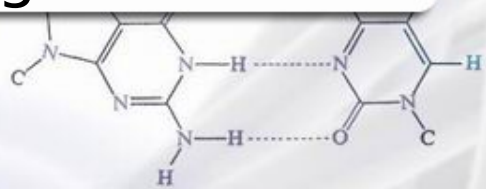
Computing the exact p-value for structured motif

Zhang Jing (Tsinghua University and university of waterloo)

Co-authors: Xi Chen, Ming Li



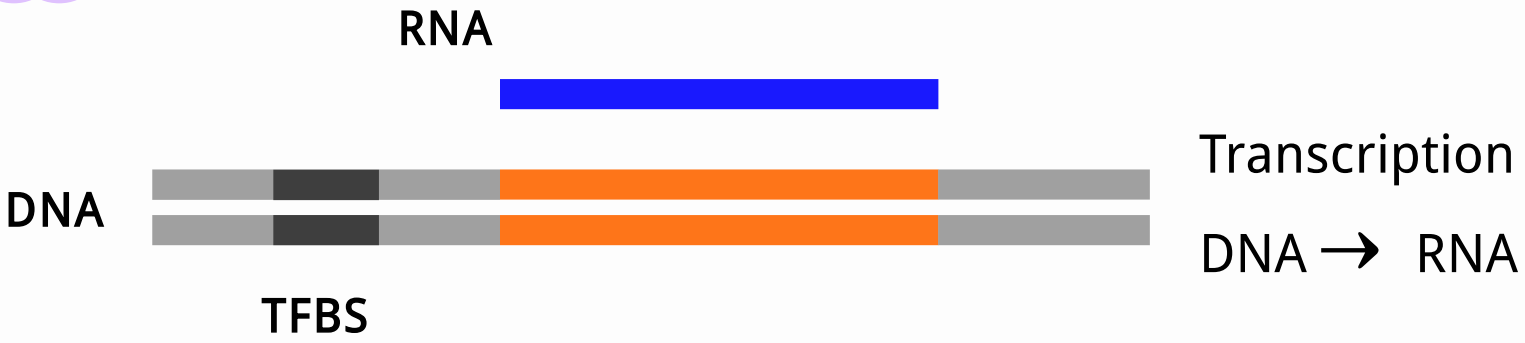
deoxyribonucleic acid

Outline

- Background
- Model and Problem Description
- Previous works and Our results
- Algorithms
- Conclusion

Biology Background

Problem: given a DNA sequence and a group of TF candidates , which one regulates the DNA transcription ?



transcription factor (TF) transcription factor binding site (TFBS)

Model of Transcription Factor

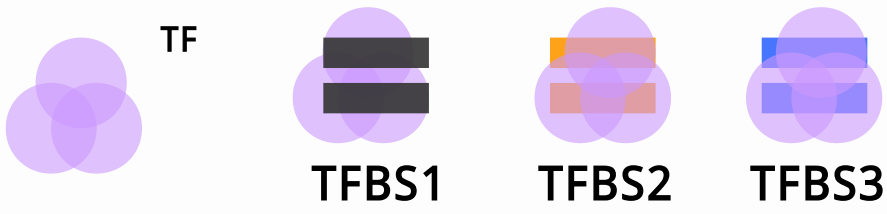
- One TF binds to a certain pattern of DNA clip (motif). We usually use binding cites to describe TF

