A **3D** pattern matching algorithm for DNA sequences

Mikk Eelmets

Journal Club
07.05.2007

DNA structure

- Biologists usually work with textual DNA sequences (A, C, G, T).
- Linear coding offers only a local and a one-dimensional vision of the molecule.
- The 3D structure of DNA is known to be very important in many essential biological mechanisms.

Observing DNA molecule

- Two experimental methods:

1. X-ray crystallography
2. nuclear magnetic resonance (NMR)

- NMR is limited to small molecules (<30 kDa), for bigger molecules, there is X-ray crystallography

3D conformation models

- Construct a 3D trajectory of a naked DNA molecule from its textual sequence.
- Do not represent DNA wrapping around nucleosomes and high level of folding inside cell.
- Provides for each dinucleotide three angular values and a raise translation.

ADN-Viewer

- INPUT - textual DNA sequences and the 3D conformation.
- OUTPUT - the 3D coordinates of each nucleotide.

3D Visualization of Scerevisiae chrMT (~50 Kbp). Genes are displayed in white and intergenic areas are in black.

Pattern matching definition

- to find all the positions of a motif M of size m in a sequence T of size n.

First stage – definition of a 3D comparison

- Data are represented by the succession of 3D coordinates or succession of 3D vectors.
- 3D coordinates transforming into angles
- One-by-one method of displacement of the motif along the sequence

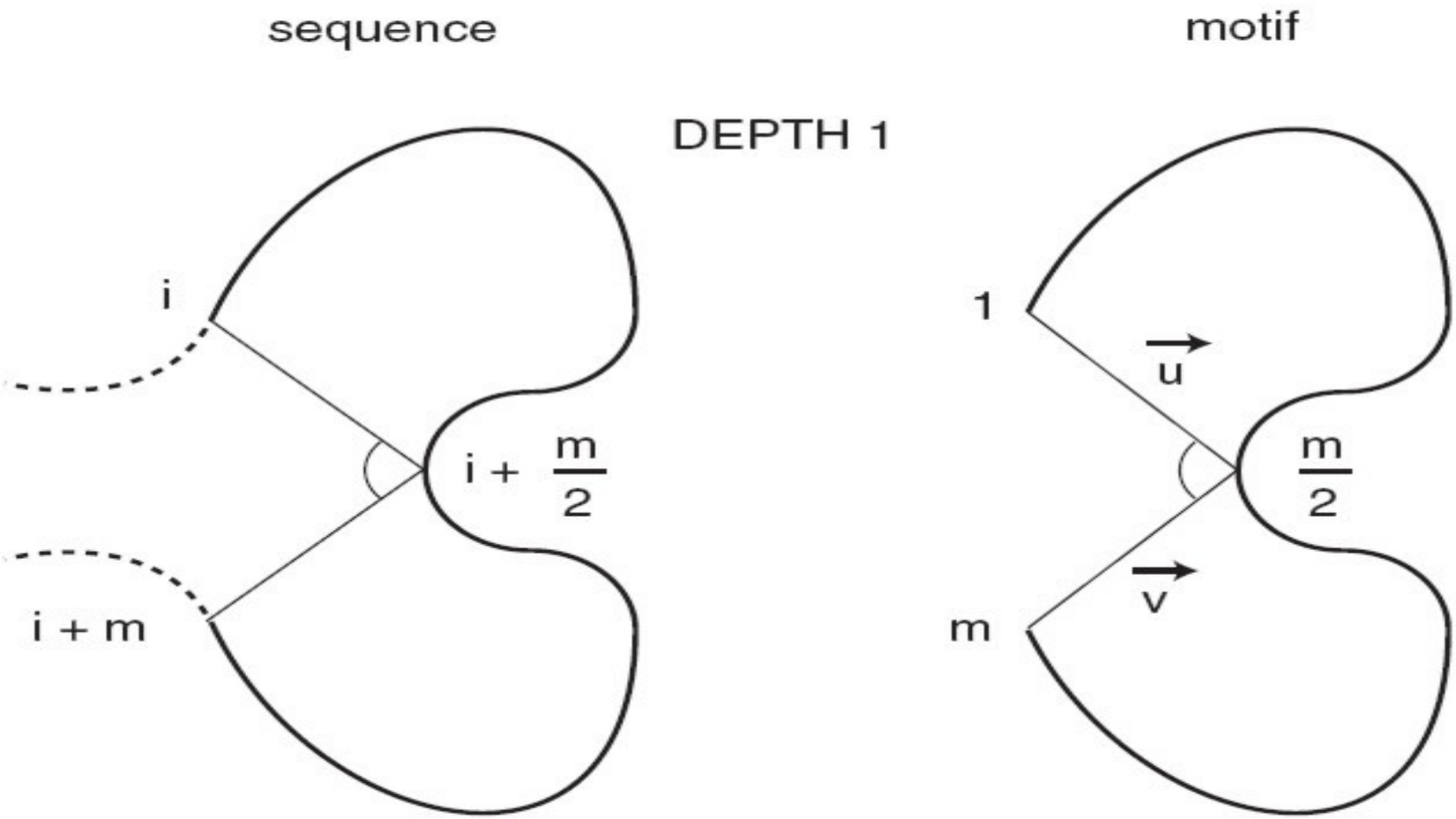
Second stage—definition of angles equality

