A beautiful game

Lecture 12 (Gene prediction) ①

The best bet for simpletons (笨人)

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 $\{0,1\}^l$, 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 (B) comes first, then A (B) wins the game.

Example, $l=3$,
$A \to 010$
$B \to 110$
$\Big|\ \langle 0,0,1,1,1,\cdots\rangle$

(Fact 1) The game will terminate with probability 1, assuming the coin is a fair one or at least both sides have positive probability. (?) 两面概率不等也可以。

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

(Fact 3) The best bet for simpletons is a non-transitive game. (难的通常比较吗?)

$A>B, B>C \not\Rightarrow A>C$.

好，可以

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

$A = 100100$
$B = 001001$

Shift B to the right!

	AB
shift 0	0
shift 1	1
shift 2	0
shift 3	0
shift 4	1
shift 5	1

"1"

the i-th bit of AB is defined to be "1" if the $(\ell-i)$-prefix of B coincides with the $(\ell-i)$-suffix of A. Otherwise, "0".

$i = 0, 1, 2, 3, 4, 5$

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

Definition The correlation polynomial of A and B is defined as $K_{AB}(t) = c_0 + c_1 t + c_2 t^2 + \cdots + c_{\ell-1} t^{\ell-1}$. We also denote $K_{AB} = K_{AB}(\frac{1}{2})$.

Fair coin!

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

(3)

John Conway

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 我到最簡潔的說明.)

A = 00

B = 10

如果A先選00，則B選10的勝算大很多！

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

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

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

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

A = 100

B = 010

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

直觀：除了00A贏外 (10)01,11都是B贏.

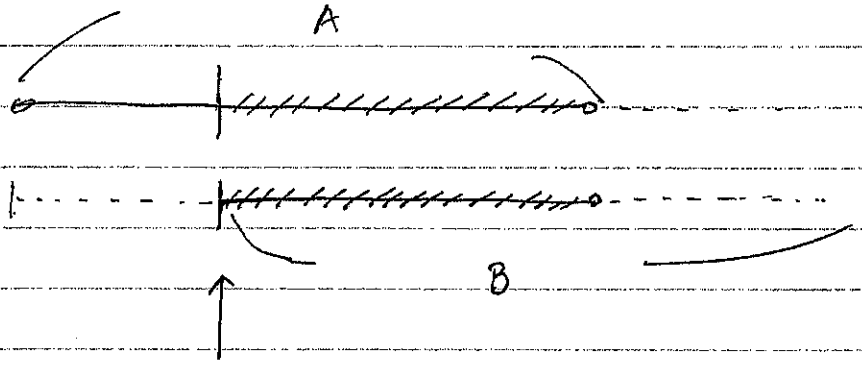
Who wins ?

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

1993, Pevzner shortest one !

Date / /

No. (3)'



(•) 在切断之後，後半段 (A) 与 B 的前段可能(基本上)会是一样，

因此，我们探讨所谓的 Self-correlation or auto-correlation：自身循环移动後相似的情況。(Optical codes)

(•) 如果 Letters 都一样，则无论如何移动都具有自身相似的情況。

(*) 所以我们讨论的对象是 Letters 出现的机率给定的情況下所产生的 Words，以如 0, 1 出现的机率各 $\frac{1}{2}$，在不断地生成 (0,1)-字串時，那种类型 (pattern) 的 Words 会比较容易出现，还是都一样！

Example 00, 01) 谁的出现机率大？

010, 101

A B

Date / No. (3)''

前三次	A	B	
000			A 先贏
001			B 先贏
010	V		
011			B —
100			A —
101		V	
110			B —
111			B —

整体而言，B 贏的机率較大。

A : 010

B : 101

If Conway is correct, then what is the "策略" to win? (B over A!)

$$\Rightarrow \frac{K_{AB} - K_{AA}}{K_{BB} - K_{BA}} \geq 1$$

(*) Either $K_{AA} \gg K_{BB}$ or

$K_{AB} \ll K_{BA}$!

⇒ Page (9)

The Conway Equation

Finding Signals in DNA (continued)

Review of Correlation Polynomial

A', B' : ℓ-letterwords

Correlation of A', B' : $A'B' = (c_0, c_1, \cdots c_{\ell-1})$ where the ith bit of $A'B'$ 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. $A = 100100$ $\quad A = A'$

$B = 001001$ $\quad B = B'$

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

$AA = (1,0,0,1,0,0)$ $\quad A = A' = B'$

$BB = (1,0,0,1,0,0)$ $\quad B = B' = A'$

以下用 A, B 代表

$BA = (0,0,1,0,0,1)$ $\quad A' = B,\ B' = A$

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

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

($Supp(AB) = \{1, 4, 5\}$.)

