The Central Dogma of Genetics

Or the Coding Theory Behind it

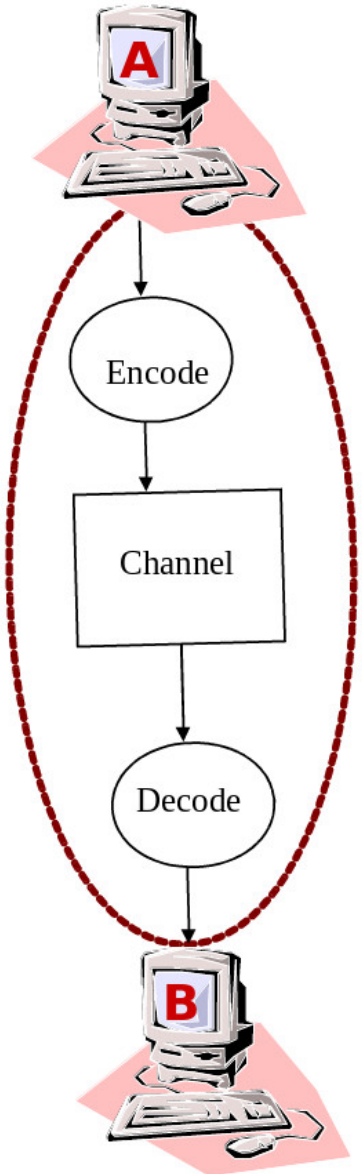
Artur Schäfer

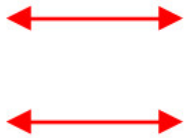
University of St. Andrews

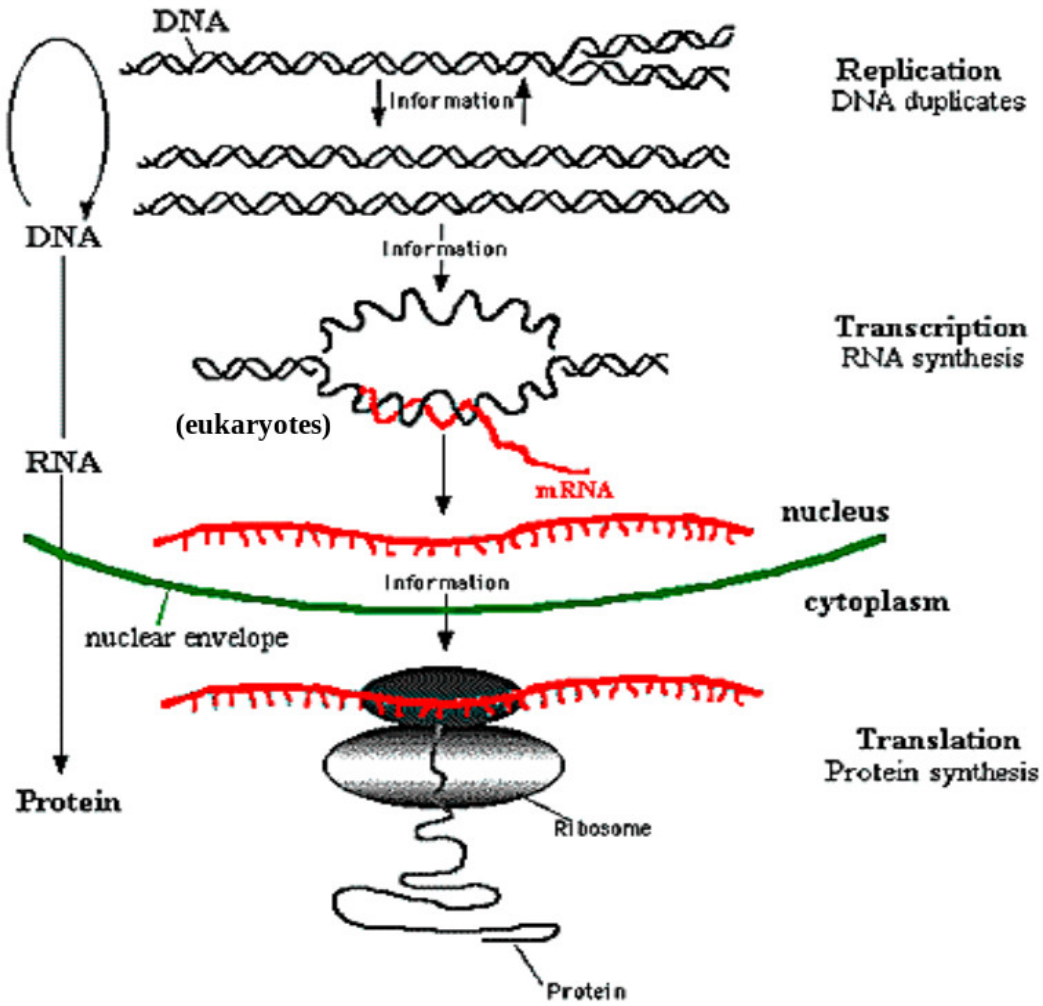
PPS 2014, Oct 17th

Quiz: What is the next number? 3, 9, 21, 45, 93,...

