ISTITUTO DI SCIENZE E TECNOLOGIE DELL'INFORMAZIONE

Computer-Aided Base Calling from Forward and Reverse Electropherograms

Valerio Freschi Alessandro Bogliolo

Istituto di Scienze e Tecnologie dell'Informazione (STI)
University of Urbino

Outline

- Introduction and scope
  - Definitions
  - Problem statement
  - DNA Base Calling from Forward and Reverse Electropherograms
- Proposed approaches
  - Consensus Generation after Base Calling (CGaBC)
  - Base Calling after Trace Merging (BCaTM)
    - Sample-Driven BCaTM (SD BCaTM)
    - Base-Driven BCaTM (SD BCaTM)
- Results and Conclusions

Scope

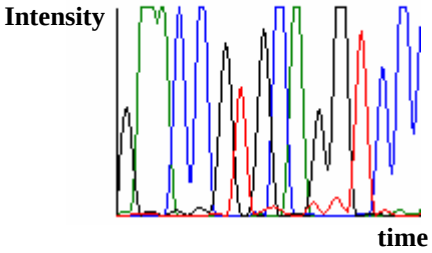
Sequencing experiment

Sanger's method

Forward and Reverse sequencing experiments for the same sample: i.e.. sequencing from opposite ends

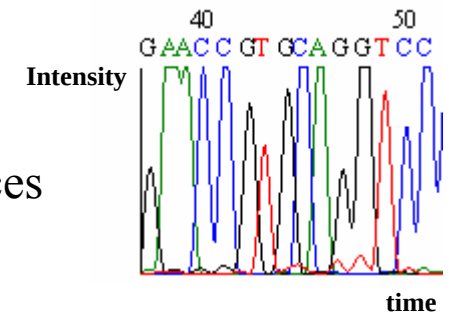
Produces electropherograms representing the detection of bases at given positions

4-component time series



Base calling

Decodes electropherogram traces in base sequences



Sequence analysis

Compares, aligns, studies called sequences

```
Query GAACCGT-GCAGGTCC
Entry GAACCGTTGCAGGTCC
```

Base calling

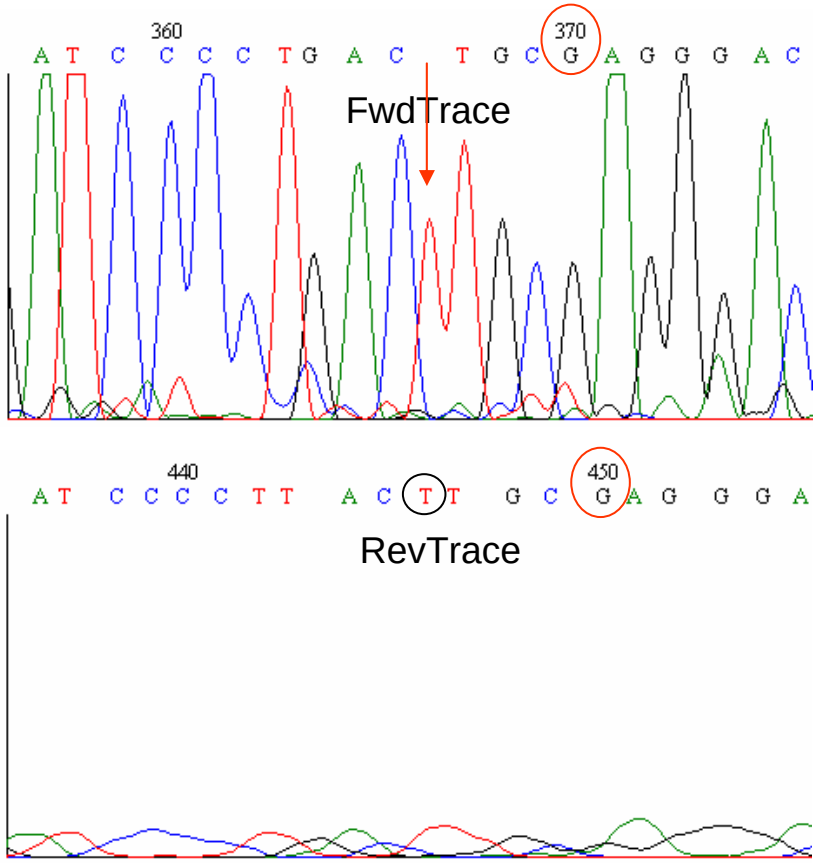
- Base calling is a computer process used for converting traces detected from automated sequencers to an inferred base read

- Base calling usually involves several steps that can be summarized as:
  - trace preprocessing (noise filtering, mobility shift, …)
  - trace decoding (actual base calling) ←

Observations

Looking at traces helps sequence alignment

```
Real sequence CCCTGACTTGCGA
Corr. Align.  CCCTGAC-TGCGA
Wrong Align.  CCCTGACT-GCGA
```



Sequence alignment helps trace comparison

```
Base 370 of FwdTrace can be
compared with base 450 of
RevTrace
```

Trace comparison helps base calling

```
RevTrace tells us that the
gap in FwdTrace is a T
```

