Lecture 9

12,11 ~ 12,12

Finding Signals in DNA

- The first signal in DNA was found in 1990 by Hamilton Smith after the discovery of the Hind II restriction enzyme.
- The palindromic site of the restriction enzyme is a signal that initiates DNA cutting.

(Fact) 1. Hamilton Smith was lucky: restriction sites are the simplest signals in DNA. (Reliably find in DNA.)

2. Most other signals (promoters, splicing sites, etc.) are so complicated that we don't yet have good models or reliable algorithms for their recognition.

(**) Understanding gene regulation is a major challenge in computational biology. For example, regulation of gene expression may involve a protein binding to a region of DNA to affect transcription of an adjacent gene.

(Fact) Since protein–DNA binding mechanisms are still insufficiently understood to allow "in silico" prediction of binding sites, the common experimental approach is to locate the approximate position of the binding site.

(***) These experiments usually lead to "identification" of a DNA fragment of length n that contains a binding site (an unknown magic word) of length $\ell \ll n$.

•• Of course, one experiment is insufficient for finding the binding site, but a "sample" of experimentally found DNA fragments gives one hope of recovering the magic word.

GATTCTTAGGC
TATACGTTCGA
TGATTGACTTC
⋮

Problem

Problem (3)

Given a sample of K sequences where an (unknown) magic word appears at different (unknown) positions in these K sequences. Find the magic word!

Common Sense Approach (Staden, 1989; Wolfertstetter et al, 1996; Tompa, 1999)

Since $l << n$, we may test all words of length l and find those that appear in all (or almost all) K sequences.

Note l 可能也不知道有多長；所以可從較小的長度測試再增長；一直到有"唯一"未現的 Magic word 為止。

① The described approach usually works fine for short continuous words such as GAATTC the restriction site of EcoRI.

② In the idea of randomness, this is also quite difficult, since we have 4^6 words to compete for the magic word.

③ The problem gets even more difficult (complicated) when the magic word has gaps.

For example,

$CCAN_qTGG$ (Xcm I restriction enzyme)

↑ a gap of length q

$(Pu)^m CN_{40-2000}(Pu)^m C$ (McrBC Endonuclease)

↑ A or G

$TTGACAN_{17}TATAAT$ (E. Coli promoters)

Enumeration and check of all patterns of the above types are hardly possibly due to computational complexity.

Remark

寻找 Magic Words 基本上就像密码学中寻找 Key 的概念，除了利用各种可能提供的讯息之外，还要发挥一些想像力来协助判断；把所有可能的答案都试一次绝非良策。

- DNA linguistics is at the heart of the "pattern-driven" approach to signal finding, which is based on enumerating all possible patterns and choosing the most frequent or the fittest among them.

(*) The fitness measures vary from estimates of the statistical significance of discovered signals to the information content of the fragments that approximately match the signal.

Steps for pattern-driven approach

Step 1. Define the frequency or fitness measure. (f.f.m.)

Step 2. Calculate the f.f.m. of each word w.r.t. the sample DNA fragments.

Step 3. Report the fittest words as potential signals.

(Fact)

Note that if A denotes the set of alphabets, then the search space for patterns of length l is $|A|^l$.

(•) DNA texts are not easy to decipher, and there is little doubt that Nature can construct an "enigma" (謎) of the kind which human ingenuity (創造力) may not resolve.

- A popular approach in DNA linguistics is based on the assumption that frequent or rare words may correspond to signals in DNA.

- A word occurs considerably more (or less) frequently than expected has the potential to become a "signal". → What is its biological meaning?

For example, Gelfand and Koonin (1997) showed that the most avoided 6-palindrome in the archaeon M. jannaschii is likely to be the recognition site of a restriction-modification system.

- To find frequent and rare words (W) in a text, one has to compute expected value $E(W)$ and the variance $\sigma^2(W)$ for the number of occurrences (frequency) of each word W.
- Afterwards, the frequent and rare words are identified as the words with significant deviations from expected frequencies.

Remark

In many DNA linguistics papers, the variance $\sigma^2(W)$ of the number of occurrences of a word in a text was erroneously assumed to be $E(W)$.

要探討 word's occurrence 的和率就不是一件直截(或簡單)的事。參考 "Correlation Polynomial" (Autocorelation)

The best bet for simpletons

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. (A) (B)

(Fact 1) The game will terminate with probability 1, assuming the coin is a fair one or at least both sides have positive probability.

(Fact 2) B has higher probability to win the game.

(Fact 3) The best bet for simpletons is a non-transitive game.

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:

$A = \not{0}\underline{XX\not{0}XX}$
$B = XX\not{0}XX\not{0}$
$\quad\ \underline{XX\not{Y}XX}Y$
$\quad\ 110110$
$\quad\ 110110$
$\quad\ 110110$
$\quad\ 110110$

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

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

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

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

John Conway

The odds that B will win over A is

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

Martin Gardner, 1974:

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.

$A = 00$

$B = 10$

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

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

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

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

$$\frac{1 + \frac{1}{2}}{1 - \frac{1}{2}} = \frac{\frac{3}{2}}{\frac{1}{2}} = \frac{3}{1}$$

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

1993, Pevzner shortest one!