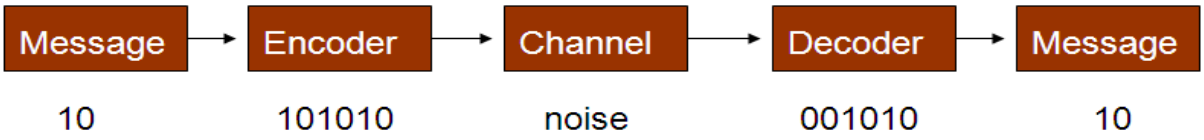
Model Evolution of DNA as a communication process. Does the DNA contain an error correction code?

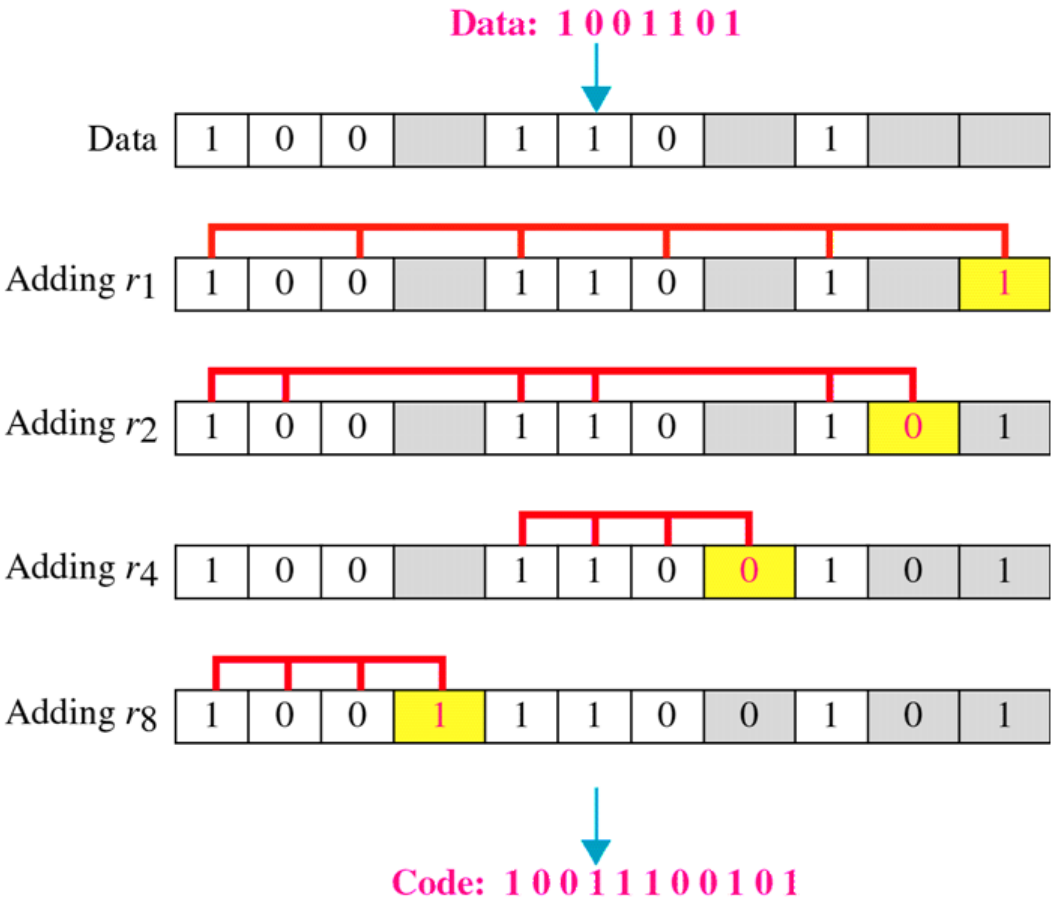
A. ECC
B. DNA + Broad Objective
C. Literature Survey
C. Proposed Work: More Specific
D. Already done, what we've learnt from that
E. New Experiment

A. Error Control Coding:

Transmission of Information (in the form of bits) content across a channel. AWGN adds to the data and corrupts it. Message will end up getting destroyed. To prevent the message from getting entirely lost some protection bits (for the sake of redundancy) called parity are added to it. The parity bits are computed using the message bits and can not just detect but also correct the errors.