- Word model

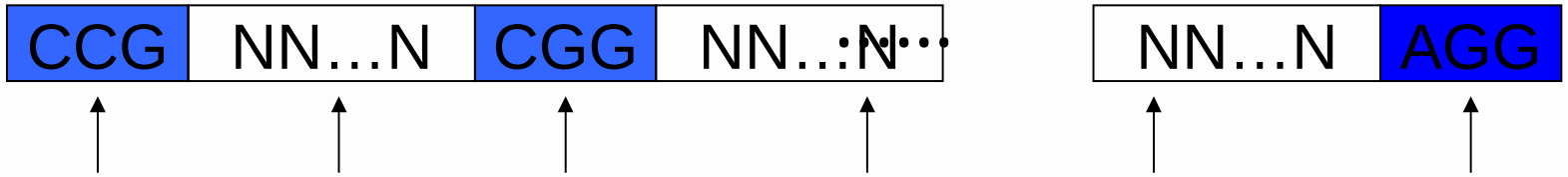
```
GACCGCTTTGTCAAC
GCTGCAGGTGTTCTC
GCAGCAGGTGTTCCC
CCCACAGCTGGGATC
```

- Matrix Model

A	0/8	2/8	3/8	3/8	0/8	7/8
C	6/8	4/8	2/8	1/8	6/8	1/8
G	3/8	2/8	0/8	4/8	1/8	0/8
T	0/8	0/8	3/8	0/8	0/8	0/8



Model of Transcription Factor(cont.)

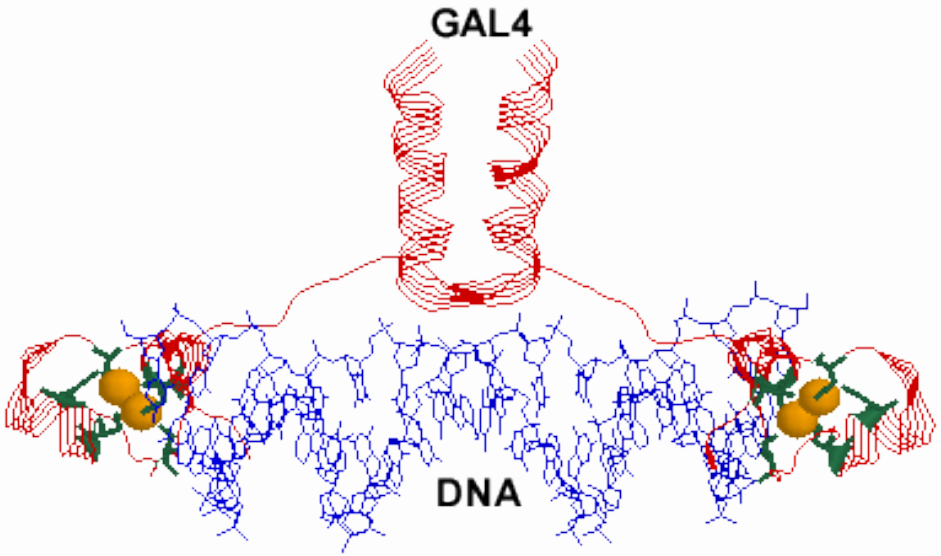
- A mixed model: structured motif introduced by Marsan and Sagot(2000):
- 

conserved spacers conserved spacers spacers conserved

- Consist of alternate conserved regions (boxes) and spacers (‘N’)

Two box structured motif

- In our paper, we consider two box structured motifs
- GAL4 is a typical two box structured motif: "CGGNNNNNNNNNNNCCG”



Models of DNA sequence

- DNA sequence

  …ATTCTAGCAAGCCTTAATTATCCAACTAAATCAGACCAGG…

- From biological view, it comes from two parts
  - background

    generated by random variables $\{X_1, X_2, \ldots, X_n\}$, which is a 1-order Markov Chain with transition matrix and stationary probability
  - binding sites

Method: Hypothesis test

- Our hypothesis is:

  the given motif *m* comes from background

  (generated by a 1-order Markov Chain R)

- Our observation is:

  m appears on DNA sequence for *k* times

- If Pr(m **appears** on R for at least k times) is very small, the hypothesis must be wrong

Problem Description (P-value calculation)

- Input: a structured motif,
  an integer $k > 0$,
  a 1-order Markov Chain of length n
  with transition matrix T and
  stationary probability u

- Output: Pr(motif appears on R for at least k times)