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

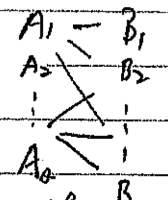
$(a_1, \cdots, a_{m_1}), (a_1, a_2, \cdots, a_{m_1}, \cdots, a_{m_2}), \cdots, (a_1, a_2, \cdots, a_{m_{k-1}}, \cdots, a_{m_k})$.

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

$B = YYXYYX$ (5)

$\mathcal{A}_{AB} = \{X, XYYX, XYYXY\}$

1 4 5



<u>Definition</u> Let $A = a_1 a_2 \cdots a_\ell$ and $B = b_1 b_2 \cdots b_\ell$, then the concatenation of A and B, $A * B = a_1 a_2 \cdots a_\ell b_1 b_2 \cdots b_\ell$. If $\mathbb{A} = \{A\}$ and $\mathbb{B} = \{B\}$, then $\mathbb{A} * \mathbb{B} = \{A * B\}$ which has possibly $|\mathbb{A}||\mathbb{B}|$ words (perhaps with repeats).

<u>Definition</u> 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. $P(\mathcal{A}_{AB}) = \frac{1}{2} + \frac{1}{2^4} + \frac{1}{2^5}$.

<u>Lemma</u> $K_{AB}(\frac{1}{2}) = P(\mathcal{A}_{AB})$.

<u>Proof</u> (Explain with an example)

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

(*) 在 $\mathcal{A}_{AB}$ 中有長度分別為 1, 4, 5 的 words.

AA

$A = XYYXYY$

$XYYXYY$ 1

$XYYXYY$ 0

$XYYXYY$ 0

$XYYXYY$ 1

$XYYXYY$ 0

$XYYXYY$ 0

$\mathcal{H}_{AA} = \{\emptyset, X, YY\}$

poly. $1+t^3$

BB

$B = YYXYYX$

$YYXYYX$ 1

$YYXYYX$ 0

$YYXYYX$ 0

$YYXYYX$ 1

$YYXYYX$ 0

$YYXYYX$ 0

$\mathcal{H}_{BB} = \{\emptyset, Y, YX\}$

$1+t^3$

BA

$YYXYYX$

$XYYXYY$ 0

$XYYXYY$ 0

$XYYXYY$ 1

$XYYXYY$ 0

$XYYXYY$ 0

$XYYXYY$ 1

$\mathcal{H}_{BA} = \{\emptyset, YY, YYXYY\}$

t^2+t^5

A

B

$XYYXYY$ 0

$YYXYYX$ 1

$YYXYYX$ 0

$YYXYYX$ 0

$YYXYYX$ 1

$YYXYYX$ 1

$\mathcal{H}_{AB} = \{X, XYYX, XYYXY\}$

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

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

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

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贏有可能A沒接完，B已經贏了。
(一般而言)

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

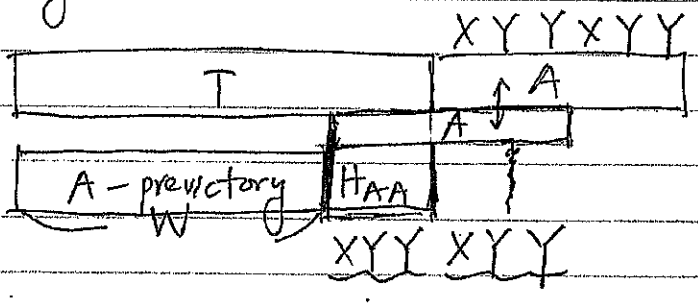
$\mathcal{T} = \{T : T \text{ is neither A-victory nor B-victory}\}$.

↑ the set of no-victor words.

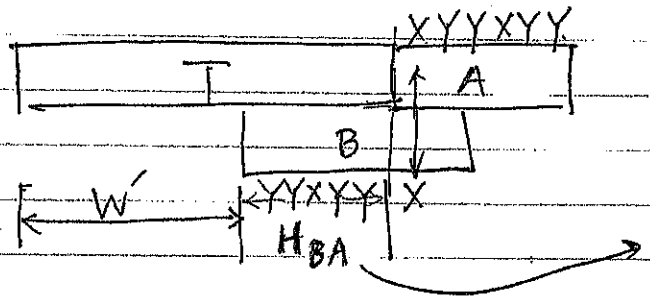
$\forall\, T \in \mathcal{T}$,

Fact 1 $T*A$ corresponds to either an A-victory or a B-victory
(接完但也是A贏)

Fact 2 If $T*A$ corresponds to an A-victory, then T can be represented as $W*H_{AA}$ where W is an A-previctory.



Fact 3 If $T*A$ corresponds to a B-victory, then $T = W' * H_{BA}$ where W' is a B-previctory.

