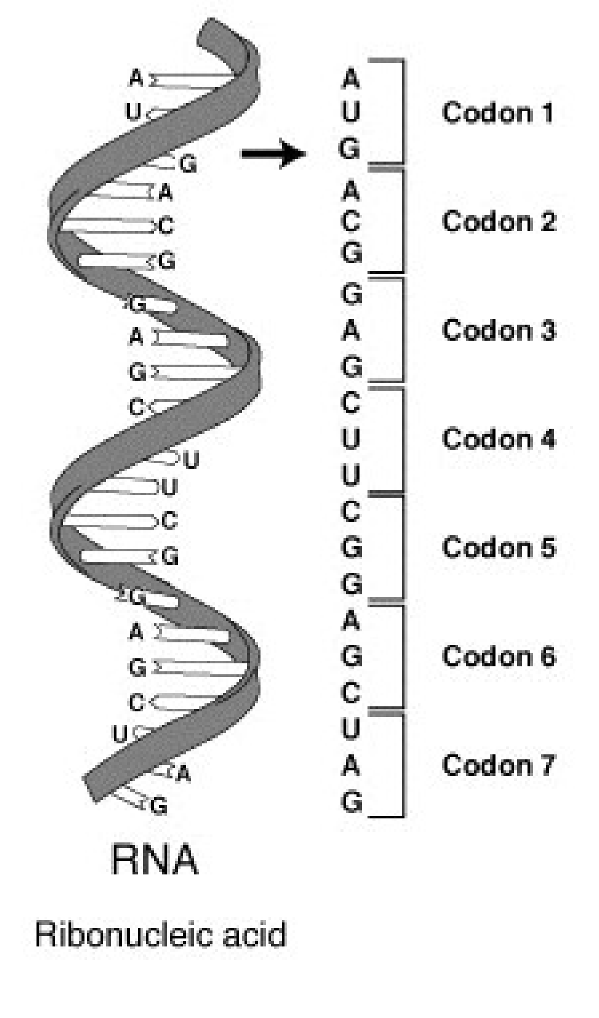
Compression of Genetic Coding Sequences

MohammadReza Ghodsi

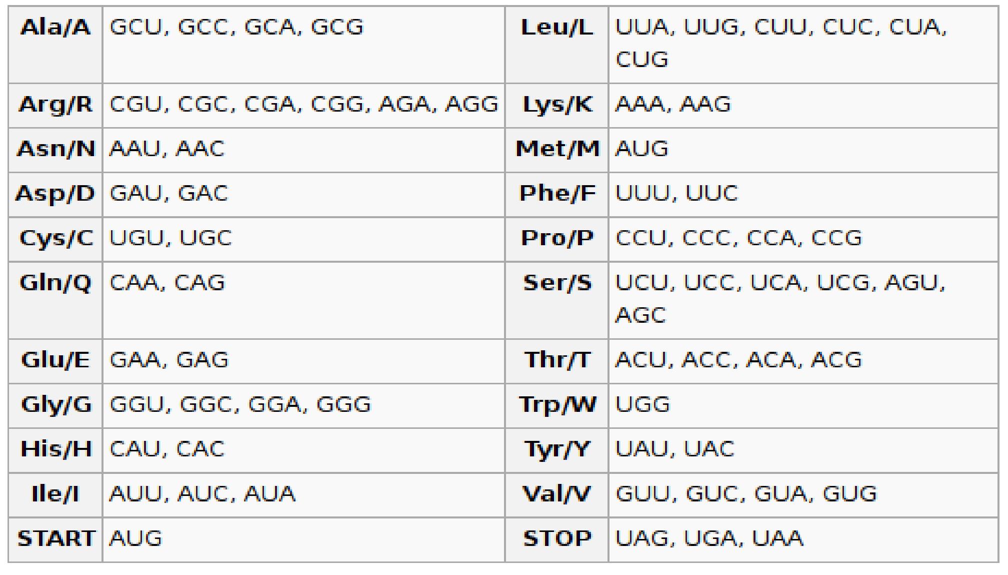
Genetic Code (Recap)

- The code defines a mapping between tri-nucleotide sequences called codons and amino acids.
- They begins with start codon (ATG), ends with a stop codon (TAG/TGA/TAA)



Genetic Code (Recap) - 2

- The genetic code has redundancy but no ambiguity. (There are 4^3=64 codons and only 20 amino acids)
- A position of a codon is said to be a *degenerate site* if different nucleotides at this position specify the same amino acid.



