

A beautiful game

Lecture 13

①

The best bet for simpletons (※人)

The best bet for simpletons starts out with two players who select words of length l in 0-1 alphabet. Player I selects a sequence A of l alphabets from $\{0,1\}$, and play II, knowing what A is, selects another sequence B of length l . The players then flip a coin to obtain heads (1) or tails (0) in turns until either A or B appears as a block of l consecutive outcomes. If A comes first, then A wins the game. (B)

Example, $l=3$,
 $A \rightarrow 010$
 $B \rightarrow 110$ | $\langle 0,0,1,1,1,\dots \rangle$

(Fact 1) The game will terminate with probability 1,

assuming the coin is a fair one or at least both sides have positive probability. (?) 两面机率不等也可以。

(Fact 2) $\star$ B has higher probability to win the game. (How?)
($l \geq 3$)

(Fact 3) The best bet for simpletons is a non-transitive game.
 $A > B, B > C \nRightarrow A > C$.
(谁的運氣比較強?)

(2)

Given two l -letter words A and B , the correlation of A and B , to be denoted by $AB = (c_0, c_1, \dots, c_{l-1})$, is an l -letter boolean word defined as follows:

$A = 100100$		
$B = 001001$		
<u>100100</u>	← shift 0	AB
001001	← shift 1	0
001001	← shift 2	1
001001	← shift 3	0
001001	← shift 4	0
001001	← shift 5	1
001001	← shift 5	1

Shift B to the right!

the i -th bit of AB is defined to be one if the $(l-i)$ -
 $i = 0, 1, 2, 3, 4, 5$
 prefix of B coincides with the $(l-i)$ -suffix of A . Otherwise,

$AB = (0, 1, 0, 0, 1, 1) = (c_0, c_1, c_2, c_3, c_4, c_5)$

Definition The correlation polynomial of A and B is defined

as $K_{AB}(t) = c_0 + c_1 t + c_2 t^2 + \dots + c_{l-1} t^{l-1}$. We also denote

$K_{AB} = K_{AB}(\frac{1}{2})$.

Fair coin!

$K_{AB} = (1 + t^4 + t^5)(\frac{1}{2}) = \frac{1}{2} + \frac{1}{16} + \frac{1}{32} = \frac{19}{32}$.

John Conway

(3)

The odds that B will win over A is

$$\frac{K_{AA} - K_{AB}}{K_{BB} - K_{BA}}$$

Martin Gardner, 1974 : (Remark)

I have no idea why it works. It just cranks out
the answer as if by magic, like so many of Conway's
other algorithms. (1993 找到最简单的证明.)

A = 00

B = 10

如果 A 先选 00,
则 B 选 10 的胜率
会大很多!

$$K_{AB}(x) = (0, 0) \sim 0 + 0x$$

$$K_{AA}(x) = (1, 1) \sim 1 + x$$

$$K_{BB}(x) = (1, 0) \sim 1$$

$$K_{BA}(x) = (0, 1) \sim 0 + x$$

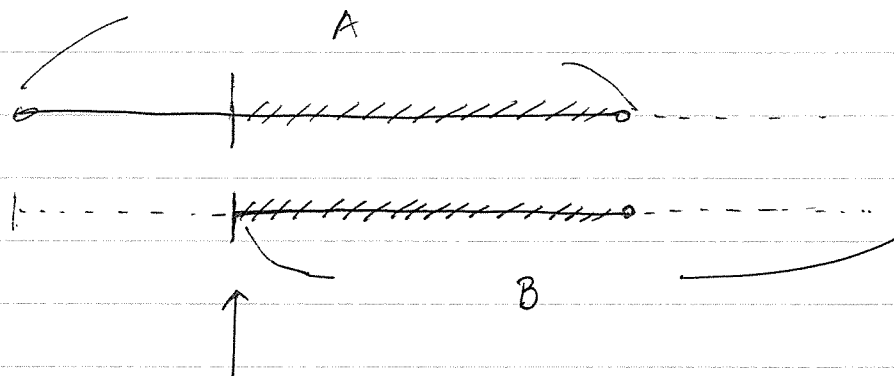
$$\frac{K_{AA} - K_{AB}}{K_{BB} - K_{BA}} = \frac{1 + \frac{1}{2}}{1 - \frac{1}{2}} = \frac{\frac{3}{2}}{\frac{1}{2}} = \frac{3}{1}$$

直觉: 除了 00 A 赢之外
(10, 01, 11) 都是 B 赢。

Proof. 1980, Li || 1981, Guibas and Odlyzko,

1993, Pevzner

shortest one!



