Error Correction in the DNA

Model Evolution as a Communication Model using the DNA

Does the DNA contain an error correction code?

A. Redundancy in the DNA

B. Communication Model and its assumptions

C. Prior Work

D. Work Done so far: Modeling code in DNA as LDPC

E. Future Work

Redundancy and Mutations in the DNA

- The Complementary Double helix- Single chain contains all information.
- Some combinations of nucleotides result in the same codon

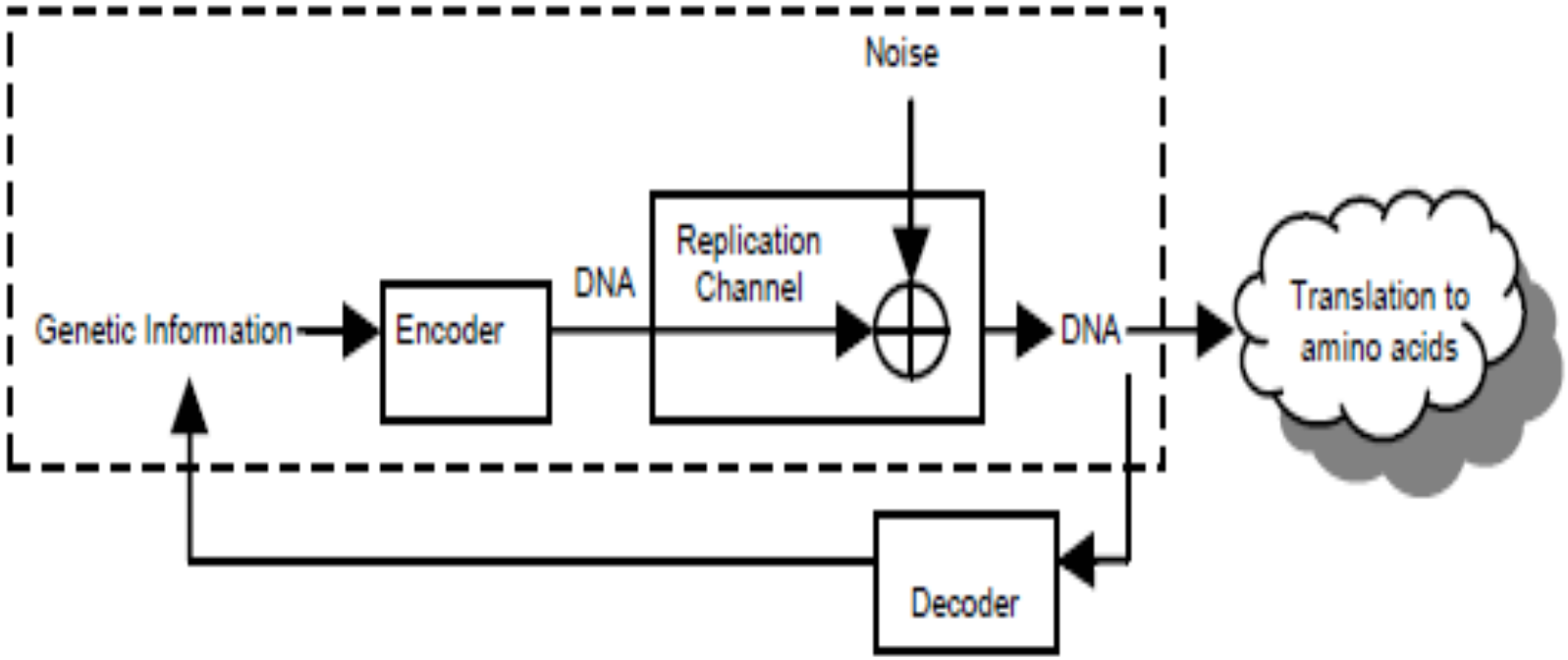
Types of mutations

- Spontaneous : due to addition, elimination, transversion, copying errors
- Induced : due to chemicals and UV radiation, viruses etc
- Another classification: Point mutation or Substitutions
  - Transition and Transversion
  - Silent mutations: which code for the same amino acid.
  - Missense mutations: which code for a different amino acid.

Gerard Battail proposed a proof that shows that **the genome length of species is limited if there is no error correction in the DNA.**
With coding, the average lifetime of a genome can assume values of the order of magnitude of geological times.

Communication model

Thus, the mutation is modeled as noise causing errors in the DNA sequence.



Generator matrix G and Parity check matrix H used in the encoding

C = { mG | m Î {0,1}^k }

c =m ·G =m0g0 + m1g0 +……+ mk-1gk-1

H is the dual of G.

$$[c_1 \quad c_2 \quad c_3 \quad c_4 \quad c_5 \quad c_6 \quad c_7] \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \mathbf{0}$$

$$\begin{aligned} c_1 + c_4 + c_6 + c_7 &= 0 \\ c_2 + c_4 + c_5 + c_6 &= 0 \\ c_3 + c_5 + c_6 + c_7 &= 0 \end{aligned} \quad \Rightarrow \quad \begin{aligned} c_1 &= c_4 + c_6 + c_7 \\ c_2 &= c_4 + c_5 + c_6 \\ c_3 &= c_5 + c_6 + c_7 \end{aligned}$$

$c_1 \oplus c_2 \oplus c_5$

$$H = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

LDPC: Formation uses a lot of randon… …ta.