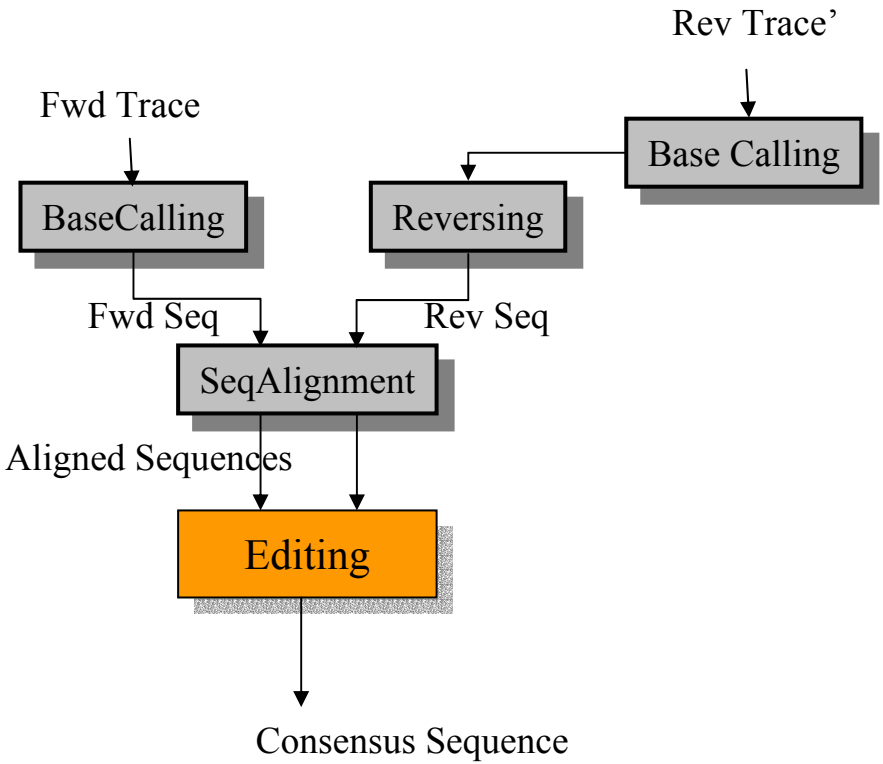
Outline

- Introduction and scope
  - Definitions
  - Problem statement
  - DNA Base Calling from Forward and Reverse Electropherograms
- Proposed approaches
  - Consensus Generation after Base Calling (CGaBC)
  - Base Calling after Trace Merging (BCaTM)
    - Sample-Driven BCaTM (SD BCaTM)
    - Base-Driven BCaTM (SD BCaTM)
- Results and Conclusions

Consensus Generation after Base Calling (CGaBC)

1. Call the bases of Forward trace and Reverse trace
2. Align the sequences obtained from Forward trace and Reverse trace
3. From alignment and comparison infer real sequence

CGaBC tool flow



ACGT trace samples Fwd and Rev
(A_n,C_n,G_n,T_n)

Called bases Fwd and Rev
$(Base_i,Position_i)$

Aligned base sequences Fwd and Rev
$(Base_i,Position_i,Order_i)$

Called bases
$(Base_j,Position_j)$

CGaBC Consensus

1. Most base callers assign with each base a quality value
2. The quality value is a measure of a correctness probability: $q=-10\log_{10}(p)$
3. The CGaBC approach consists of taking automated decisions based on base qualities
4. Trivial case: matching bases are assigned to consensus
5. Mismatch handling: base with higher quality is assigned to the consensus
6. Gaps handling: quality values assigned also to gaps by averaging qualities of preceding and following calls

Outline

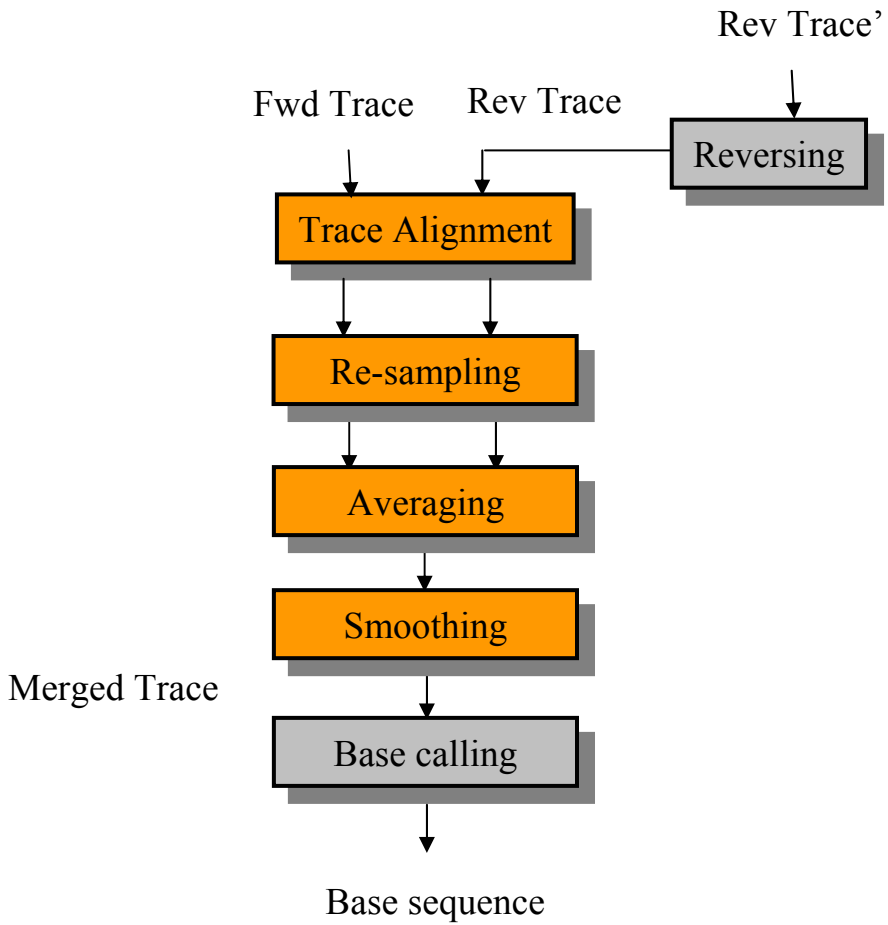
- Introduction and scope
  - Definitions
  - Problem statement
  - DNA Base Calling from Forward and Reverse Electropherograms
- Proposed approaches
  - Consensus Generation after Base Calling (CGaBC)
  - Base Calling after Trace Merging (BCaTM)
    - Sample-Driven BCaTM (SD BCaTM)
    - Base-Driven BCaTM (SD BCaTM)
- Results and Conclusions

Base Calling after Trace Merging (BCaTM)

1. Average the available electropherograms
2. Apply base calling to the combined trace

- Issue: averaging of independent traces.
- Proposed procedures:
  - Sample-driven alignment
  - Base-driven alignment

BCaTM tool flow



ACGT trace samples Fwd and Rev
(A_n,C_n,G_n,T_n)

Aligned trace samples Fwd and Rev
on common axis (x_n, A_n,C_n,G_n,T_n)

Re-sampled traces Fwd and Rev
(A_n,C_n,G_n,T_n)

Averaged trace samples Fwd and Rev
(A_n,C_n,G_n,T_n)

Called bases
($Base_j$,$Position_j$)