(•) 在切断之後，後半段 (A) 与 B 的前段 可能会是一样，
(基本上)

因此，我们探讨所谓的 Self-correlation or

auto-correlation: 自身循環移动後相似的情况。
(Optical codes)

(•) 如果 Letters 都一样，则無論如何移动都具有自身相似的情况。

(*) 所以我们对论的对象是 Letters 出现的机率给定的情况下所產生的 Words，例如 0, 1 出现的机率各 $\frac{1}{2}$ ；在不断地生成 (0, 1)-字串时，那种类型 (pattern) 的 Words 会比较容易出现，还是都一样！

Example

00, 01

010, 101

A

B

) 谁的出現机率大？

Date

1

前三次

A

B

No.

(3)

000

001

010

011

100

101

110

111

V

V

A 先赢

B 先赢

B —

A —

B —

B —

整体而言, B 赢的几率较大。

The Conway Equation

Review of Correlation Polynomial

Finding Signals in
DNA (continued)

A', B' : l -letter words

Correlation of A', B' : $AB' = (c_0, c_1, \dots, c_{l-1})$ where the i th bit of AB' is defined to be one "1" if the $(n-i)$ -prefix (the first $n-i$ letters) of B' coincides with the $(n-i)$ -suffix of A' , and "0" otherwise.

e.g.
+.

$$A = \underline{100100}$$

$$A = A'$$

$$B = \underline{001001}$$

$$B = B'$$

$$AB = (0, 1, 0, 0, 1, 1)$$

$$AA = (1, 0, 0, 1, 0, 0)$$

$$A = A' = B'$$

$$BB = (1, 0, 0, 1, 0, 0)$$

$$B = B' = A'$$

以下用 A, B 代表
from

$$BA = (0, 0, 1, 0, 0, 1)$$

$$A' = B, B' = A$$

(*) Let $AB = (c_0, c_1, \dots, c_{l-1})$ and $c_{m_1}, c_{m_2}, \dots, c_{m_k}$ be the bits of AB equal to 1.

In AB , $c_1 = 1, c_4 = 1, c_5 = 1$ and thus $m_1 = 1, m_2 = 4, m_3 = 5$.

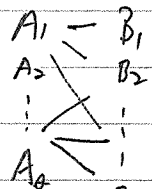
⑤

(*) Denote as $\mathcal{H}_{AB}$ the set of k prefixes of $A = a_1 a_2 \dots a_\ell$ of length $m_1, m_2, \dots, m_k$:

$$(a_1, \dots, a_{m_1}), (a_1, a_2, \dots, a_{m_1}, \dots, a_{m_2}), \dots, (a_1, a_2, \dots, a_{m_{k-1}}, \dots, a_{m_k}),$$

e.g. If $A = XYXYXY$, then

⑤' $\mathcal{H}_{AB} = \{X, XYXY, XYXYXY\}$



Definition Let $A = a_1 a_2 \dots a_\ell$ and $B = b_1 b_2 \dots b_\ell$, then

the concatenation of A and B , $A * B = a_1 a_2 \dots a_\ell b_1 b_2 \dots b_\ell$. If

$\mathcal{A} = \{A\}$ and $\mathcal{B} = \{B\}$, then $\mathcal{A} * \mathcal{B} = \{A * B\}$ which has possibly $|\mathcal{A}||\mathcal{B}|$ words (perhaps with repeats).

Definition If W is an ℓ -letter word, then let $P(W) = \frac{1}{2^\ell}$. For

a set $\mathcal{W} = \{W\}$, let $P(\mathcal{W}) = \sum_{W \in \mathcal{W}} P(W)$.

e.g. $P(\mathcal{H}_{AB}) = \frac{1}{2} + \frac{1}{2^4} + \frac{1}{2^5}$.

Lemma $K_{AB}(\frac{1}{2}) = P(\mathcal{H}_{AB})$.

Proof. (Explain with an example)

In the above example $K_{AB}(t) = t + t^4 + t^5 = \frac{1}{2} + (\frac{1}{2})^4 + (\frac{1}{2})^5$.