C1 = {000, 111} binary repetition code of length 3
C2 = { 00000, 01100, 10110} binary code of length 5
C3 = {0000, 0111, 0222, 1012, 1020, 1201, 2021,2102, 2210} ternary code of length 4



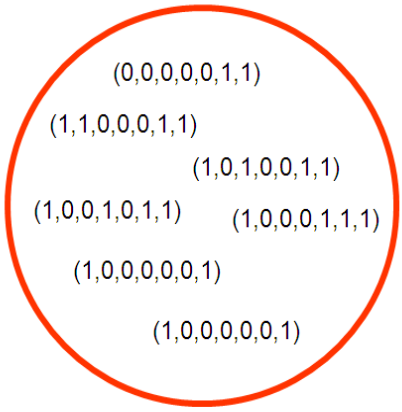
Basic assumptions

- If i < j then i errors are more likely than j errors
- Errors occur randomly
- The Hamming distance between two words over the same alphabet is the number of places where the symbols differ. Example : d(100111, 001110) = 3
- A code C can correct up to t errors if d(C) >= 2t + 1
- Nearest neighbor decoding

The *sphere of radius r* about a vector *u* is defined as:

$$S_r(u) = \{v \in V \mid dist(u,v) \le r\}$$

e.g. *u*=(1,0,0,0,0,1,1)



Goal: Correct as many errors as possible while using as little redundancy as possible

Generator matrix of code C: A k X n matrix with elements from GF(2) whose rows form a basis of a linear [n,k] -code (k dimensional subspace) C. Example G of C2:

$$C_2 = \begin{Bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{Bmatrix} \quad \text{is} \quad \begin{Bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{Bmatrix}$$

C = { mG | m Î {0,1}^k }
c =m ·G =m0g0 + m1g0 +……+ mk-1gk-1
gi is the i-th row vector of G. The rows of G are linearly independent since G is assumed to have rank k.

$$\boldsymbol{G} = \begin{bmatrix} \overline{\boldsymbol{g}}_0 \\ \overline{\boldsymbol{g}}_1 \\ \vdots \\ \overline{\boldsymbol{g}}_{k-1} \end{bmatrix} = [\boldsymbol{I}_k \boldsymbol{P}]$$

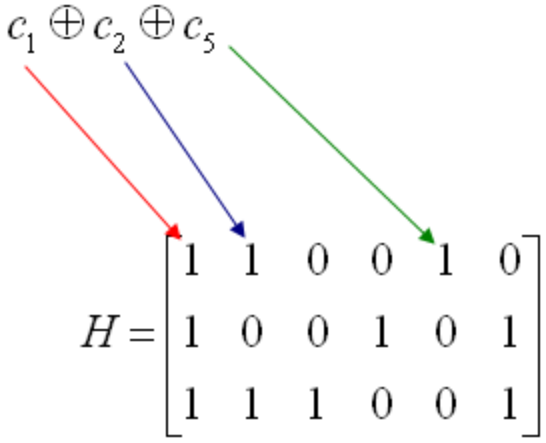
Systematic form: $\underbrace{\quad}_{n-k \text{ check bits}} \quad \underbrace{\quad}_{k \text{ information bits}}$

Parity-check matrix H for an [n,k] -code C is an (n-k)xn matrix that is the dual of G.
C = {c Î V(n,q) | $cH^T = 0$} as $GH^T = 0$, c = mG and so $cH^T = mGH^T = 0$.

H = [I_{n-k} | P^T]

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

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



Low Density Parity Check codes (LDPC): Class of linear block code. The term “Low Density” refers to the characteristic of the parity check matrix which contains only few ‘1’s in comparison to ‘0’s. LDPC codes are arguably the best error correction codes in existence at present. Not gonna intuite why it has great efficiency, performance and achieves shannon capacity, something that was once only theoretical.
Important thing is that LDPC parity check matrix can be graphically represented as a Tanner (Bipartite) graph. This formation uses a lot of randomness which resembles real life data.

