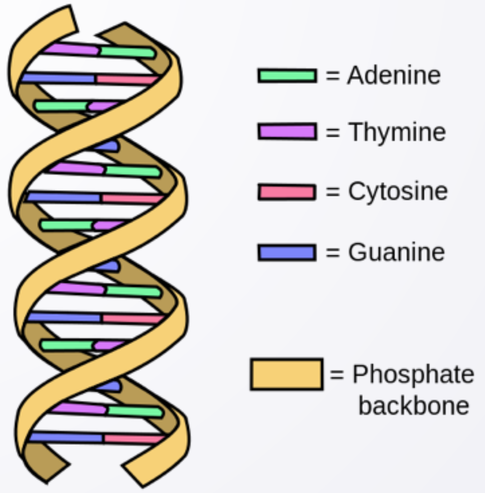
Finding a Gene in DNA

Developing an Algorithm

Duke
UNIVERSITY

Algorithms for Finding Genes

- Genes help make proteins; finding genes models searching in a string
  - Could search for tags on a web page
  - Could search for videos on YouTube
- Simple version of gene finding
  - String methods
  - Integer arithmetic
  - Seven Step Process



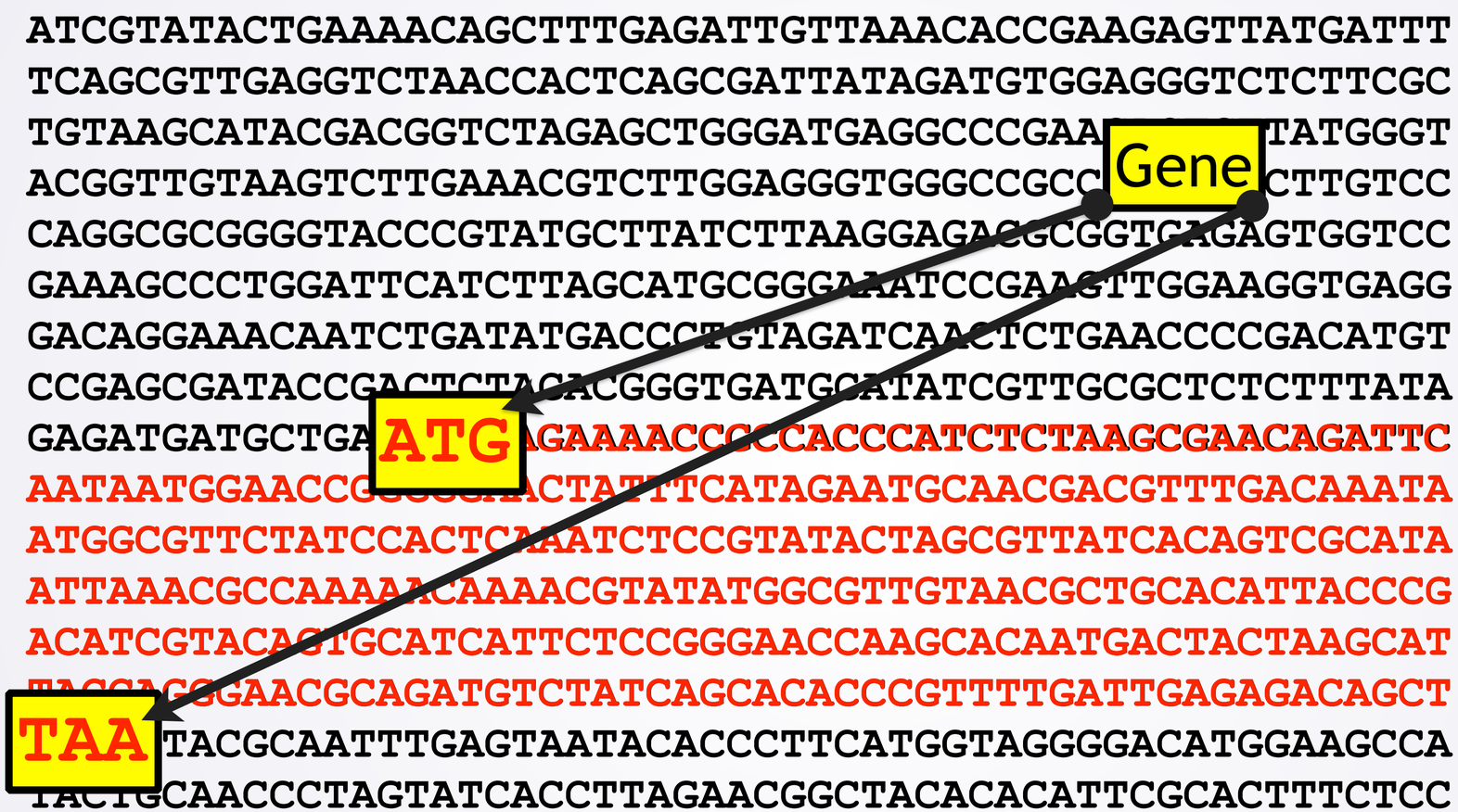
DNA

Needles in Haystacks—Genes in DNA

ATCGTATACTGAAAACAGCTTTGAGATTGTTAAACACCGAAGAGTTATGATTT
TCAGCGTTGAGGTCTAACCACTCAGCGATTATAGATGTGGAGGGTCTCTTCGC
TGTAAGCATACGACGGTCTAGAGCTGGGATGAGGCCCGAACACTGTTATGGGT
ACGGTTGTAAGTCTTGAAACGTCTTGGAGGGTGGGCCGCCCAAGTACTTGTCC
CAGGCGCGGGGTACCCGTATGCTTATCTTAAGGAGACGCGGTGAGAGTGGTCC
GAAAGCCCTGGATTCATCTTAGCATGCGGGAAATCCGAAGTTGGAAGGTGAGG
GACAGGAAACAATCTGATATGACCCTGTAGATCAACTCTGAACCCCGACATGT