(*) 在 $\mathcal{H}_{AB}$ 中 长度分别为 1, 4, 5 的 words.

$$A = XYYXY$$

$$XYYXY$$

$$XYYXY$$

$$XYYXY$$

$$XYYXY$$

$$XYYXY$$

AA

1

0

0

1

0

0

$$H_{AA} = \{XYY\}$$

$$\text{poly. } 1+t^3$$

$$B = XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

BB

1

0

0

1

0

0

$$H_{BB} = \{XYXYX\}$$

$$1+t^3$$

BA

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$H_{BA} = \{XY, XYXYXY\}$$

$$t^2 + t^5$$

A

$$XYXYXYX$$

B

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

$$XYXYXYX$$

0

1

0

0

1

1

$$H_{AB} = \{X, XYXYX, XYXYXY\}$$

$$t + t^4 + t^5$$

$$\frac{K_{AB} - K_{BA}}{K_{BB} - K_{BA}} = \frac{\frac{9}{8} - \frac{19}{32}}{\frac{9}{8} - \frac{9}{32}}$$

$$= \frac{17}{27}$$

(6)

A word W is an A -victory if it contains A in the end and does not contain B .
(B) (B)
(A)

A word W is an A -previctory if $W * A$ is an A -victory.
(B) (B) (B)

(*) $W * A$ 不一定是 A 赢 (有可能 A 没接完, B 已经赢了).
(一般而言)

Let $S_A = \{A\text{-previctories}\}$, $S_B = \{B\text{-previctories}\}$ and

$\mathcal{T} = \{T : T \text{ is neither } A\text{-victory nor } B\text{-victory}\}.$

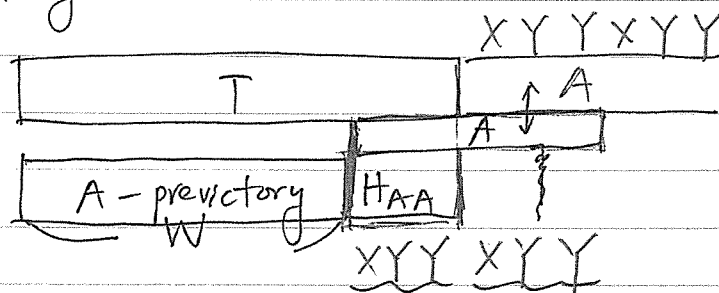
↑ the set of no-victor words.

$\forall T \in \mathcal{T},$

Fact 1 $T * A$ corresponds to either an A -victory or a B -victory
(接完当然是 A 赢)

Fact 2 If $T * A$ corresponds to an A -victory, then

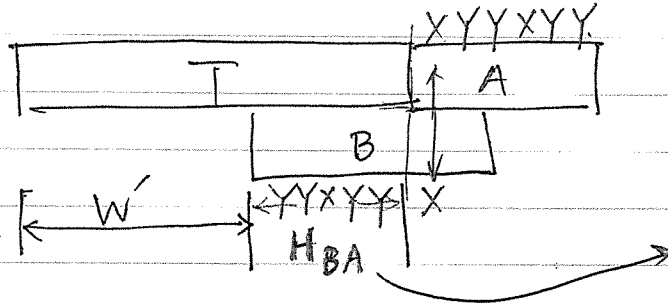
T can be represented as $W * HAA$ where W is an A -previctory.



⑦

Fact 3 If $T * A$ corresponds to a B-victory, then

$T = W' * H_{BA}$ where W' is a B-previctory.



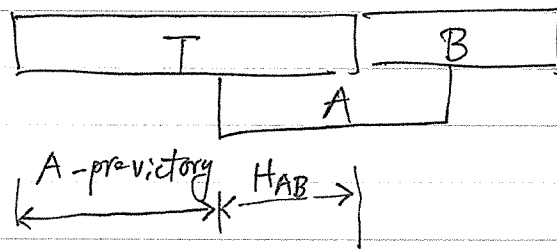
B的前:段与A的后:段相同

Fact 4 (Combine Fact 2 and Fact 3.)

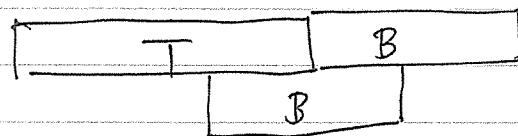
$$\mathcal{G} = \mathcal{G}_1 = (S_A * H_{AA}) \cup (S_B * H_{BA})$$

Fact 5 Similar to Fact 4,

$$\mathcal{G} = \mathcal{G}_2 = (S_A * H_{AB}) \cup (S_B * H_{BB})$$



$T * B \rightarrow A\text{-victory}$



$T * B \rightarrow B\text{-victory}$

(8)

Theorem The odds that B wins over A is

$$\frac{K_{AA} - K_{AB}}{K_{BB} - K_{BA}} \quad (K_{AB} = K_{AB}(\frac{1}{2}))$$

Proof.