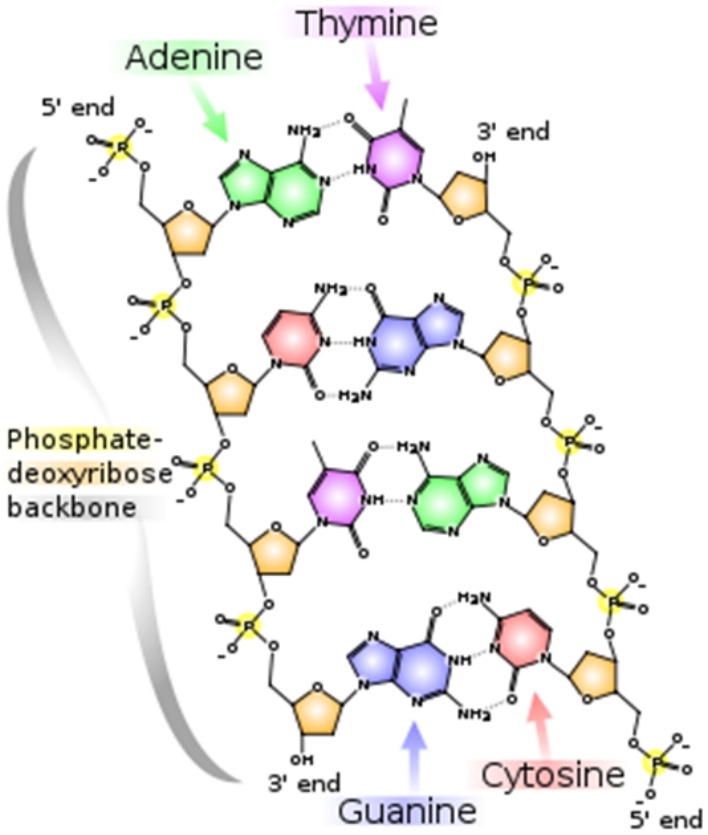
Row weight - number of ‘1’s in a row Number of- Number of symbols taking part in a parity check
Column weight - number of ‘1’s in a column- Number of times a symbol takes part in parity checks

B. DNA + Broad Objective

DNA is genetic code of life, found on chromosome located in nucleus. Gene is fundamental functional and physical unit of heredity. It carries information from one generation to another. *This is what we would like to model as a communication channel.*

- DNA molecule consists of two unidimensional chains (or strands) each made of alternating phosphate and deoxiribose groups where each sugar group is covalently bound to one of four nitrogenous “nucleic bases” or “nucleotides,” namely adenine A, thymine T, guanine G, and cytosine C.
- The nucleic bases A and G are purines, while T and C are pyrimidines.
- The bases A and T bound to the facing strands are tied together by two hydrogen bonds, while three hydrogen bonds tie together G and C.
- Law of complementary base pairing, one strand determines base sequence of other. *Is a simple repetition code.*

- Genome is the entirety of an organism's heredity information, contain both coding regions i.e. genes and non-coding region i.e. junk-DNA. 98% of DNA noncoding – "junk" or regulatory. *This much redundancy indicates the possibility of the implementation of an ECC.*
- DNA is a systematic code where the message appears by itself.