CCGAGCGATACCGACTCTACACGGGTGATGCATATCGTTGCGCTCTCTTTATA
GAGATGATGCTGAATGGAAGAAAACCGCCACCCATCTCTAAGCGAACAGATTC
AATAATGGAACCGGCCGAACTATTTCATAGAATGCAACGACGTTTGACAAATA
ATGGCGTTCTATCCACTCAAATCTCCGTATACTAGCGTTATCACAGTCGCATA
ATTAAACGCCAAAAACAAAACGTATATGGCGTTGTAACGCTGCACATTACCCG
ACATCGTACAGTGCATCATTCTCCGGGAACCAAGCACAATGACTACTAAGCAT
TACCAGGGAACGCAGATGTCTATCAGCACACCCGTTTTGATTGAGAGACAGCT
TAATGTACGCAATTTGAGTAATACACCCTTCATGGTAGGGGACATGGAAGCCA
TACTGCAACCCTAGTATCACCTTAGAACGGCTACACACATTCGCACTTTCTCC
TACGCGGCAACTTGTCGACGTTCTTGAGACGCTGTCGAGTGTTCCCAGCTAGC
CTGGTCGGGACAATTATGACAACGGCAGTCCAGCATCATATGCCGCGAGCCGC
ACATTGGCTCCGTGTCACGCGCGATTGCTAGATCCGGGCA

Duke
UNIVERSITY

Needles in Haystacks—Genes in DNA



String Method Overview

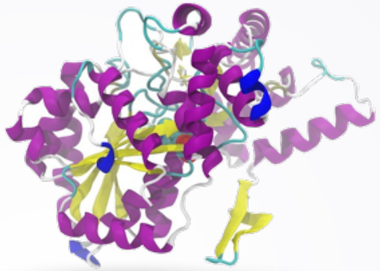
- `String s = "dukeprogramming";`

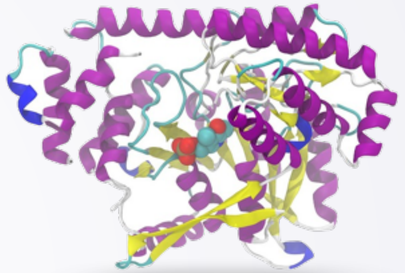
```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4
```