$$P(\mathcal{G}_2) = P(S_A * \mathcal{H}_{AB}) + P(S_B * \mathcal{H}_{BB})$$

$$= P(S_A)P(\mathcal{H}_{AB}) + P(S_B) \cdot P(\mathcal{H}_{BB})$$

$$= P(S_A)K_{AB} + P(S_B) \cdot K_{BB}$$

$$P(\mathcal{G}_1) = P(S_B)K_{BA} + P(S_A)K_{AA}$$

Since $P(\mathcal{G}_1) = P(\mathcal{G}_2)$,

$$P(S_A)K_{AB} + P(S_B)K_{BB} = P(S_B)K_{BA} + P(S_A)K_{AA}$$

$$\frac{P(S_B)}{P(S_A)} = \frac{K_{AA} - K_{AB}}{K_{BB} - K_{BA}}$$

(Note that the odds that B wins over A is equal to the ratio of $P(S_B)$ over $P(S_A)$, i.e., the probability of B-previctories over the probability of A-previctories.)

(~~Handwritten~~) Signals are everywhere!



A 0101

$$(t = \frac{1}{2})$$

B 選什麼?

Sol. Find a B such that $\frac{K_{AA} - K_{AB}}{K_{BB} - K_{BA}} \geq 1$. (" > 1 " is preferable.)

 K_{AA}

0101

0101

0101

0101

0101

1

0

1

0

$$\text{poly.} \approx 1 + t^2$$

$$K_{AA} = \frac{5}{4}$$

(*) Find B with smaller K_{BB} . (策略: $K_{AA} \geq K_{BB}$
 $K_{AB} \geq K_{BA}$)

$$B = 1101$$

1101

1101

1101

1101

1101

1

0

0

1

$$\text{poly.} \approx 1 + t^3$$

$$K_{BB} = \frac{9}{8}$$

0101

1101

1101

1101

1101

0

0

0

0

1

$$t^3$$

 K_{AB}

$$K_{AB} = \frac{1}{8}$$

1101

0101

0101

0101

0101

0

0

0

1

0

$$t^2$$

$$K_{BA} = \frac{1}{4}$$

$$\text{Odds} = \frac{\frac{5}{4} - \frac{1}{8}}{\frac{9}{8} - \frac{1}{4}} = \frac{\frac{9}{8}}{\frac{7}{8}} = \frac{9}{7}$$

Problem: Can we use the patterns (codewords) to determine the odds (B over A)?

What if $A = 1101$? $K_{AA} = \frac{9}{8}$

(*) Smaller K_{BB} . $B = 0001$

$$K_{BB} = 1$$

$$\begin{array}{r} 0001 \\ 0001 \\ 0001 \\ 0001 \\ 0001 \end{array} \quad \begin{array}{r} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \quad \begin{array}{r} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array}$$

$$\begin{array}{r} K_{AB} \\ 1101 \\ 0001 \\ 0 \end{array}$$

$$\text{Odds (B over A)} = \frac{\frac{9}{8}}{\frac{7}{8}} = \frac{9}{7}$$

$$\begin{array}{r} K_{BA} \\ 0001 \\ 1101 \\ 1 \\ 8 \end{array}$$

A

C

$$\text{rank}(0101) < \text{rank}(1101) < \text{rank}(0001)$$

$$K_{AA} = \frac{5}{4}, K_{CC} = 1$$

$$\begin{array}{r} 0101 \\ 0001 \end{array} \quad K_{AC} = 0, K_{CA} = \frac{1}{4}$$

$$\begin{array}{r} 0001 \\ 0101 \end{array} \quad \frac{1}{4}$$

$$\frac{\frac{5}{4} - 0}{1 - \frac{1}{4}} = \frac{5}{3} \quad (\text{even larger!})$$

So, can you find an example such that the transitivity of oeds does not hold?

Problem

$$\in \mathbb{Z}_2^n$$

1. For which A and B , $\frac{K_{AA} - K_{AB}}{K_{BB} - K_{BA}} =_{\text{def}} K(A, B)$ is maximum?

2. For which $x \in \mathbb{Z}_2^n$, $K_{xx} = 1$?

Gene Prediction (Comments)

Charles Yanofsky, + Sydney Brenner et al. (1960's)

Fact 2

Showed that a gene and its protein product are colinear structures with direct correlation between triplets of nucleotides in the gene and amino acids in the protein.

Later (1960's)

Overlapping genes and genes-within-genes were discovered.

(*) 如此, 註定要預測基因是一件很困難的工作。

尤其是在1977, Split human genes 的發現, 終於造成
"computational gene prediction" puzzle."

1977

(Phillip Sharp and independently Richard Roberts)

Fact 1

① Most human genes are interrupted by junk DNA

and are broken into pieces call exons.