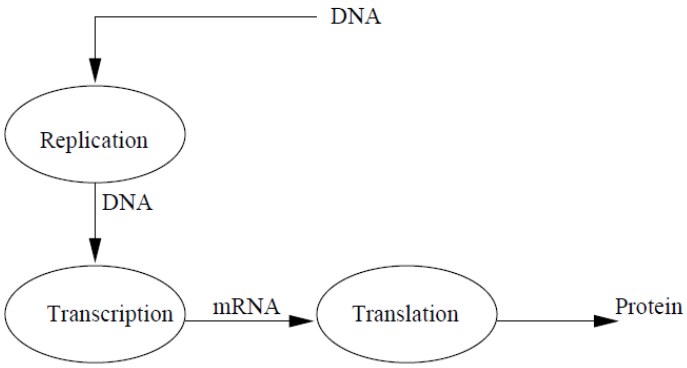


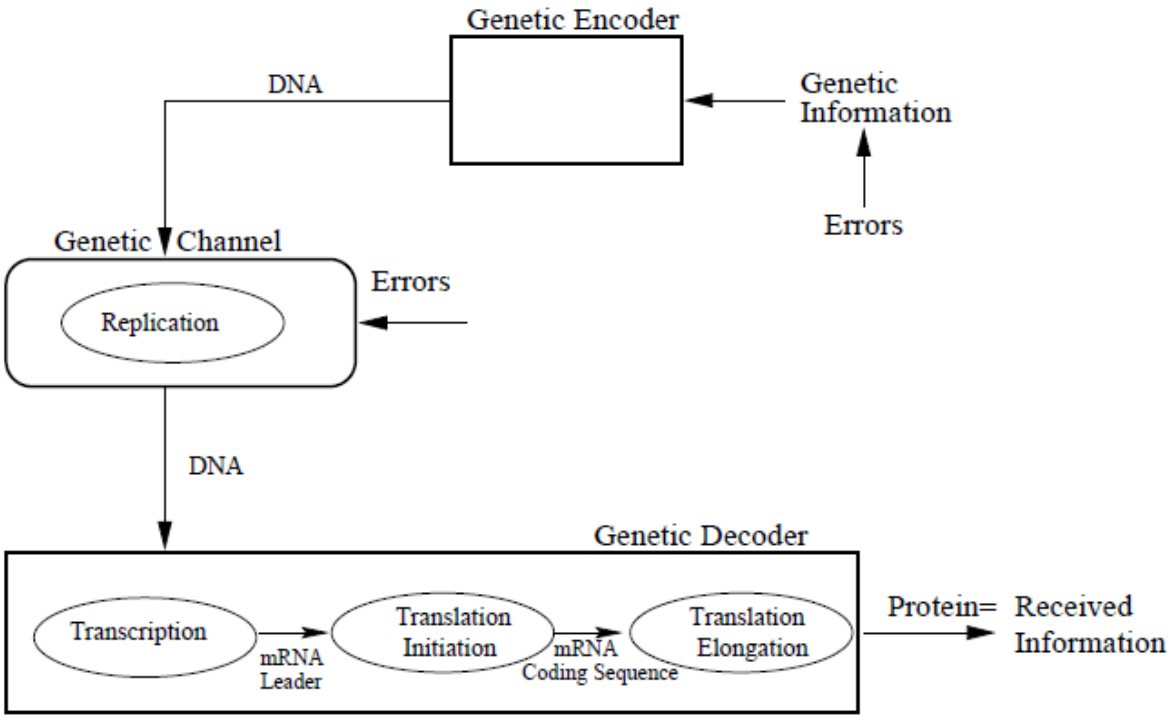
- Genes generally express their functional effect through the production of proteins, which are complex molecules responsible for most functions in the cell.
- Each group of three nucleotides in the sequence, called a codon, corresponds either to one of the twenty possible amino acids in a protein. One amino acid can have approx 1-6 combinations. For example alanine can have 4 combinations i.e. GCU, GCC, GCA, GCG. *Again, this indicates that there is some inherent redundancy at this stage too.*
- Can we also quantify the actual information content.

Central Dogma and Comm Model

¨The process of The majority of genes are expressed as the proteins they encode. The process occurs in two steps: Transcription = DNA → RNA
Translation = RNA → protein

Taken together, they make up the "central dogma" of biology: DNA → RNA → protein.





Mutations:

Two Basic Things can happen in which DNA molecule can:

1. Replicate: Then You have 2 from 1 (DNA REPLICATION---Leads to MITOSIS. one cell goes to two cells- their full genome is copied and each daughter cell inherits one copy.

2. Selected sections of the DNA strand (genes) can be expressed (transcribed and into mRNA and Translated into proteins)

Mutations are changes in a genomic sequences, the DNA sequences of a cell.

- Somatic mutations occur in somatic cells and only affect the individual in which the mutation arises. Not passed to offspring. Some type of skin cancers and leukemia result from somatic mutations
- Germ-line mutations alter gametes and passed to the next generation.

Mutations are quantified in two ways:
1.Mutation rate = probability of a particular type of mutation per unit time (or generation).
2.Mutation frequency = number of times a particular mutation occurs in a population of cells or individuals.