Central Dogma of Genetics = Genetic Information Transmission







The Central Dogma of Molecular Biology

(http://www-stat.stanford.edu/~susan/courses/s166/central.gif)

1. Introduction to Coding Theory
2. Linear Codes and Related Codes
3. Group Ring Codes of extra-special Groups
4. Orthogonal Array's and Codes

Contents

Definition

Let A be an alphabet. A code C of length n over the alphabet A is a set of n-tuples with entries in A.

Beispiel

Using $A = \{0, 1, 2\}$, then $C = \{000, 121, 212\}$ is a ternary code of length 3.

Recall: The **Hamming distance** $d(x_1, x_2)$ between two n-tuples is the number of coordinates, where x_1 and x_2 do not coincide.

The Hamming distances for C are $\{3, 3, 3\}$.

Definition

The **minimum distance** of a code C is $\min\{d(x_1, x_2) : x_1, x_2 \in C\}$.

To each code we can attach parameters (n, M, d, q).

- $n =$ length of C
- $M = \#$ elements in C
- $d =$ minimum distance of C
- $q =$ size of A

Beispiel

$C = \{000, 121, 212\}$ is a $(3, 3, 3, 3)-$code.

Goal of Coding Theory

Given n, d and q.

Find a code C with M as big as possible!

Reason:

- A code with $d > 2t + 1$, can correct t errors, for any t.
- A code with big M is more useful than for small M.

Contents

Definition

1. Let F be a finite field and n a non negative integer. A **linear code** C is a subspace $C \leq F^n$.
2. Let C be a linear code. C is **cyclic** if
$$(c_0, ..., c_{n-2}, c_{n-1}) \in C \Rightarrow (c_{n-1}, c_0, ..., c_{n-2}) \in C.$$
3. If C is cyclic then, via $(c_0, ..., c_{n-1}) \mapsto \sum_{i=0}^{n-1} c_i x^i$, we can identify C as an ideal $C \trianglelefteq F[x]/(x^n - 1)$.

Since the code is subspace we use a matrix G to describe it, where the rows of G form a basis of this space.

Beispiel

Let $C \leq \mathbb{F}_4^3$ be given by $G = \begin{pmatrix} 1 & 0 & \alpha^2 \\ 0 & 1 & \alpha \end{pmatrix}$, with $\mathbb{F}_4^\times = \langle \alpha \rangle$.

(**Reed-Solomon**). C is cyclic and satisfies

$$(c_1, ..., c_{n-1}, c_n) \in C \Rightarrow (c_n, c_1, ..., c_{n-1}) \in C.$$

Other cyclic codes: BCH-codes, binary Hamming-codes.

Alternative description

Also, it is common to provide a **parity check matrix** H. The code C is the kernel of this matrix. $\Rightarrow GH^T = 0$.

Beispiel

$G = \begin{pmatrix} 1 & 0 & \alpha^2 \\ 0 & 1 & \alpha \end{pmatrix}$, with $\mathbb{F}_4^\times = \langle \alpha \rangle$. We get

$$H = \begin{pmatrix} \alpha^2 & \alpha & 1 \end{pmatrix}$$

Definition

$C \leq F^n$ a code

$$\begin{aligned} Perm_n(F) &= n \times n \text{ permutationmatrices} \cong S_n \\ Mon_n(F) &= n \times n \text{ monomial matrices} \cong \left(F^\times\right)^n \rtimes S_n \\ Perm(C) &= \{M \in Perm_n(F) \mid C.M = C\} \\ Mon(C) &= \{M \in Mon_n(F) \mid C.M = C\}. \end{aligned}$$

If two codes $C, D \leq F^n$ satisfy $C.M = D$, for $M \in Mon_n(F)$, then they are equivalent.

Lemma

Equivalent codes have the same parameters.

$C = \{000, 121, 212\}$ and $D = \{000, 111, 222\}$ are equivalent $(3, 3, 3, 3)-$codes.

Cyclic Codes are the most Important ones

So, far cyclic codes have the best parameters (n, M, d, q) for practical use so far.
$\rightarrow$ Reed-Solomon codes for CD's.

Does a generalization of cyclic codes lead to better codes?