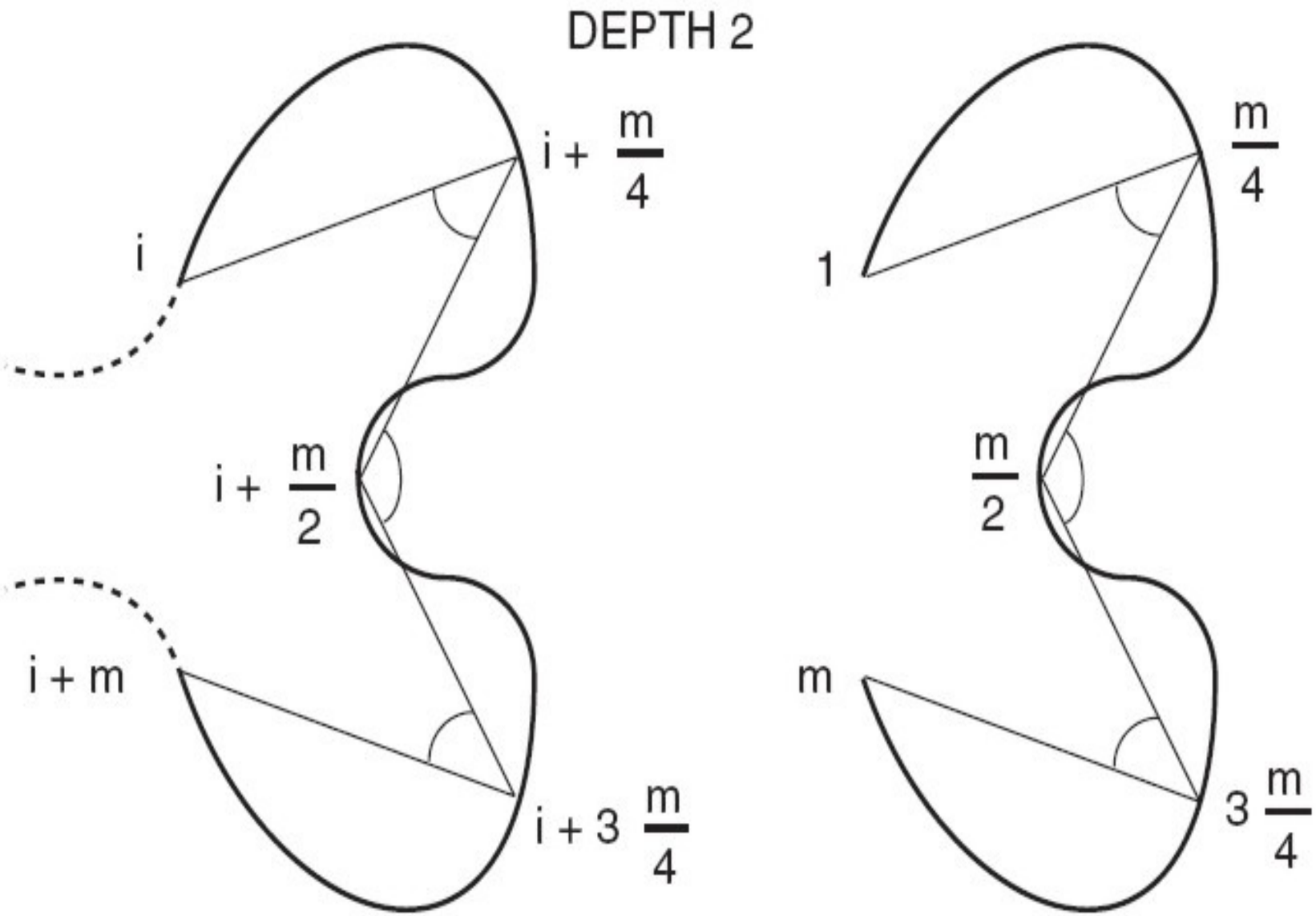
- The angles of motif and sequence will be equal when both sequences have the same succession of nucleotides
- Flexible comparison - error parameter ε
- Strong similarity – small values of ε
- Less selective detection –greater values of ε