- Their effects may not be serious unless they affect an amino acid that is essential for the structure and function of the finished protein molecule (e.g. sickle cell anaemia)
- Many mutations are repaired by enzymes
- Some mutations may improve an organism's survival (beneficial)

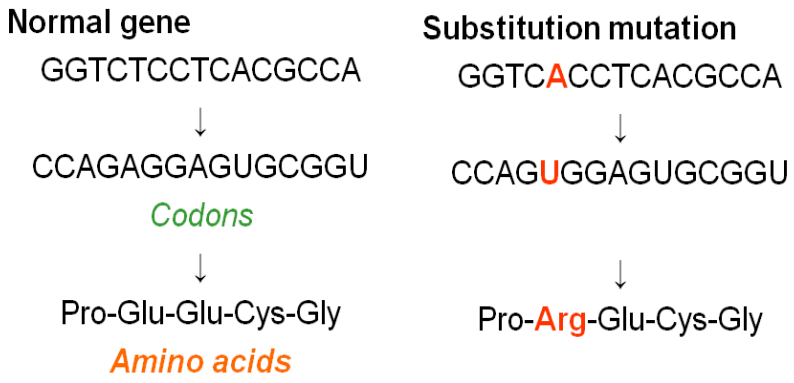
Types of mutations

- Spontaneous : due to addition, elimination, transversion, copying errors
- Induced : due to chemicals and UV radiation, viruses etc

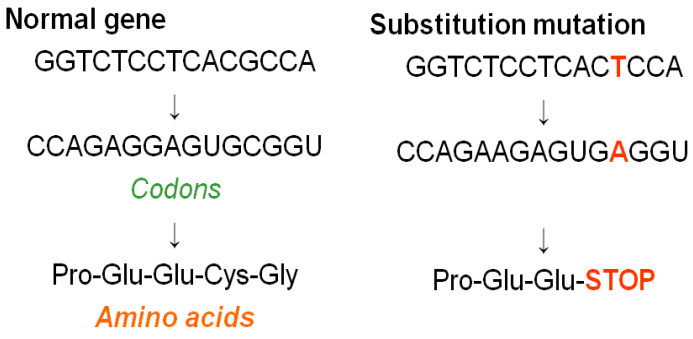
Another Classification:

I. Point mutation or Substitution will only affect a single codon (transition and transversion): basically caused by silent mutations, missense and non sense mutations. Most transitions results in synonymous substitution.

1. Transition : when a purine base is replaced by purine or a pyrimidine is replaced by pyrimidine. 4 types of transitions; A « G and T « C
1. Transversion: when a purine is replaced by pyrimidine or vice-versa. 8 types of transversions; A « T, G « C, A « C, and G « T. Transversion more likely to result in nonsyn substitution.
2. Silent mutations: which code for the same amino acid.
3. Missense mutations: which code for a different amino acid.