Previous works

- Exact algorithm:
  - The non-overlapped (Helden et. al.[1])
- Approximation algorithms:
  - Marsan et. al. [2] Robin S. et al. [3]

- [1]Van Helden, et. al.J. Rios, A.F. and Collado-Vides,J. Discovering and Regulatory elements in non-coding sequences by analysis of spaced dyads.Nucl. Acids Res. 28 1808-1818
- [2]Marsan, L., and Sagot, M.-F. 2000. Algorithms for extracting structured motifs using a suf. x tree with an applicationto promoter and regulatory site consensus identi. cation. J. Comp. Biol. 7, 345–362.
- [3] Robin, S., Daudin, J.-J., Richard, H., Sagot, M.-F. and Schbath, S. (2002). Oc-currence probability of structured motifs in random sequences. J. Comp. Biol. 9 761-773.

Our Contributions

- We give the first non-trivial algorithm to calculate the exact probability value of two-boxes structure motif
- The way to we do decomposition and dynamic programming is totally new and may be helpful to similar probability calculation.

Algorithms: main idea

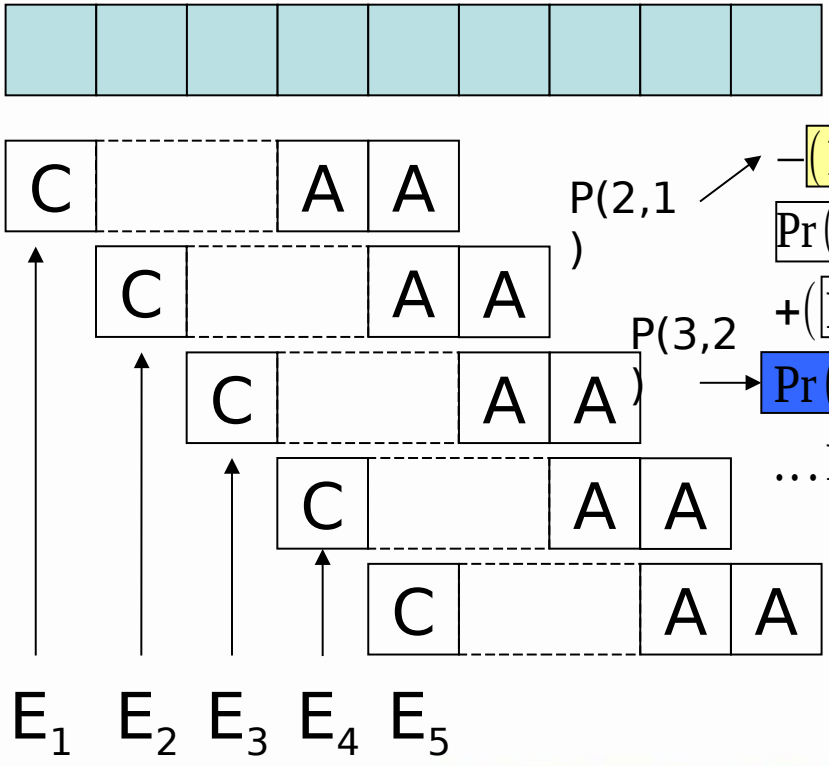
- Transformation and Decomposition to the target probability
- Do Dynamic Programming to calculate the terms in decomposition

Example

- Structured motif: CNNAA
- Hit times: at least once
- Sequence model: 1-order Markov Chain
- Sequence length: 9