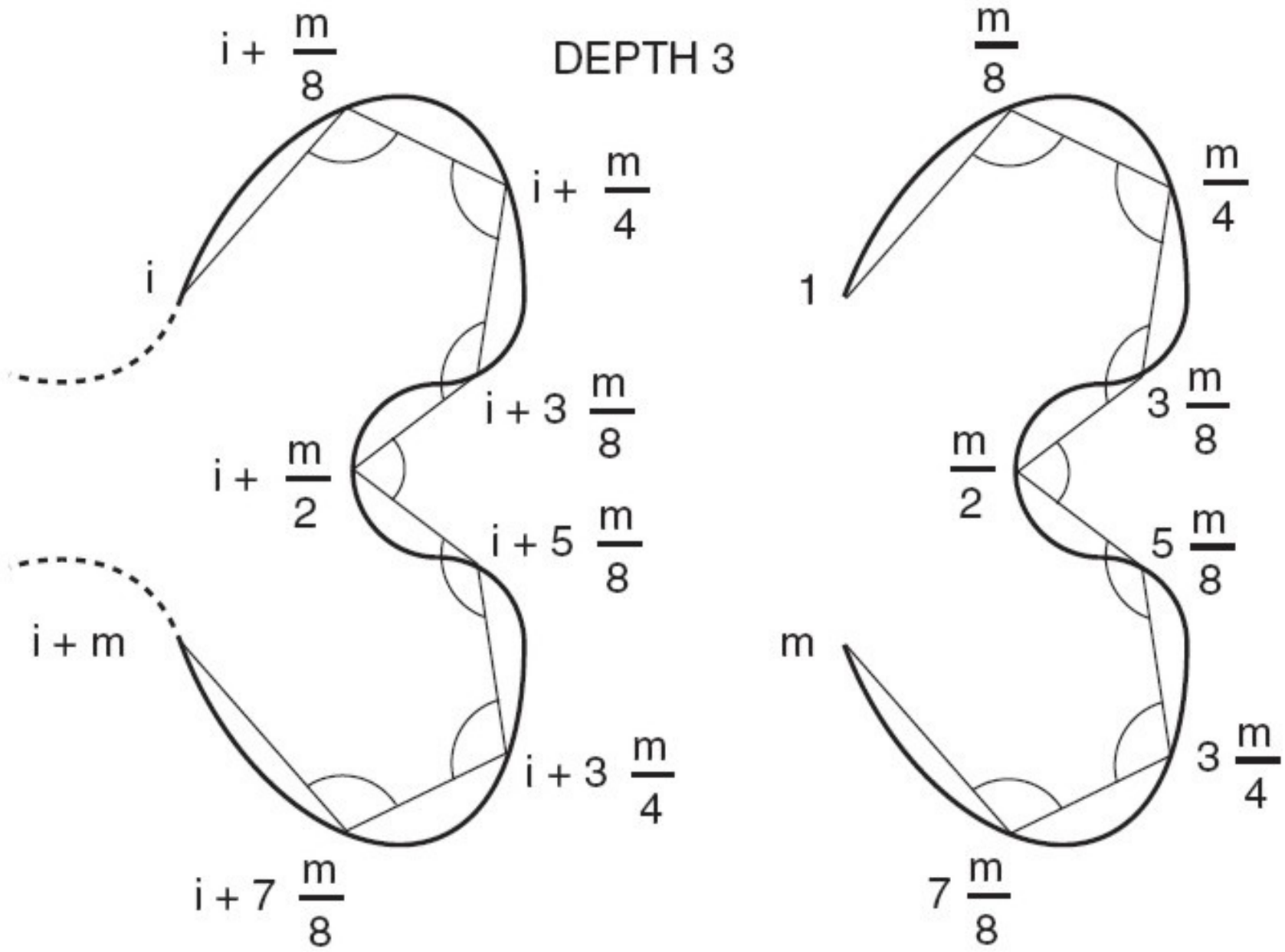
The approach

- For each motif position on the sequence and the each position on the fragment of the compared sequence, the algorithm calculates the angle
- Uses vector-based cutting by dichotomy

$$\widehat{(\vec{u}, \vec{v})} \quad \text{where} \quad \vec{u} = \begin{pmatrix} x_{\frac{m}{2}} - x_1 \\ y_{\frac{m}{2}} - y_1 \\ z_{\frac{m}{2}} - z_1 \end{pmatrix} \quad \text{and} \quad \vec{v} = \begin{pmatrix} x_m - x_{\frac{m}{2}} \\ y_m - y_{\frac{m}{2}} \\ z_m - z_{\frac{m}{2}} \end{pmatrix}$$