Ala/A	GCU, GCC, GCA, GCG	**Leu/L**	UUA, UUG, CUU, CUC, CUA, CUG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	**Lys/K**	AAA, AAG
Asn/N	AAU, AAC	**Met/M**	AUG
Asp/D	GAU, GAC	**Phe/F**	UUU, UUC
Cys/C	UGU, UGC	**Pro/P**	CCU, CCC, CCA, CCG
Gln/Q	CAA, CAG	**Ser/S**	UCU, UCC, UCA, UCG, AGU, AGC
Glu/E	GAA, GAG	**Thr/T**	ACU, ACC, ACA, ACG
Gly/G	GGU, GGC, GGA, GGG	**Trp/W**	UGG
His/H	CAU, CAC	**Tyr/Y**	UAU, UAC
Ile/I	AUU, AUC, AUA	**Val/V**	GUU, GUC, GUA, GUG
START	AUG	**STOP**	UAG, UGA, UAA

Lossless Data Compression

- Completely random data streams cannot be compressed.
- Many different algorithms exist that are designed either with a specific type of input data in mind or with specific assumptions about what kinds of redundancy the uncompressed data are likely to contain.
- I am planning to use Ziv-Lempel + Huffman

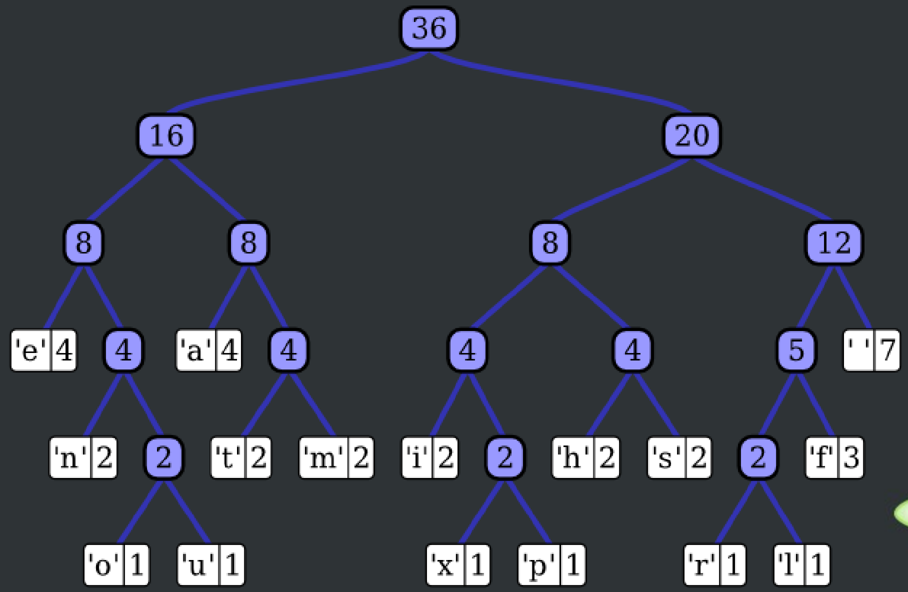
Lempel-Ziv LZ77 (Recap)

- Idea: Find the longest prefix of S[i..n] that is a substring of S[1..i-1]
- Compression can be done using suffix trees. In linear time and space. This is the method that I intend to use
- Most practical implementations use a sliding window algorithm to achive a more space efficient algorithm (online algorithms).

Huffman coding

- A variable-length code table for encoding a source symbol (Codons in our case) where the variable-length code table has been derived based on the frequency of each possible value of the source symbol.

Huffman tree generated from the exact frequencies in the sentence "this is an example of a huffman tree"



Hypothetical Application

- Folding@Home is a distributed computing project designed to perform computationally intensive simulations of protein folding.
- The client periodically connects to a server to retrieve "work units," which are packets of data upon which to perform calculations. Each completed work unit is then sent back to the server.