For a cyclic code C a monogenetic inverse semigroup $C_n = \langle g \rangle$, we have the following correspondence:

$$\begin{array}{ccccc} C \leq F^n & \leftrightarrow & I \trianglelefteq F[x]/(x^n - 1) & \leftrightarrow & \widetilde{I} \trianglelefteq FC_n \\ (c_1, ..., c_n) & \leftrightarrow & \sum_{i=1}^{n} c_i x^i & \leftrightarrow & \sum_{i=1}^{n} c_i g^i \end{array}$$

$\Rightarrow$ Consider codes of group rings FG

Definition

Let G be a finite group, F a finite field and $\{g_1, ..., g_n\}$ a basis of the vector space FG. Then, any ideal I of FG defines a code $C \leq F^n$ by:

$$(a_1, a_2, ..., a_n) \in C \Leftrightarrow a_1 g_1 + a_2 g_2 + \cdots + a_n g_n \in I.$$

Any code equivalent to C, for any ideal is a G-code.

Definition

If G is abelian/cyclic, then we say the G-code is abelian/cyclic.

Contents

We consider semi-simple group rings over a field with sufficiently many roots of unity.

Definition

An extra-special group is a group G, with $Z(G) = G' = \{1\}$ (cetre subgroup, commutator subgroup).

Cosets

We get the coset decomposition G/G'

$$G = \{1, g, ..., g^{p-1}, \underbrace{t_2, t_2 g, ..., t_2 g^{p-1}}_{p \text{ elements}}, ..., \underbrace{t_{p^{2n}}, t_{p^{2n}} g, ..., t_{p^{2n}} g^{p-1}}\}$$

Theorem

An extra-special group G has the structure of a symplectic vector space V with

$$V = \langle a_1, b_1 \rangle \perp \cdots \perp \langle a_n, b_n \rangle.$$

Corollary

$$FG = \underbrace{\bigoplus_{i=1}^{p^{2n}} \underbrace{e_i FG}_{\dim(\cdots)=1}}_{I} \oplus \underbrace{\bigoplus_{k=1}^{p-1} \underbrace{f_k FG}_{\dim(\cdots)=p^{2n}}}_{I^\perp}.$$

$$I = \begin{pmatrix} \overbrace{1 \cdots 1}^{p\text{-times}} & 0 \cdots 0 & 0 \cdots 0 & \cdots & 0 \cdots 0 \\ 0 \cdots 0 & 1 \cdots 1 & 0 \cdots 0 & \cdots & 0 \cdots 0 \\ 0 \cdots 0 & 0 \cdots 0 & 1 \cdots 1 & \cdots & 0 \cdots 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & \cdots & 1 \cdots 1 \end{pmatrix} \in F^{p^{2n} \times p^{2n+1}}.$$

$\Rightarrow$ I is a semi-cyclic code, see **repetition code**.

$$\Rightarrow I^{\perp} = \perp_{i=1}^{p^{2n}} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}.$$

$\Rightarrow$ $I^{\perp}$ is a semi-cyclic code.

Contents

Definition

A $t-(v,k,\lambda)$ orthogonal array (OA) is a $\lambda v^t \times k$ array whose entries are chosen from a set X with v points such that in every subset of t columns of the array, every t-tuple of points of X appears in exactly λ rows.

Beispiel

$$OA = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 3 \\ 2 & 1 & 2 \\ 2 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 3 \\ 3 & 2 & 1 \\ 3 & 3 & 2 \end{pmatrix}$$

Definition

A **Latin hypercube** of order n and dimension m is $\mathbb{Z}_n^m$, where each tuple is labelled with an additional entry from $1, .., n$ such that each line contains an entry only once.

Latin hypercubes are orthogonal arrays, but not every OA is a Latin hypercube.

$$OA = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 3 & 3 & 3 \\ 2 & 1 & 2 & 3 \\ 2 & 2 & 3 & 1 \\ 2 & 3 & 1 & 2 \\ 3 & 1 & 3 & 2 \\ 3 & 2 & 1 & 3 \\ 3 & 3 & 2 & 1 \end{pmatrix}$$

Definition

Two LHC are **mutually orthogonal**, if their labels form tuples where each tuple appears exactly n^{m-1} times.

New Stuff!

Definition

A **Hamming graph** $H^S(m, n)$ is a graph with vertices $\mathbb{Z}_n^m$, where two vertices have an edge if their Hamming distance is in the set $S \subseteq \{1, ..., m\}$.

Proposition

If $S = \{k + 1, ..., m\}$, then maximal cliques of size n^{m-k} of $H^S(m, n)$ are orthogonal arrays.

$\Rightarrow$ finding mutually orthogonal Latin squares is equivalent to finding maximal cliques!

Thank You for Your Attention