DEPTH 2
i
i + m/4
i + m/2
i + m
i + 3 m/4
1
m/4
m/2
m
3 m/4

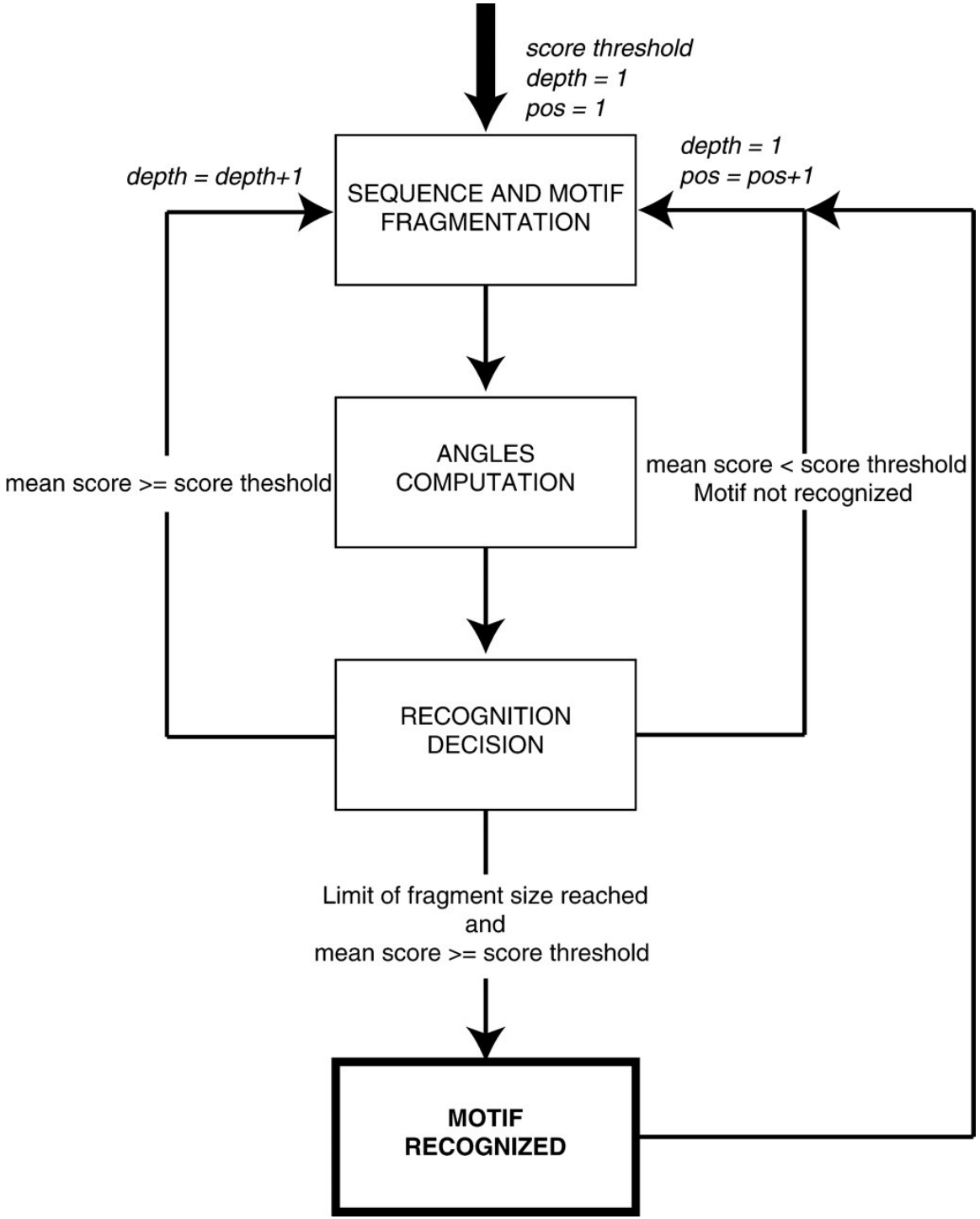

DEPTH 3
i + m/8
i + m/4
i
i + 3 m/8
i + m/2
i + 5 m/8
i + m
i + 3 m/4
i + 7 m/8
m/8
m/4
1
3 m/8
m/2
5 m/8
m
3 m/4
7 m/8

The approach

- The angles are compared in linear way
- STOP

  - angels are too different

  - vecotrs formed by 10 nucleotides
- Comparison score 0..100


score threshold
depth = 1
pos = 1
depth = depth+1
SEQUENCE AND MOTIF FRAGMENTATION
depth = 1
pos = pos+1
ANGLES COMPUTATION
mean score >= score theshold
mean score < score threshold
Motif not recognized
RECOGNITION DECISION
Limit of fragment size reached
and
mean score >= score threshold
MOTIF RECOGNIZED

The approach

- User defines two parameters:
  - the score threshold
  - the bonus percentage

Benefit of 3D approach

- to discover hidden phenomena from the textual sequence
- to reveal phenomena easier/faster, as compared to textual sequence

Results

- *Arabidopsis thaliana*
  - five choromosomes
  - 157 million bp
  - 25500 genes found so far
  - first sequenced plant genome, in 2000

Results

AT3G24310: 1232 bp

A. thaliana chromosome 3

match = 8 811 138 – 8 812 369: **100%** of 3D similarity

(AT3G24310 gene: auto-detection)

match = 21 348 410 – 21 349 641: **90.9**% of 3D similarity

(AT3G57620 gene: 21 348 541 – 21 350 278)

A. thaliana chromosome 2

(none)

A. thaliana chromosome 1

match = 2 564 260 – 2 565 491: **90.3%** of 3D similarity

(AT1G08180 gene: 2 564 738 – 2 565 073)

A. thaliana chromosome 4

(none)

A. thaliana chromosome 5

(none)

Results

NCBI **Blast 2 Sequences results**

PubMed Entrez BLAST OMIM Taxonomy Structure

BLAST 2 SEQUENCES RESULTS VERSION BLASTN 2.2.15 [Oct-15-2006]

