles
- Methylated cytosines remain un-converted

$\Rightarrow$ after PCR un-methylated cytosines appear as thymines and methylated ones remain unaltered

Aligning BS-treated sequences: the problem



Aligning BS-treated sequences: the problem

Determine Methylated C's (i.e. C^m's) along the reference R

- Sodium Bisulphite converts un-methylated cytosines to uraciles
- Methylated cytosines remain un-converted

$\Rightarrow$ after PCR un-methylated cytosines appear as thymines and methylated ones remain unaltered

Two (main) issues

1. *some* mismatch conveys information
2. we are working with a 5-characters alphabet

Characters and mismatches

"Natural" approach

Ignore $T - C$ mismatches
($\Rightarrow$ identify T's and C's $\Rightarrow$ use a 3-characters alphabet!)

Characters and mismatches

"Natural" approach

Ignore $T - C$ mismatches
($\Rightarrow$ identify T's and C's $\Rightarrow$ use a 3-characters alphabet!)

Hamming distance to "render" methylated C's

d_{GH}: the Generalized Hamming distance for BS-treated sequences, assigns

0 distance to T-read/C-reference

mismatches (as well as G-read/A-reference)

ERNE-BS5

d_{GH} indirectly reduces the alphabet to 3 characters

- Many reads do not reach a threshold to be reliably aligned
- Many misalignment

Idea: use the *fifth* character to improve alignment

Methylated *C*'s can be used to disambiguate multiple alignments

ERNE-BS5

N_i^C number of C's read at position i; N_i^T number of T's read at position i;

Definition (Methylation level)

$$\mu(i) = \frac{N_i^C}{N_i^C + N_i^T}$$

Definition (Context-aware Hamming distance d_α)

Watson strand:

$$\alpha(i,x) = \begin{cases} 1-\mu(i) & \text{if } R[i] = C \wedge x = C; \\ \mu(i) & \text{if } R[i] = C \wedge x = T; \\ 0 & \text{otherwise.} \end{cases}$$

Crick strand analogous.

ERNE-BS5

The full strategy

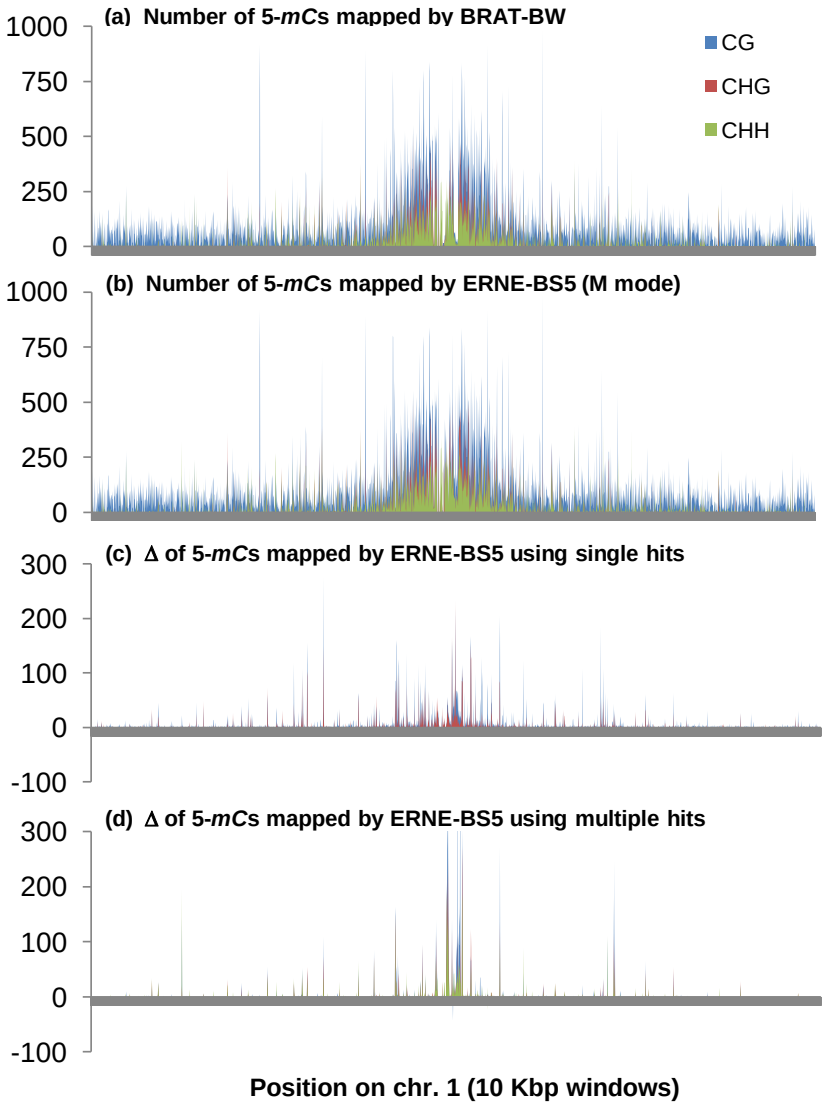
step 1 Align reads using d_{GH}

step 2 On-line compute d_α

step 3 Align reads multiply aligning using d_α

step 4 If new alignments are found, go to step 2

Results on *Arabidopsis thaliana*, real data



Further work

- Extensively test ERNE-BS5
- Encode more in d_α
- Integrate more knowledge
- Study the *two* methylation's patterns
- Compress