

Gene Prediction

- In the 1960s, Charles Yanofsky and their collaborators showed that a gene and its product are colinear with direct correlation between triplets of nucleotides in the gene and amino acids in the protein.
- However, the concept of the gene as a synthetic string of nucleotides did not live long.
- Overlapping genes and genes-within-genes were discovered in the late 1960s.
- These studies demonstrated that the computational problem of gene prediction is far from simple.

Approaches

There are ways to discover genes:

(1). Statistical Approach

(2). Similarity-Based Approach

(3). Spliced Alignment

(4). Assembling Idea

(5). Hidden Markov Models

(*) See the reference book: **Computational Molecular Biology by Pavel A Pevzner.**

Possible Genes

- By the experience of finding more genes over past years, we may have idea in predicting which combination of nucleotides are genes by way of “**smart**” **guess**.
- So, the problem is to predict genes through a **beautiful mind**.

A beautiful game: The best bet for simpletons

The best bet for simpletons starts out with two players who select words of length k in 0-1 alphabets. Player 1 selects a sequence A of k alphabets from $\{0, 1\}^k$, and Player 2, after knowing what A is, selects another sequence of length k also from $\{0, 1\}^k$. The players then flip a coin to obtain heads (1) or tails (0) in terms until A or B appears as a block of k consecutive outcomes. If A (B) comes first then A (B) wins the game.

Comments and Examples

- For example, $A = 100100$ and $B = 001001$.

Do you believe that they have different probability of winning the game?

- Comments

1. The game will terminate with probability 1 as long as the length of sequences chosen by A and B are finite, assuming both sides of the coin have positive probabilities. (?) In general, we assume they have the same probability $\frac{1}{2}$.
2. B has higher probability to win the game. (How?)

An Observation

- Let $A = 010$ and $B = 101$.

Now, we flip the coin 3 times assuming each side has probability $\frac{1}{2}$.

Then we have 8 outcomes:

000 (A^+)

001 (B^+)

010 (A)

011 (B^+)

100 (A^+)

101 (B)

110 (B^+)

111

A^+ : A has better chance to win.

Correlation of A and B

- In order to find which sequence (A and B) has better chance to win, it is essential to find the relationship between A and B.

Definition (Correlation)

Given two k -letter words A and B , the correlation of A and B (ordered), denoted by $AB = (c_0, c_1, \dots, c_{k-1})$, is a k -letter Boolean word $((0, 1)$ -vector) defined as follows: the i -th component of AB , c_{i-1} , is equal to 1 if the $(k-i)$ -prefix of B coincides with the $(k-i)$ -suffix of A , and 0 otherwise.

For example, $A = 100100$ and $B = 001001$. Then, $AB = (0, 1, 0, 0, 1, 1)$, $BA = (0, 0, 1, 0, 0, 1)$, $AA = (1, 0, 0, 1, 0, 0)$ and $BB = (1, 0, 0, 1, 0, 0)$.

Correlation Polynomial of A and B

Definition (Correlation polynomial)

The correlation polynomial of A and B is defined as $K_{AB}(t) = c_0 + c_1 t + c_2 t^2 + \dots + c_{k-1} t^{k-1}$ provided $AB = (c_0, c_1, \dots, c_{k-1})$. For convenience, we let $K_{AB}(1/2) = K_{AB}$.

(*) The $\frac{1}{2}$ comes from flipping a fair coin.

Therefore, if $AB = (0, 1, 0, 0, 1, 1)$, then $K_{AB} = 1/2 + 1/16 + 1/32 = 19/32$.

We can also find K_{AA} , K_{BA} , and K_{BB} if AA , BA and BB are determined.

(*) They play an important role in finding whether A or B has a better chance to win in the simpletons' game.

John Conway's Equation

Theorem (John Conway)

The odds that B will win over A in the simpletons' game is $(K_{AA} - K_{AB}) / (K_{BB} - K_{BA})$.

This result was first obtained by J. Conway in early 1970s, but it just comes out as if by magic, not easy to follow. The shortest proof was obtained later in 1993 by P. A. Pevzner even there are other proofs given earlier. (Li, 1980; Guibas and Odlyzko, 1981)

Proof of Conway's Equation

Notations

- Let $AB = (c_0, c_1, \dots, c_{k-1})$. If $m_1, m_2, \dots, m_h$ are the indices of c_i 's such that $c_i = 1$, then the support of AB , $\text{supp}(AB) = \{m_1, m_2, \dots, m_h\}$.
- For example, $AB = (0, 1, 0, 0, 1, 1)$, $\text{supp}(AB) = \{1, 4, 5\}$.
- Let H_{AB} denote the set of h prefixes of $A = (a_0, a_1, \dots, a_{k-1})$ of length $m_1, m_2, \dots, m_h$.
- For $AB = (0, 1, 0, 0, 1, 1)$, $K_{AB}(t) = t + t^4 + t^5$, $H_{AB} = \{0,0100, 01001\}$,
 $P(H_{AB}) = K_{AB} = K_{AB}(1/2) = 1/2 + 1/16 + 1/32 = 19/32$.
- $A*B$ is the **concatenation** of A and B .

Notations - Continued

- A word W is an **A-victory** if it contains A in the end and does not contain B .
- A word W is an A-previctory if W^*A is an A-victory. Note that W^*A may have contained B before getting all letters of A .
- S_A denotes the set of all A-previctories.
- B-victories, B-previctories and S_B can be defined accordingly.
- **T** = { T is neither A-victory nor B-victory}. (**The set of no-victory words.**)

Important Facts

1. For every T in $\mathbf{T}$, T^*A corresponds to either an A-victory or a B-victory.
2. If T^*A corresponds to an A-victory, then T can be represented as W^*H_{AA} where W is an A-previctory.
3. If T^*A corresponds to a B-victory, then T can be represented as W'^*H_{BA} where W' is a B-previctory.
4. If T^*B corresponds to a B-victory, then T can be represented as W^*H_{BB} where W is an B-previctory.
5. If T^*B corresponds to a A-victory, then T can be represented as W'^*H_{AB} where W' is a A-previctory.

Conclusion

- Consider the set of all T in $\mathcal{T}$. Then $\mathcal{T}$ can be written as follows:

$$\mathcal{T} = (S_A * H_{AA}) \cup (S_B * H_{BA}) = (S_A * H_{AB}) \cup (S_B * H_{BB}) \quad (\text{The set of no victory words } T, \text{ from facts } 2 + 3, 4 + 5 \text{ respectively.})$$

Now, the probability of both sides of the above equation are equal. Therefore, $P(S_A)K_{AA} + P(S_B)K_{BA} = P(S_A)K_{AB} + P(S_B)K_{BB}$. Since the odds that B wins over A is equal to the ratio of $P(S_B)$ over $P(S_A)$, that is the probability of B-previctories over the probability of A-previctories, we conclude that $P(S_B)/P(S_A) = (K_{AA} - K_{AB}) / K_{BB} - K_{BA}$.