Transition Probabilities

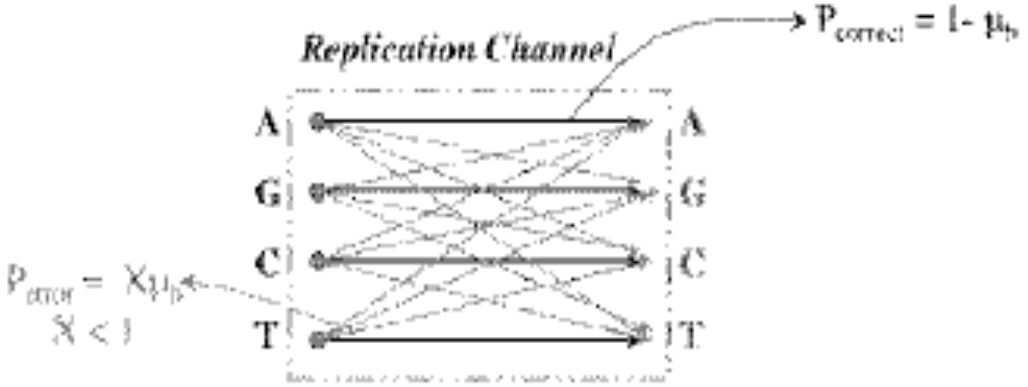
A, G, T, C represented as roots of a primitive polynomial in GF(4).

$$\alpha^0 = 1 \Leftrightarrow 1 \Leftrightarrow C$$

$$\alpha^1 = \alpha \Leftrightarrow 2 \Leftrightarrow T$$

$$\alpha^2 = \alpha + 1 \Leftrightarrow 3 \Leftrightarrow G$$

$$0 = 0 \Leftrightarrow 0 \Leftrightarrow A$$



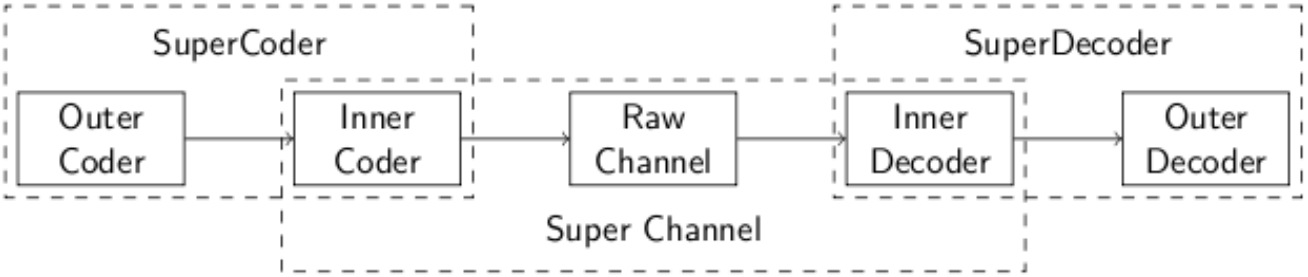
	A	G	C	T
A	$1-\mu_b$	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$
G	$\frac{\mu_b}{3}$	$1-\mu_b$	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$
C	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$	$1-\mu_b$	$\frac{\mu_b}{3}$
T	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$	$1-\mu_b$

	A	G	C	T
A	$1-\mu_b$	$\frac{2\mu_b}{3}$	$\frac{\mu_b}{6}$	$\frac{\mu_b}{6}$
G	$\frac{2\mu_b}{3}$	$1-\mu_b$	$\frac{\mu_b}{6}$	$\frac{\mu_b}{6}$
C	$\frac{\mu_b}{6}$	$\frac{\mu_b}{6}$	$1-\mu_b$	$\frac{2\mu_b}{3}$
T	$\frac{\mu_b}{6}$	$\frac{\mu_b}{6}$	$\frac{2\mu_b}{3}$	$1-\mu_b$

Nested codes

Battail further proposed:

- Some proteins like haemoglobin are found from simple to complicated species
- Characteristics that vary from person to person are uncoded



Prior Work

Authors have investigated the presence of algebraic codes in the DNA by solving a set of equations for each parity bit. They check if that code is valid for other sections of the DNA, and for longer codes.

This has not been very successful.

Our Work

- Fit an LDPC code in parts of the DNA that code to proteins. (Battail).
- Portions around it were taken to be the parity bits.
- Resilient proteins were fit with nested codes
- Reverse engineer: Assume that the DNA was coded with an LDPC code, and decode part around protein to a meaningful message- which is the protein itself.
- Problem: Using "real" mutation rates caused hardly any change (errors) in the DNA sequence.

Future work

1. Introduce more mutations than the prescribed rates in parts around resilient proteins to show they can withstand more errors than other proteins.
2. Not necessarily develop an explicit code: Make theoretical predictions about its nature
3. Check mutual information between nucleotides: Conditional probability of a nucleotide occurring at a given location based on another.
4. Use characteristics of LDPC codes to support or refute the assumption that the DNA is made of such codes.
5. Use information about specific mutations