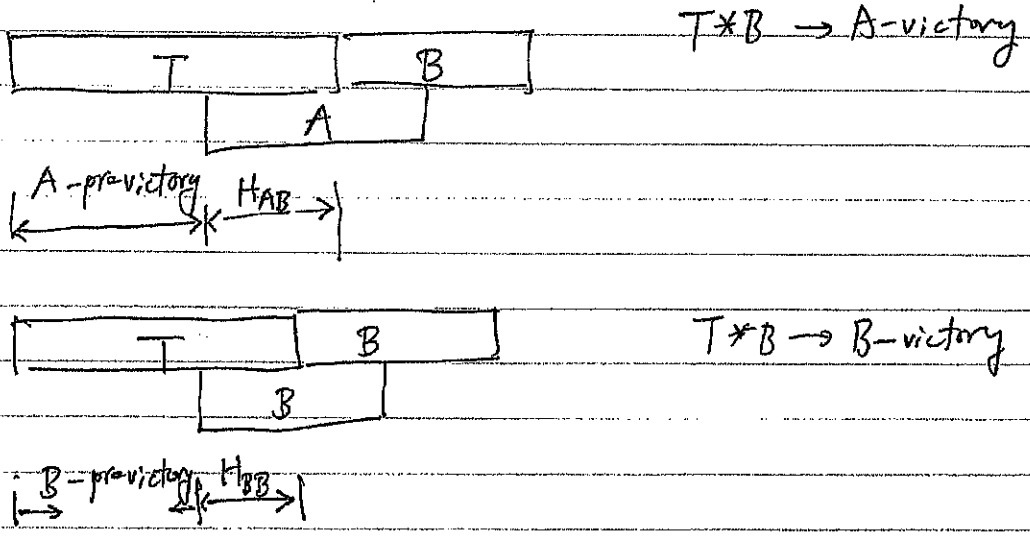


Fact 4 (Combine Fact 2 and Fact 3.)

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

Fact 5 Similar to Fact 4,

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



Theorem The odds that B wins over A is

$$\frac{K_{AA} - K_{AB}}{K_{BB} - K_{BA}}. \qquad (K_{AB} \doteq K_{AB}(\tfrac{1}{2})).$$

Proof.

$$p(\mathcal{T}_2) = p(S_A * \mathcal{A}_{AB}) + p(S_B * \mathcal{A}_{BB})$$

$$= p(S_A)\,p(\mathcal{A}_{AB}) + p(S_B) \cdot p(\mathcal{A}_{BB})$$

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

$$p(\mathcal{T}_1) = p(S_B)\,K_{BA} + p(S_A)\,K_{AA}$$

Since $p(\mathcal{T}_1) = p(\mathcal{T}_2)$,

$$p(S_A)\,K_{AB} + p(S_B)\,K_{BB} = p(S_B)\,K_{BA} + p(S_A)\,K_{AA}.$$

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

(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.)

Signals are everywhere!

Date / /

No. 9

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 1 poly. $\approx 1+t^2$

0101 0

0101 1 $K_{AA}=\frac{5}{4}$

0101 0

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

B = 1101

1101

1101 1 poly. $\approx 1+t^3$

1101 0

1101 0 $K_{BB}=\frac{9}{8}$

1101 1

0101 1101 0

1101 0 0101 0

1101 0 t^3 0101 0 t^2

1101 1 0101 1

K_{AB} 0101 0

$K_{AB}=\frac{1}{8}$ $K_{BA}=\frac{1}{4}$

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

Date / /

No. 10

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

$K_{BB} = 1$

0001
0001
0001
0001
0001

1
0
0
0
0

1

K_{AB} = 0 (1101 vs 0001)

K_{BA} = $\frac{1}{8}$ (0001 vs 1101)

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

A C

rank(0101) < rank(1101) < rank(0001)

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

0101 vs 0001: $K_{AC} = 0$, $K_{CA} = \frac{1}{4}$ (0001 vs 0101) t

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

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

Problems

1. For which A and B, $\frac{K_{AA}-K_{AB}}{K_{BB}-K_{BA}} =_{def} K(A,B)$ is maximum? $\in \mathbb{Z}_2^n$

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

Gene Prediction (Comments)

Charles Yanofsky & Sydney Brenner et. al (1960年p)

Fact2

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年後)

Overlapping genes and genes-within-genes were discovered.

(*) 如此，註定要預測基因是一件極困難的工作。

直到在1977, Split human genes的發現，終於造成

"computational gene prediction" puzzle".

1977 (phillip Sharp and independently Richard Roberts)

Fact1

① Most human genes are interrupted by junk DNA and are broken into pieces call exons.