Outline

- Introduction and scope
  - Definitions
  - Problem statement
  - DNA Base Calling from Forward and Reverse Electropherograms
- Proposed approaches
  - Consensus Generation after Base Calling (CGaBC)
  - Base Calling after Trace Merging (BCaTM)
    - Sample-Driven BCaTM (SD BCaTM)
    - Base-Driven BCaTM (SD BCaTM)
- Results and Conclusions

Sample Driven trace alignment

1. Trace labeling
   - The 4 components of each sample are sorted and compared with preceding and subsequent samples. Labels of 4 character are assigned with each sample to represent the order of its components and the occurrence of local maxima
2. Dynamic programming alignment
   - The basic Needleman-Wunsch algorithm for sequence alignment is extended to handle labels of 4 characters each, instead of single symbols

Trace labeling

- Label: ordered set of 4 symbols taken from {A,a,C,c,G,g,T,t}.
  - e.g: k-th sample

$$A_k = 100,\ C_k = 150,\ G_k = 0,\ T_k = 850$$

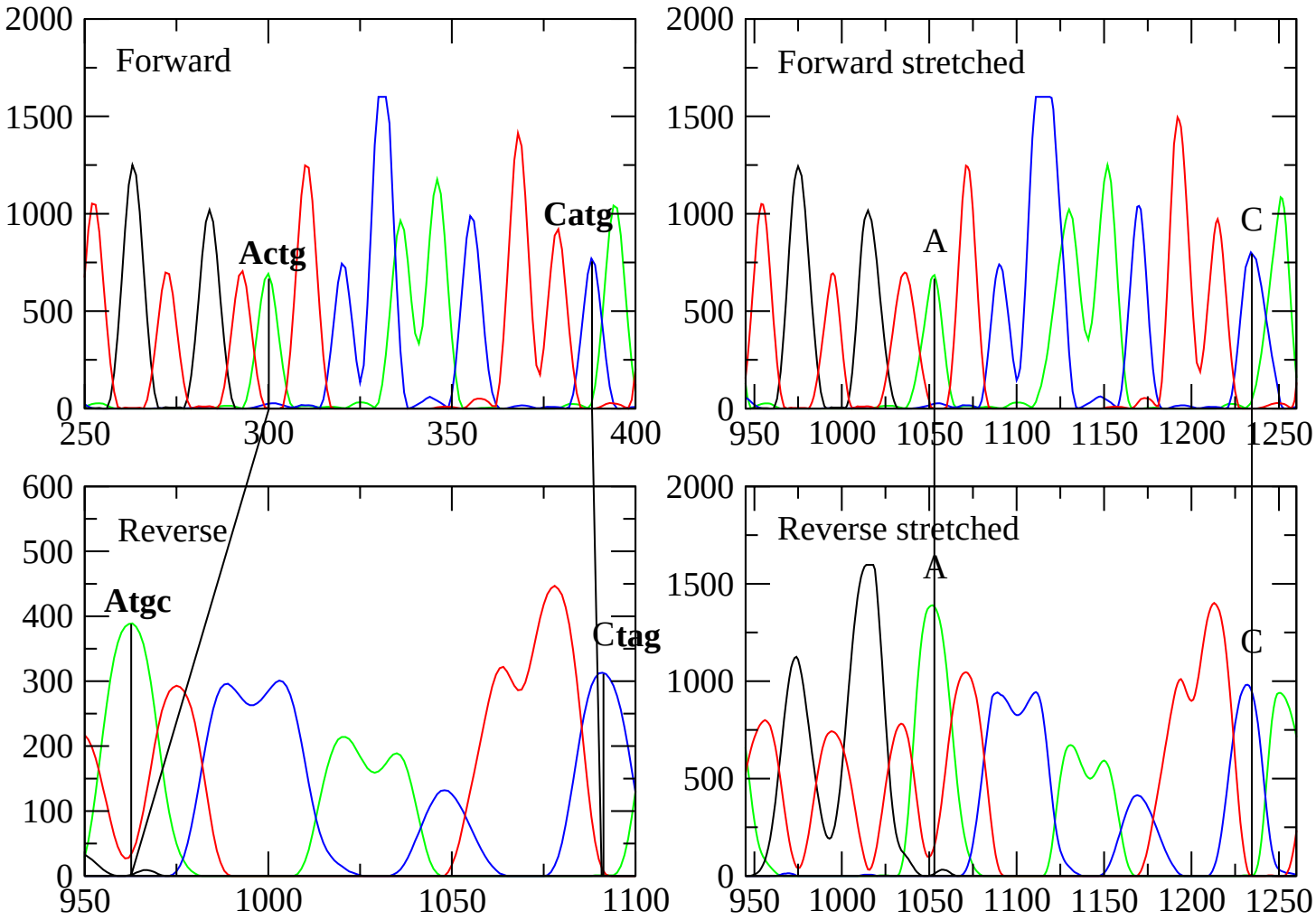
where $C_{k-1} \leq C_k$ and $C_k \geq C_{k+1}$

is assigned with label "tCag"

Capital letters denote local maxima

Dynamic programming trace alignment

- Inputs:
  - two sequences of labels
  - a substitution score table (*ST*)
- Algorithm:
  - NW-like algorithm aligning the two sequences in order to maximize the global score according to the score table
- Output:
  - aligned traces (gaps are filled by component-wise linear interpolation)