Transformation

markov region R, |R|=9



$$\Pr(mhits\ R)=\Pr(E_1\cup E_2\cup E_3\cup E_4\cup E_5)$$

$$=(\Pr(E_1)+\Pr(E_2)+\ldots+\Pr(E_5))$$

The probabilities for intersections of two events with minimum index 1

P(2,1)

$$-(\Pr(E_1\cap E_2)+\Pr(E_1\cap E_3)+\ldots+\Pr(E_1\cap E_5)+$$

$$\Pr(E_2\cap E_3)+\ldots+\Pr(E_2\cap E_5)\ldots+\Pr(E_4\cap E_5))$$

P(3,2)

$$+(\Pr(E_1\cap E_2\cap E_3)+\ldots+\Pr(E_1\cap E_4\cap E_5)+$$

$$\Pr(E_2\cap E_3\cap E_4)+\ldots+\Pr(E_2\cap E_4\cap E_5)+$$

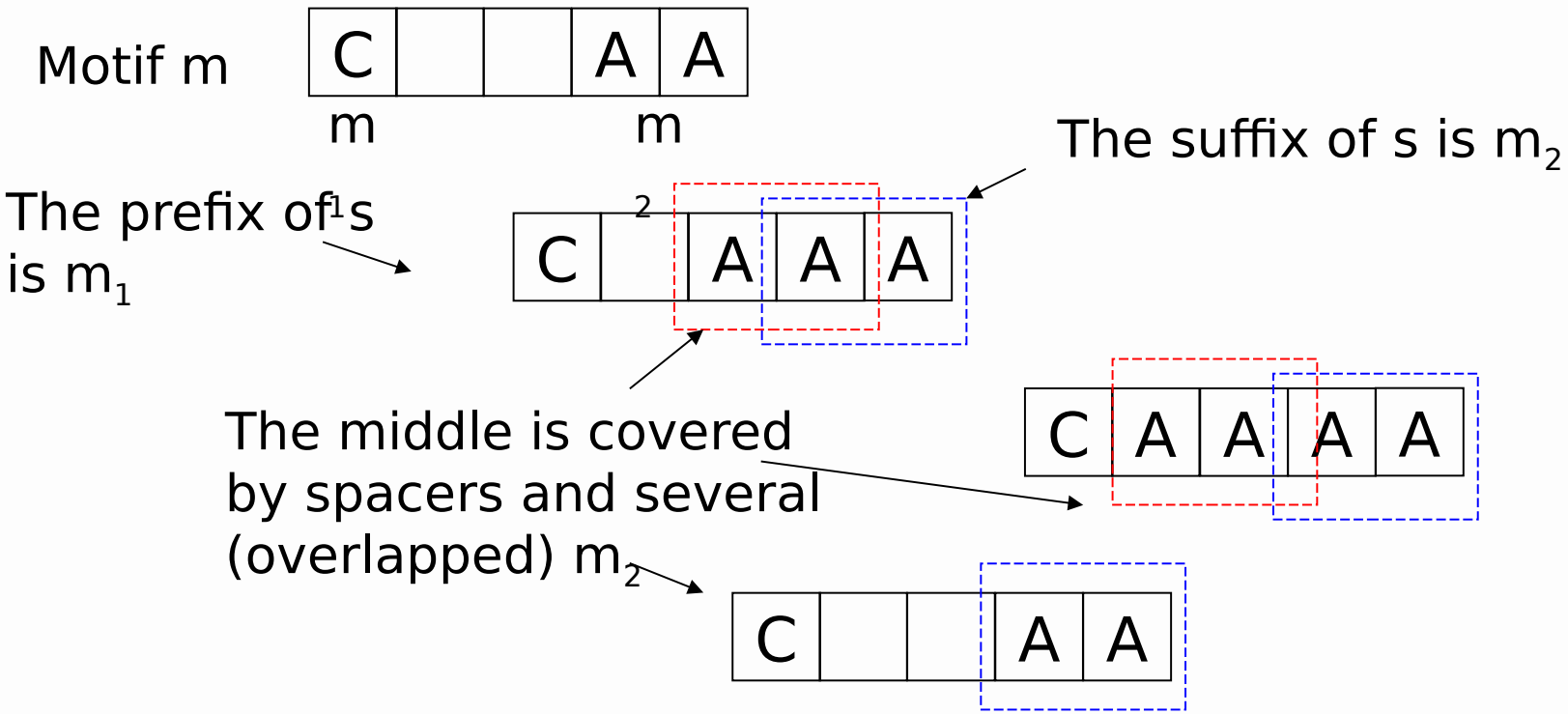
$$\ldots\Pr(E_3\cap E_4\cap E_5))$$

The probabilities for intersections of three events with minimum index 2

P(a,b) denotes the sum of all the probabilities for intersections of ***a*** events with minimum index ***b***