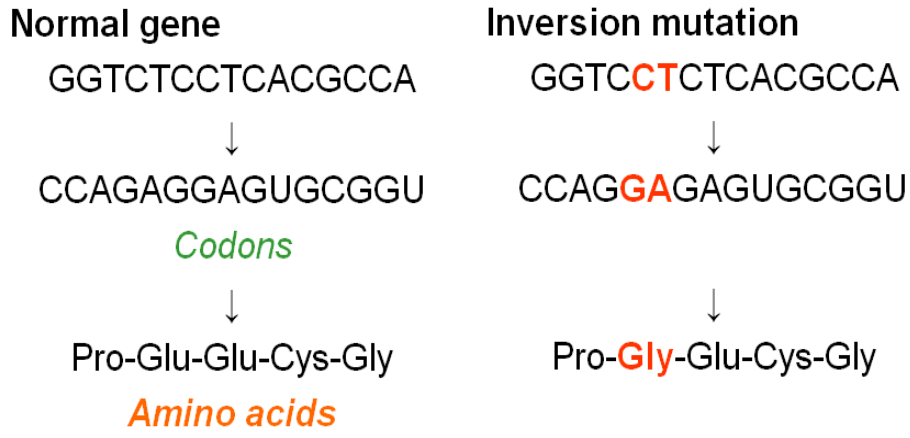


4. Nonsense mutations: which code for a stop and can truncate the protein.

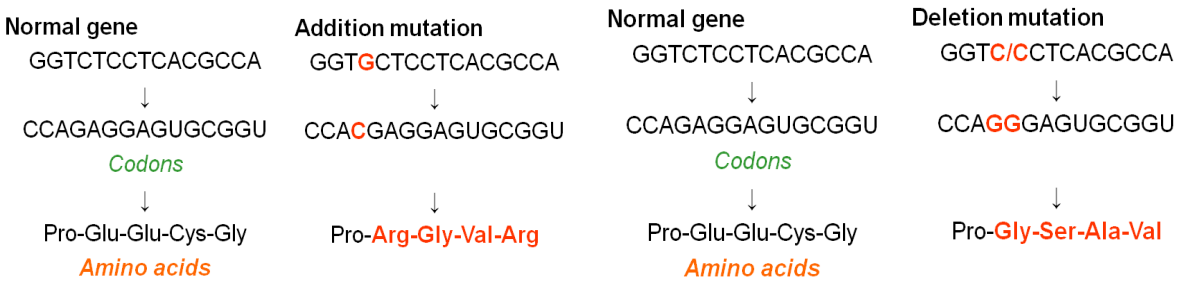


5. Changes in the third base of a codon often have no effect.

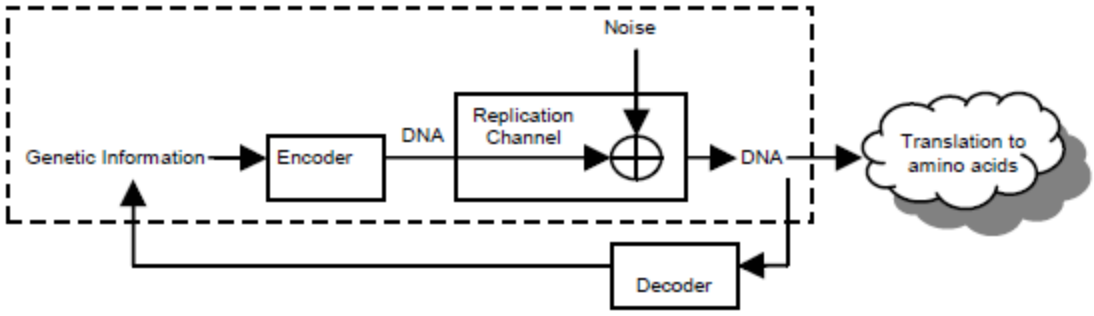
II. Inversion:



III. Insertion or Additions and deletions or frame shift mutation - Changes the “reading frame” like changing a sentence causing proteins to be built incorrectly.

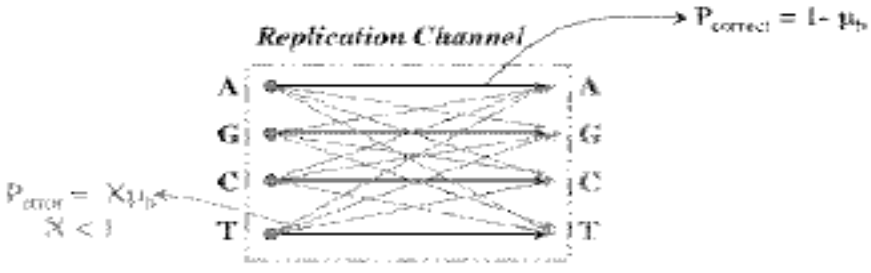


Thus, the mutation is what we are trying to model as noise causing errors in the DNA sequence.
Note, we don't mean that one base is converted to another. Just that one base may indicate the presence of another.



Other basic assumptions:
A, G, T, C represented as roots of a primitive polynomial in GF(4). When we allocate them values 0,1,2,3 modulo 4 or 00,01,10,11 we are disadvantaging some bases for other.

$$\alpha^0 = 1 \Leftrightarrow 1 \Leftrightarrow C$$
$$\alpha^1 = \alpha \Leftrightarrow 2 \Leftrightarrow T$$
$$\alpha^2 = \alpha + 1 \Leftrightarrow 3 \Leftrightarrow G$$
$$0 = 0 \Leftrightarrow 0 \Leftrightarrow A$$