Forward
Actg
Catg
Forward stretched
A
C
Reverse
Atgc
Ctag
Reverse stretched
A
C

Outline

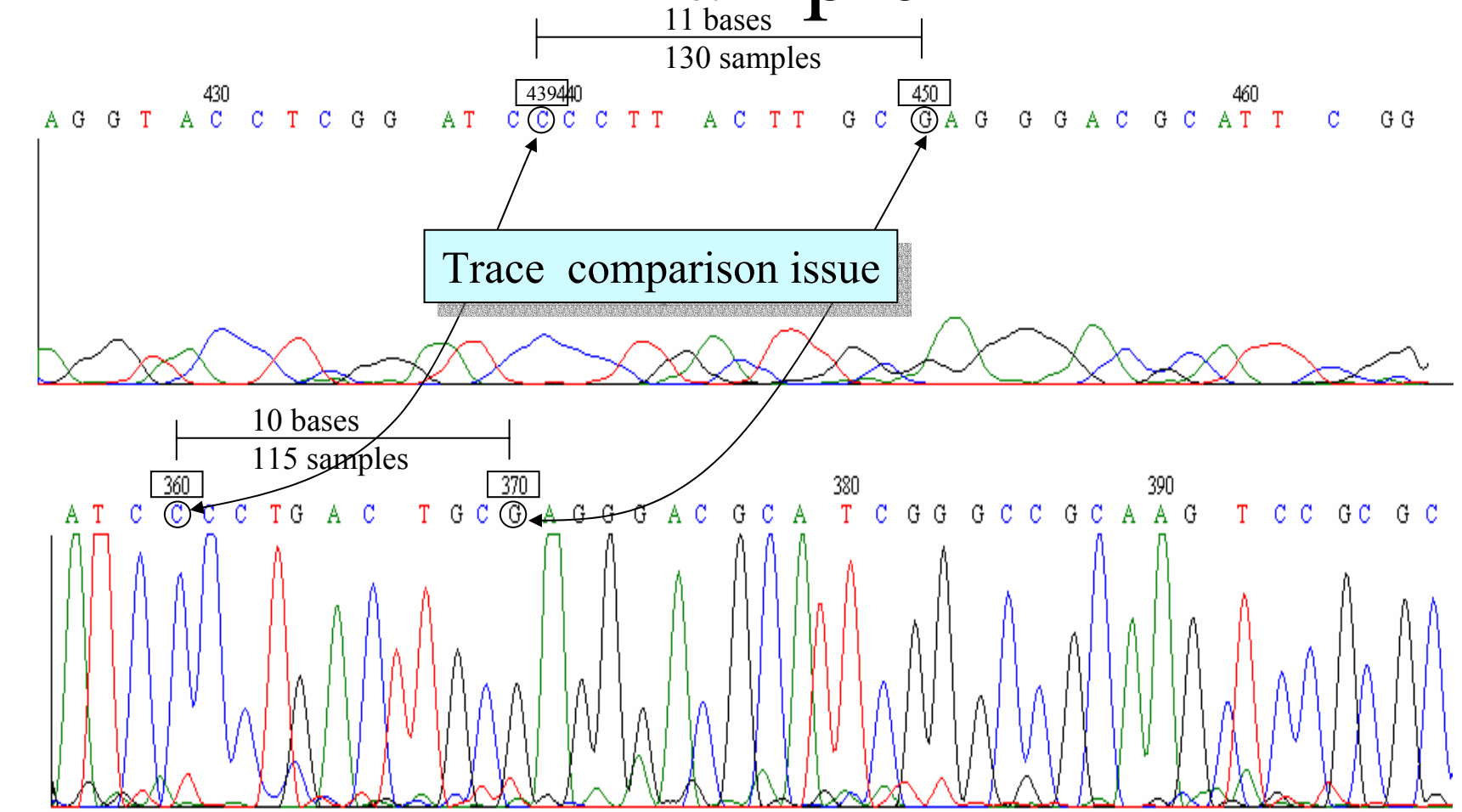
- Introduction and scope
  - Definitions
  - Problem statement
  - DNA Base Calling from Forward and Reverse Electropherograms
- Proposed approaches
  - Consensus Generation after Base Calling (CGaBC)
  - Base Calling after Trace Merging (BCaTM)
    - Sample-Driven BCaTM (SD BCaTM)
    - Base-Driven BCaTM (SD BCaTM)
- Results and Conclusions

Base Driven trace alignment

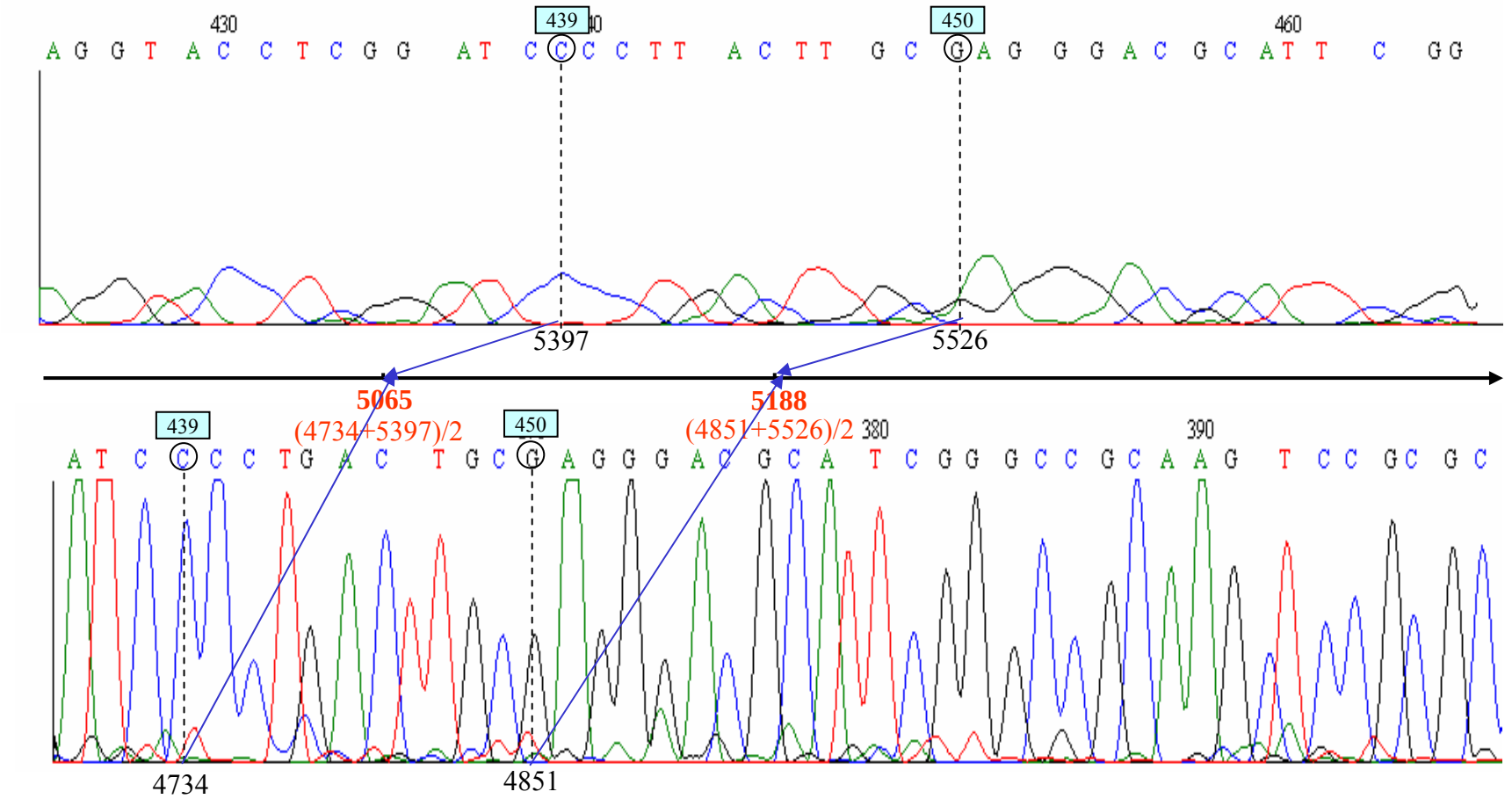
1. Base calling is performed on each trace
2. Dynamic programming alignment:

   aligned bases are annotated with their positions in the original traces
3. Bases are repositioned on a common axis, so that homologous bases have the same position
4. Original traces are stretched (shrunk) according to the position of aligned bases on the common x axis.

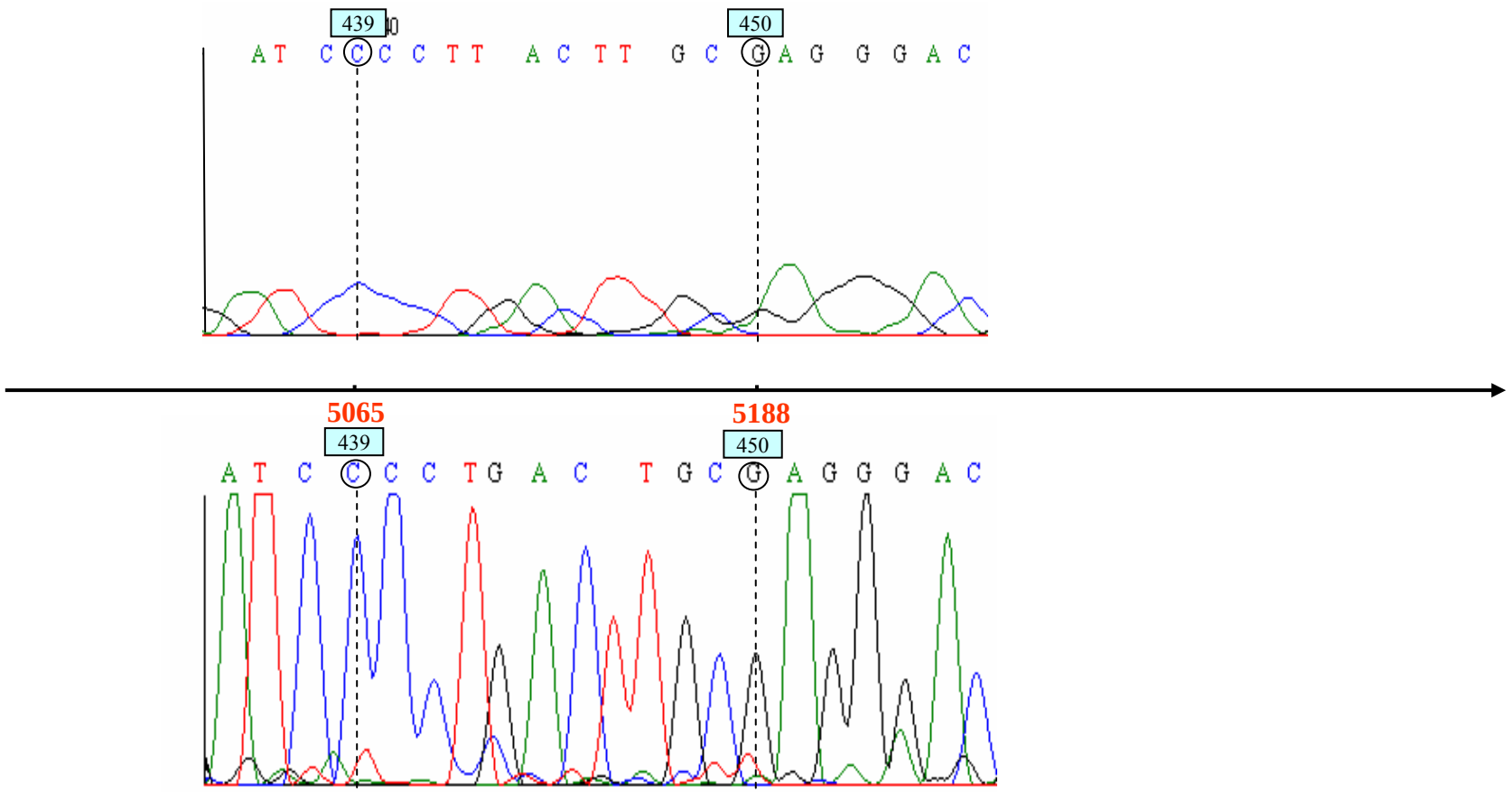
Example



Example



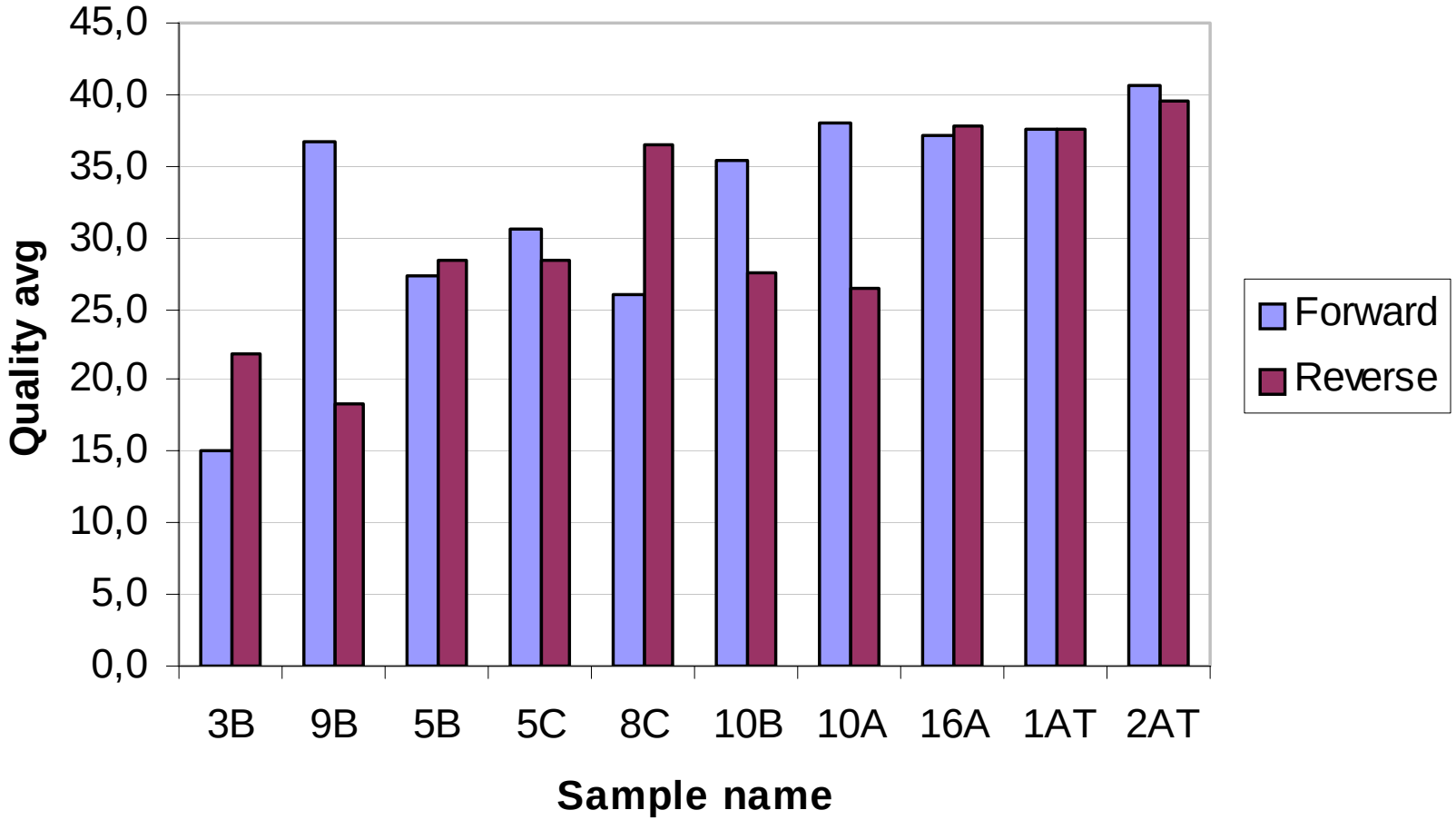