Terms in Dynamic Programming

- Structured prefix



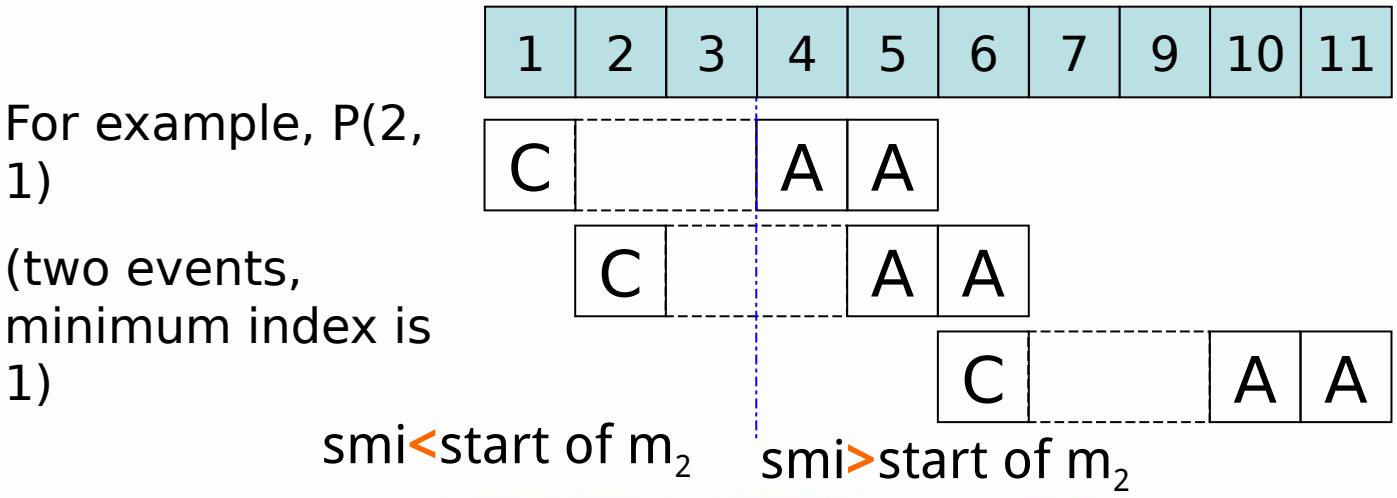
Terms in Dynamic Programming

- Recall the definition of P(a,b) and structured prefi x
  - *P(a,b)* denotes the sum of all the probabilities for inter sections of ***a*** events with minimum index ***b***
  - *structured prefix:* three constraints
- Terms in Dynamic Programming
  - P(a,b) and
  - I(a,b,z) = the sum of all the probabilities for intersectio ns of ***a*** events with minimum index ***b*** and structured p refix **z** is the prefix of subregion R[b,n]

a :from 0 to 5
b :from 5 to 1
z: arbitrary order

Dynamic Programming (conti.)

