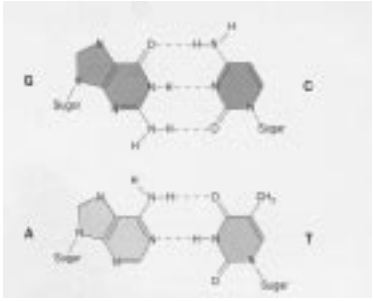
# ❑On Similarity Codes

Author : Arkadii G. D'yachkov
David C. Torney
From : IEEE Transactions on Information theory,
Vol.46, No.4, July 2000

Reporter : Institute of Electronic Engineering, NTHU
g893939 Te-Ming Chen

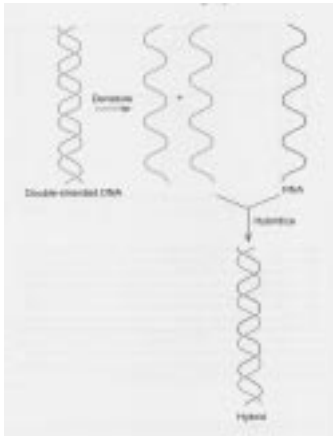
# ❑Biological motivation

- The base pairs of DNA
- The models of DNA structure



## ❑Biological motivation

- DNA Hybridization: technique for determining the similarity of 2 DNAs (or DNA and RNA) by reassociating single strands from each molecule and determining the extent of double-helix formation (indicating similar base sequences).



## ❑Definition

- Let $A \overset{def}{=} \{0,1,2,3\}$ be the quaternary alphabet, and let $\mathbf{x} = (x_1, x_2, \cdots, x_N), \quad x_i \in A$ and $\mathbf{y} = (y_1, y_2, \cdots, y_N), \quad y_i \in A.$
- For a pair $(\mathbf{x},\mathbf{y})$ of quaternary words, define a similarity

$$S(\mathbf{x}, \mathbf{y}) \overset{def}{=} \sum_{i=1}^{N} \varsigma(x_i, y_i) \qquad (1.1)$$

where the alphabetic similarity $\varsigma(x, y)$ of $x, y \in A$ is defined as follows:

$$\varsigma(x,y) \overset{def}{=} \begin{cases} 3, & \text{if } x = y = 0 \text{ or } x = y = 3 \\ 2, & \text{if } x = y = 1 \text{ or } x = y = 2 \\ 0, & \text{otherwise.} \end{cases} \qquad (1.2)$$

## ❑Definition

- If $\mathbf{x} \neq \mathbf{y}$, then the number $S(\mathbf{x},\mathbf{y})$ will be called a cross-similarity of pair $(\mathbf{x},\mathbf{y})$.
- Denote by $D(\mathbf{x},\mathbf{y})$ the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$.
- Definitions (1.1) and (1.2) mean that the following inequalities are true:

$$S(\mathbf{x}, \mathbf{y}) \leq 3[N - D(\mathbf{x}, \mathbf{y})]$$

$$D(\mathbf{x}, \mathbf{y}) \geq N - \frac{S(\mathbf{x}, \mathbf{y})}{2} \qquad (1.3)$$

## ❑Definition

- Let $N \geq 2$ and $t \geq 2$ be integers and $X = (\mathbf{x}(1), \mathbf{x}(2), \cdots, \mathbf{x}(t))$ be an arbitrary quaternary code of length $N$ with $t$ codewords $\mathbf{x}(u) = (x_1(u), x_2(u), \cdots, x_N(u))^T$,

  $x_i(u) \in A, \quad i = 1,2,\cdots,N, \quad u = 1,2,\cdots,t.$
- Let S, $0 \leq S < 2N$, be a fixed integer called a cross-similarity threshold. We say that the code $X$ is an S-similarity code of length $N$ if for any pair $(u, v)$ of distinct indices $u \neq v$, the cross-similarity

$$S(\mathbf{x}(u), \mathbf{x}(v)) \leq \mathrm{S} < 2N \qquad (1.4)$$

  By virtue of (1.3) and (1.4), all $t$ codewords of an S-similarity code are distinct.

## ❑Bounds on the rate of similarity codes

- Denoted by $t(N,S),\ 0<S<2N,$ the maximal possible size of the S-similarity code of length N. For an arbitrary fixed $s,\ 0<s<2,$ defined the function

$$R(s) \overset{def}{=} \overline{\lim_{N\to\infty}} \frac{\log_2 t(N,\lfloor sN \rfloor)}{N} = \overline{\lim_{N\to\infty}} \frac{\log_2 t(N,\lceil sN \rceil)}{N}$$

called an asymptotic binary rate of S-similarity codes.

## ❑Bounds on the rate of similarity codes

- Upper bound:

  If $0 < s < \frac{3}{5},$ then $R(s)=0.$

  If $\frac{3}{5} < s < 2,$ then $R(s) \le \overline{R}(s) \overset{def}{=} \frac{10}{7}(s-\frac{3}{5}).$ (2.1)

- Lower bound:

  For any s, $\frac{3}{5} < s < 2$, the asymptotic binary rate

$$R(s) \ge \underline{R}(s) \overset{def}{=} E(\frac{1}{5}, s).$$ (2.2)

## ❑Liner similarity codes

- The binary rate and the asymptotic binary rate of any linear $\lfloor sN \rfloor$-similarity code satisfy the inequalities

$$\frac{2K}{N} \le \frac{16}{19}\left(s\frac{t_K - 1}{t_K - 4} - \frac{5}{8}\right) + \frac{\log_2 K}{K}, \quad t_K = 4^{\left\lfloor \frac{1}{2}\log_2 K \right\rfloor}, \quad K \ge 16. \tag{3.1}$$

$$R_{Lin}(s) \le \overline{R}_{Lin}(s) \overset{def}{=} \begin{cases} 0, & \text{if } 0 < s \le \frac{5}{8} \\ \frac{16}{19}\left(s - \frac{5}{8}\right), & \text{if } \frac{5}{8} < s < 2 \end{cases} \tag{3.2}$$

- For linear S-similarity codes, (3.2) obviously improves the upper bound of (2.1).

## ❑Conclusion

- The similarity codes will have the property that stable duplexes will not form between a reverse complement of any codeword and any other codeword.
- The similarity code embodies a molecular specificity of duplex formation.

## ❑Further problems

- Obvious shortcoming : the restriction to “aligned hybridization”.
- DNA sequences could hybridize:

  (1) shifting on sequence along another,

  (2) deletions of several of letters from sequence,

  (3) forming a self-duplex from one strand.
- The pertinent measure of sequence similarity could be chosen to be a score for a variant of sequence alignment.