Example



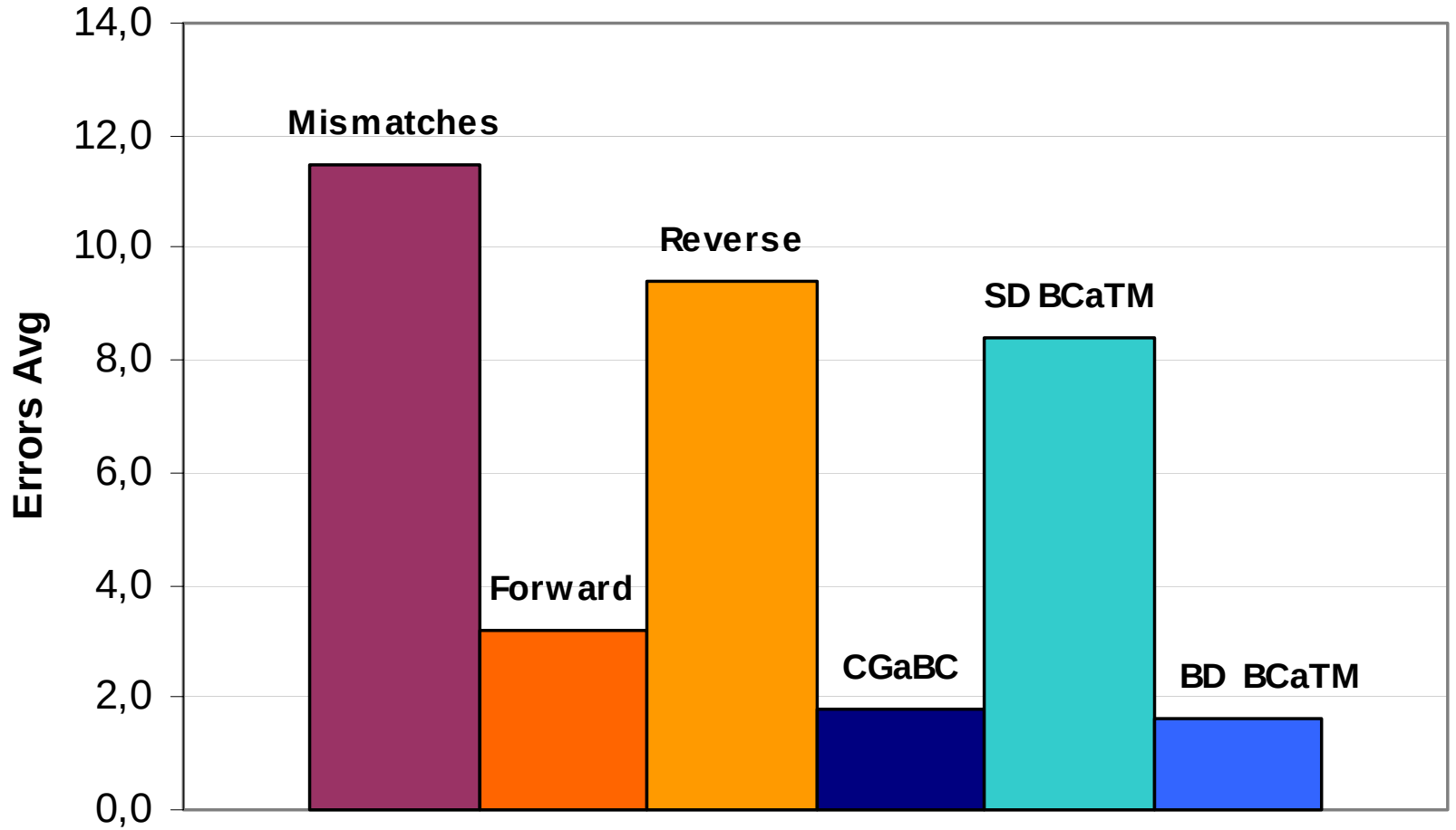
Results and conclusions

- Experimental results:
  - All approaches has been tested on a set of known DNA samples
  - Accuracy has been evaluated in terms of:
    - Number of errors (E)
    - Max. quality value assigned to a wrong call (MQ_w)
    - Min. quality value assigned to a correct call (mQ_c)