- Main idea: calculate P(a,b) and I(a,b,z) interactively
- Key idea: decompose P(a,b) and I(a,b,z) according t o the second minimum index(smi) of the events in t he sum



Recurrence Formula

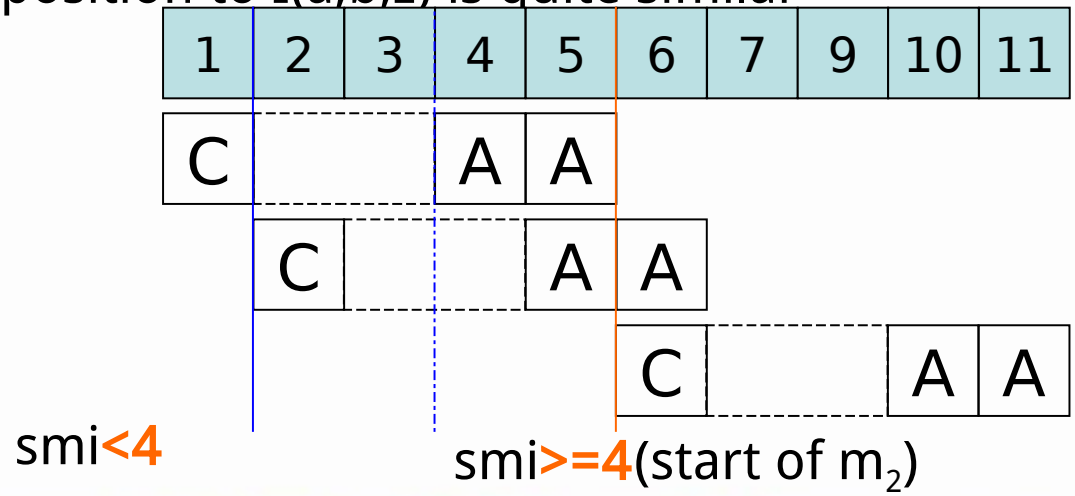
- For P(2,1)

  P(2,1) = Pr(‘CNNAA’)*P(1,6)+

  Pr(‘C’)*I(1,2,’CNAAA’)+

  Pr(‘CN’)*I(1,3,’CAAAA’)

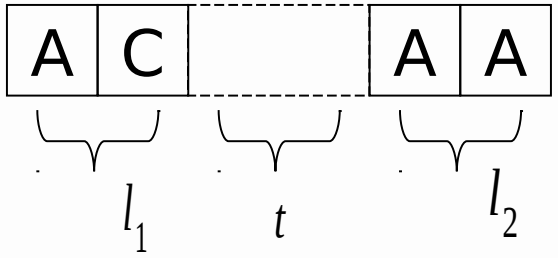
- The decomposition to I(a,b,z) is quite similar



Time complexity

- Key: estimate the size of the structured motif set
- We can prove that

$$¿\ structured\ motif = O\left(l_2^{t/l_2}\right)$$



- Total time complexity: $O\left(n^3 \times l_2^{t/l_2}\right)$
  - n is the length of the DNA sequence

Experiment Results

- In SCPD, transcription factor GAL4 is reported to bind to 7 genes. We extract upstream sequence, of length 1000 bp, for these 7 genes
- The p-value and consumed time are shown in Table 1

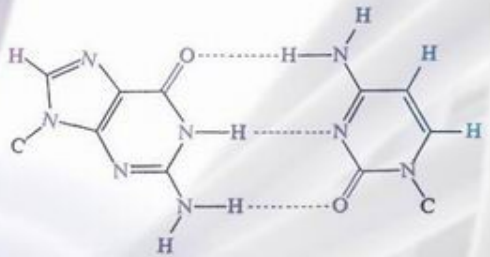
Gene Name	P-value of motif GAL4	Time
GAL1	9.61361E-06	12813ms
GAL2	2.16228E-09	12734ms
GAL4	0.006578594	12609ms
GAL7	0.034349462	12532ms
GAL10	9.61361E-06	12641ms
GAL80	0.092351709	12282ms
GCY1	0.045182862	12640ms

Table 1. The p-value of motif *GAL4* on corresponding Genes

Conclusion

- In this paper, we present a non-trivial and efficient algorithm to calculate the probability of the occurrence of a structured motif *m*
- One problem is that that is still an exponential time algorithm in worst case. Finding a polynomial time algorithm or proving that it is NP-hard are two main directions in the future work.

Thank you!



deoxyribonucleic acid