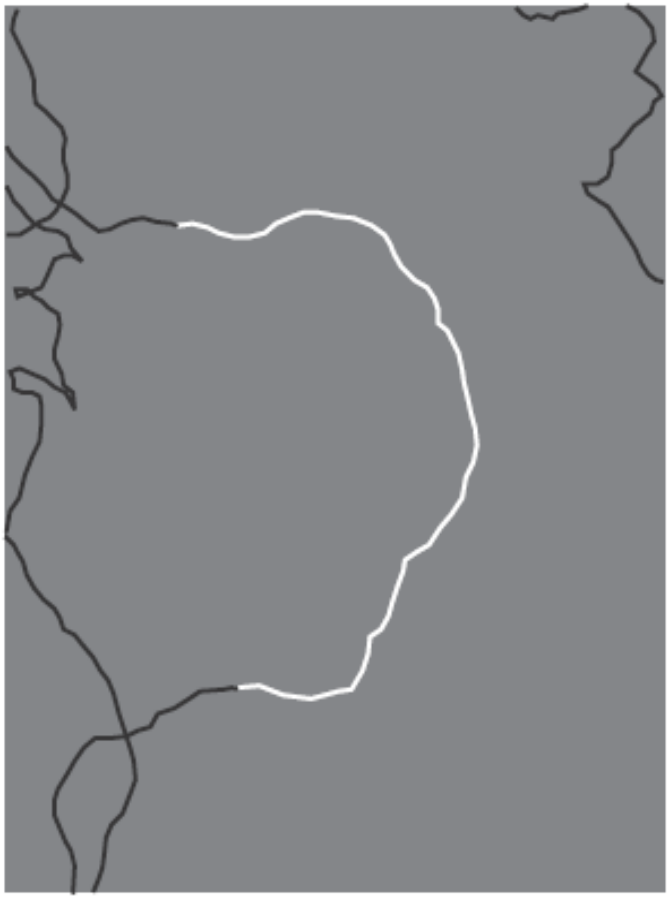
Match: 1 Mismatch: -2 gap open: 5 gap extension: 2
x_dropoff: 50 expect: 10.00 wordsize: 11 Filter View option Standard
Masking character option X for protein, n for nucleotide Masking color option Black
Show CDS translation Align

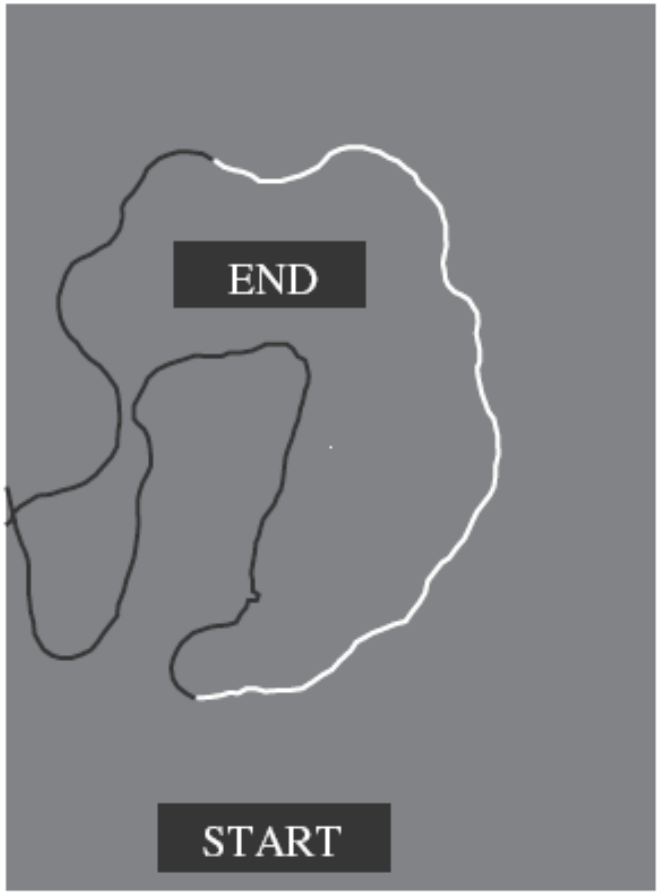
Sequence 1: lcl|1_seq_1
Length = 1232

Sequence 2: lcl|2_seq_2
Length = 1738

No significant similarity was found

Results





Reference

Herisson J,Payen G, Gherbi R.

A 3D pattern matching algorithm for DNA sequences

Bioinformatics. 2007 Mar 15;23(6):680-6

THANK YOU