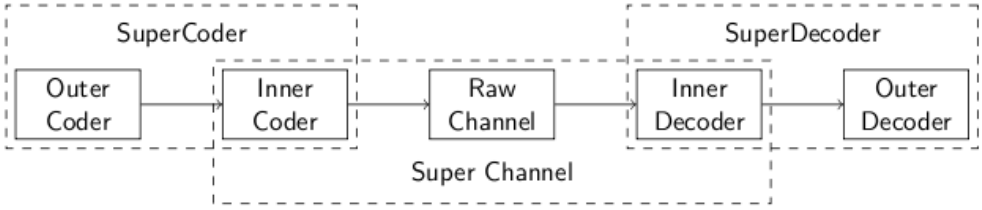
Channel transition probability assuming:
p(TransitionMuation)=p(TransversionMutation)
p(Transition Mutation) not=p(TransversionMutation)

	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	$\frac{\mu_b}{6}$	$\frac{\mu_b}{6}$	$\frac{2\mu_b}{3}$	$1-\mu_b$

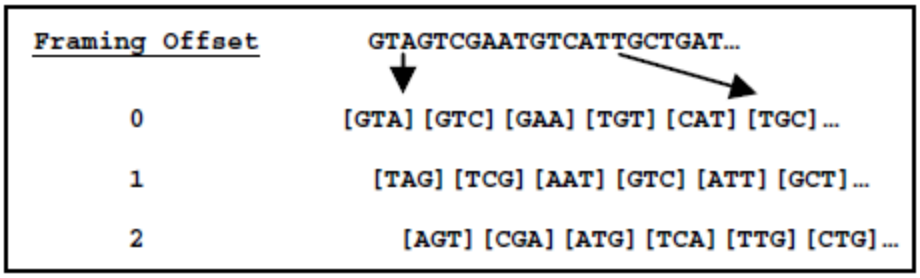
Nested Codes:

- Some proteins like haemoglobin are observed across the simplest to the most complicated species.Since mutations can occur anywhere in the sequence, the fact that some regions are intact suggests that they are protected more.
- Nested codes are codes coded on top of other codes. Simple 2 step concatenation is shown.
- Here different combinations can give variation in protein for eg in 2 species say A n B we have a common protein insulin. Insulin can be formed by ATC in one species and by CTC in another.
- Also characteristics that vary from person to person are uncoded.



Related Prior Work:

Solved a set of parity equations investigate the presence of a single parity check using modulo 4 arithmetic Gaussian elimination. Considered a parity for every set of codons. No rationale for why a certain nucleic base must be information/parity bit. Needn't be the case, there can be a code for a set of proteins alone separately. Also 0 is closer to 1 than to 2. And, Assume a uniform code, but it can vary for different sections of the DNA.
Not taken into consideration any existing redundancy like that of the amino acids.



USing a subspace partitioning approach

Algorithm Outline

1. *Obtain the orthonormal basis,* $\{e_1, e_2, \ldots, e_n\}$*, by Gram-Schmidt orthogonalization of j number of* v_i *frames where* $j \geq n$*. Form the transform matrix,* $\mathbf{G}$*, from this set.*
2. *Decompose the sequence into its basis components,* $\{t_1, t_2, \ldots t_{N-j}\}$*, across all possible framing offsets.*
3. *Note the persistence of nulls in* t_i*'s. Calculate confidence by comparing against the probability of sequential sets of randomly chosen vectors having the same subspace partitioning.*

Have used only single or few parities. There could be more.

More Specific Details:

Reverse Engineering:

Assuming that the DNA had LDPC codes we tried to decode them to a meaningful message.

Easier to reverse engineer (decode with an LDPC code and check if that was right) than to explicitly search. Coz LDPC codes have more constraints than normal linear codes.

Two LDPC codes with permutations on the bits, with same wr, wc configuration have equivalent behaviour.

Wasn't enough coz the low mutation rates caused hardly any change (errors) in the DNA sequence.

Assumed that resilient proteins that were present across many generations had more *parity bits* than use a nested code system.

Future:

Do not convert to binary. Use as symbols.

Implement concatenated structure. Many insightful properties will help crack the code easily.

Example: The code itself is a matrix rather than a vector. Every row of the matrix corresponds to G1 and every column to G2. Properties like this can be used to verify the code.

Introduce more mutations than the prescribed rates just to prove which proteins are more resilient to mutations than other. Quantify this value.

Perturb the input sequence a little to let a few samples of the coe fit into the framework of a particular parity check matrix.

Deletion = Erasure correction.