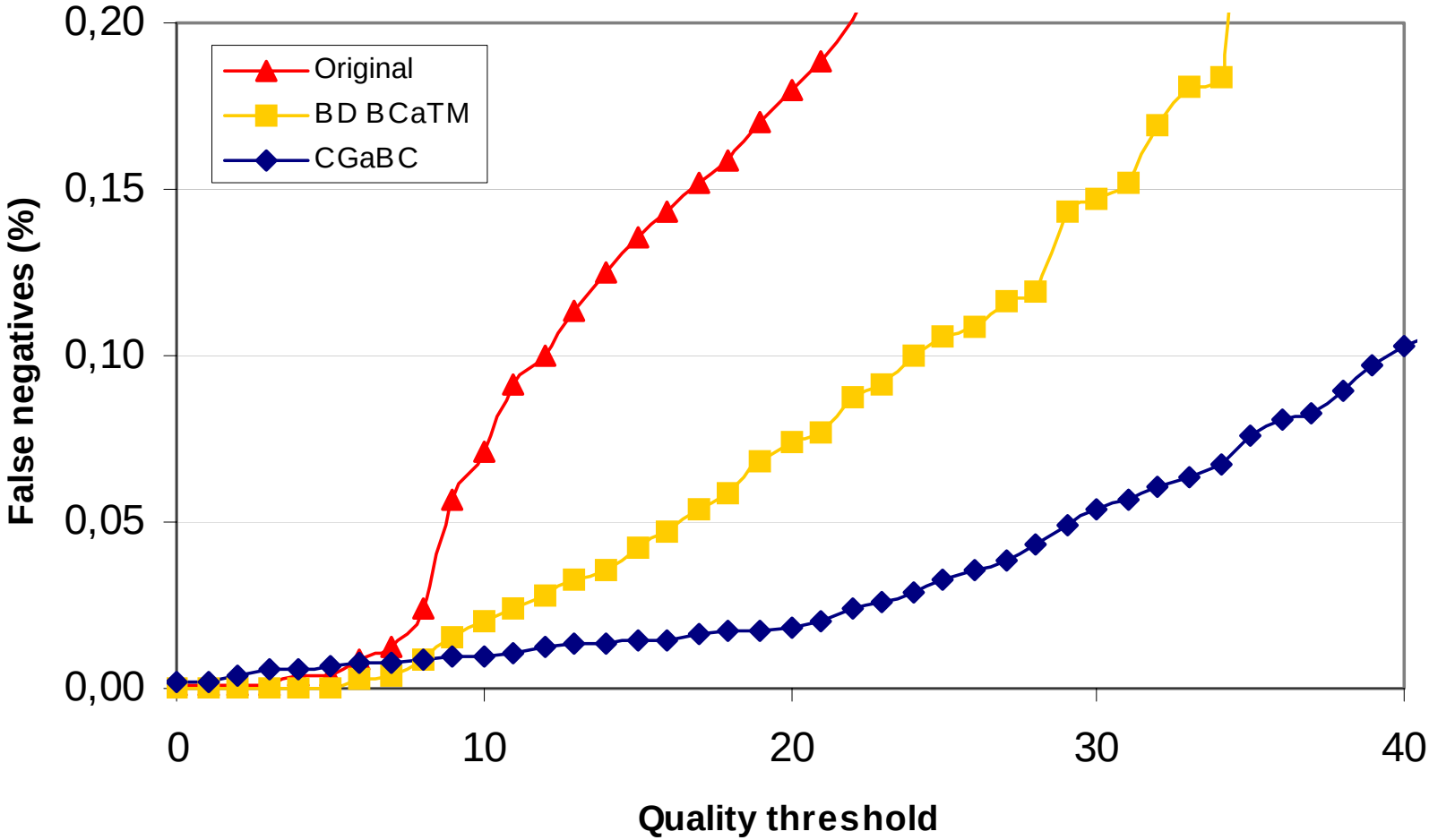
Benchmark set



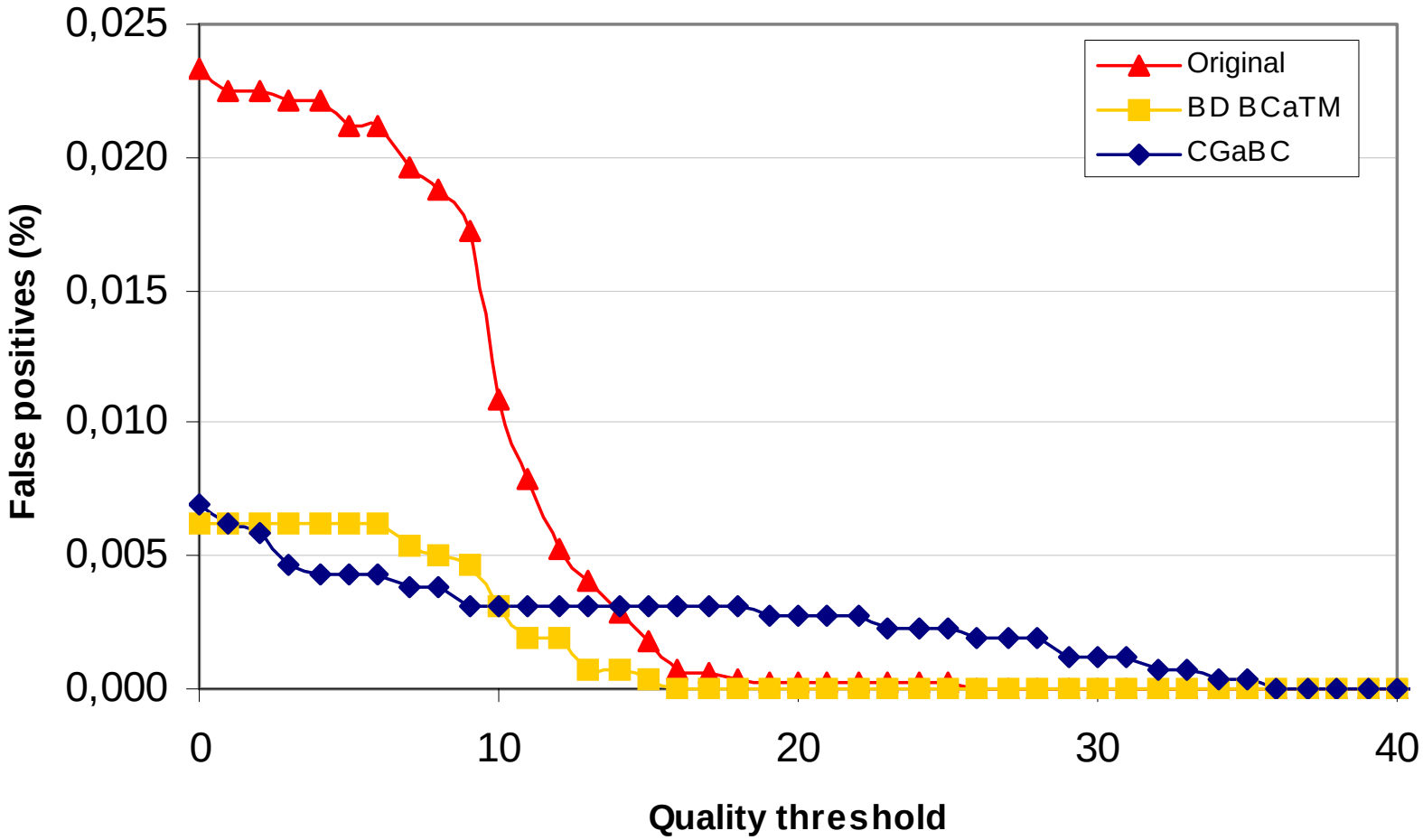
Results and conclusions



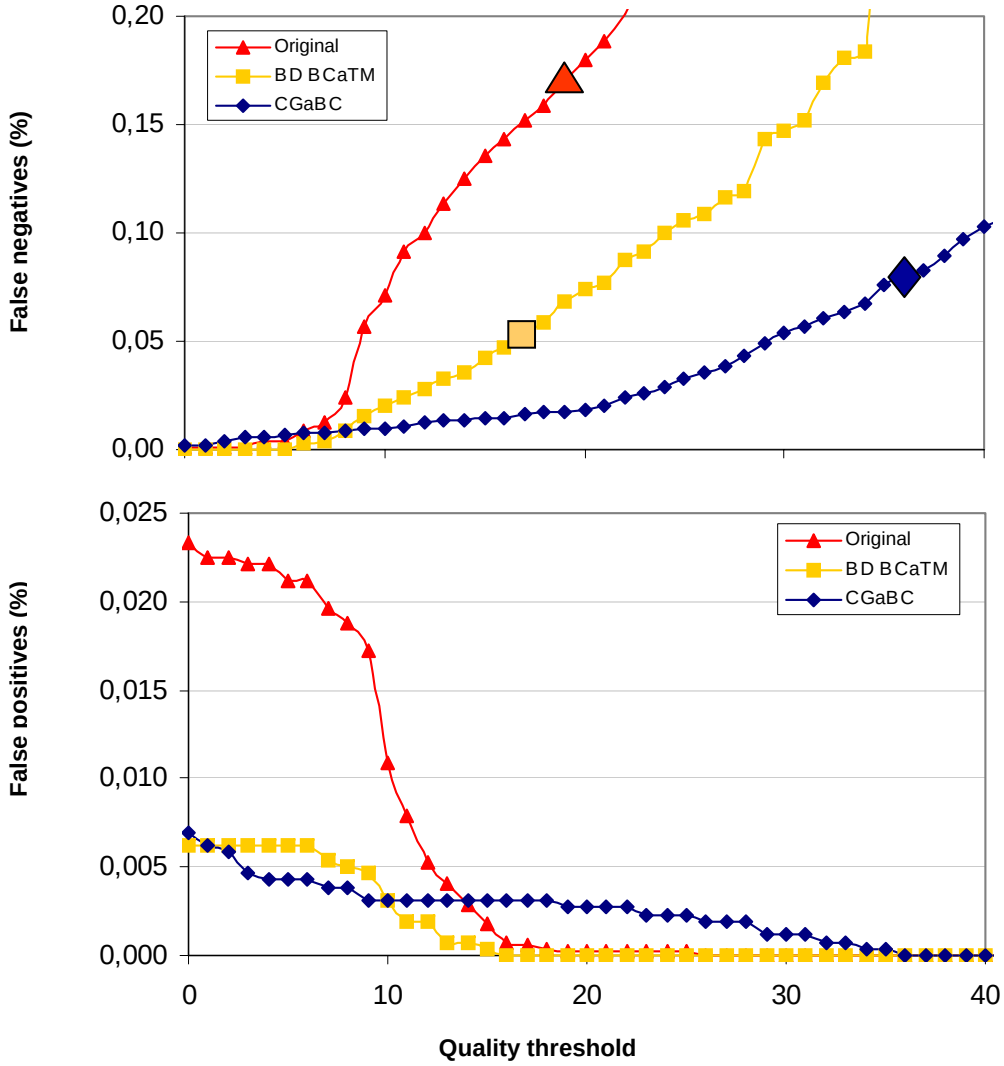
Results and conclusions



Results and conclusions



Results and conclusions



Results and conclusions

- Computer-aided approaches can minimize the need for human intervention
- Trace combination may improve base calling accuracy
- The improved quality of averaged electropherograms makes it easier to discriminate between correct and incorrect base calls