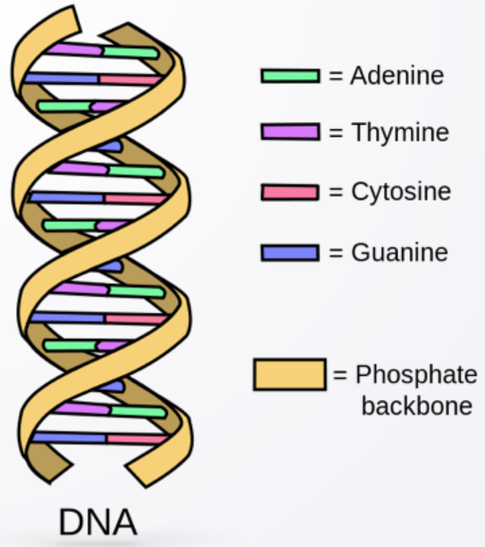
method	value
`s.length()`	15
`s.indexOf("program")`	4
`s.indexOf("g")`	7
`s.indexOf("f")`	-1
`s.indexOf("g",8)`	14
`s.startsWith("duke")`	true
`s.endsWith("king")`	false
`s.substring(4,7)`	"pro"
`s.toUpperCase()`	DUKEPROGRAMMING

Problem and Algorithm

- Write a method that returns the index of a gene in a strand of DNA
  - Start codon "ATG"
  - Stop codon "TGA"
    - Later also "TAA", "TAG"
  - Separated by 3N nucleotides
- ..ATGCGATACGCTTGA..
- ..ATGCGATACGTGA..

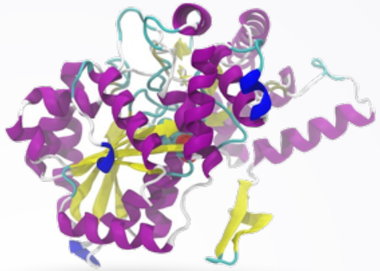


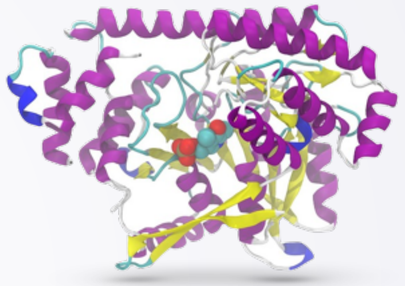


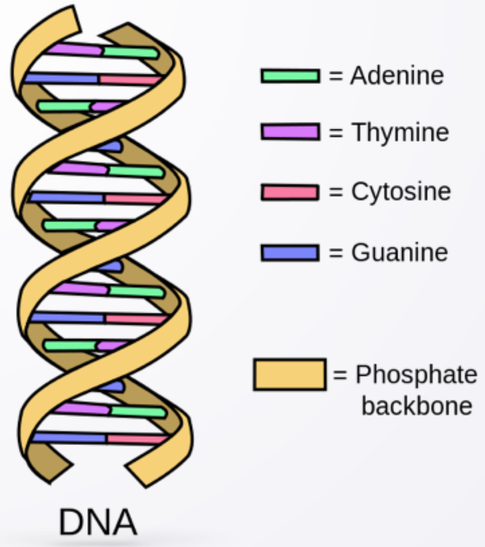


Problem and Algorithm

- Write a method that returns the index of a gene in a strand of DNA
  - Start codon "ATG"
  - Stop codon "TGA"
    - Later also "TAA", "TAG"
  - Separated by 3N nucleotides
- ..ATGCGATACGCTTGA..
- ..ATGCGATACGTGA..



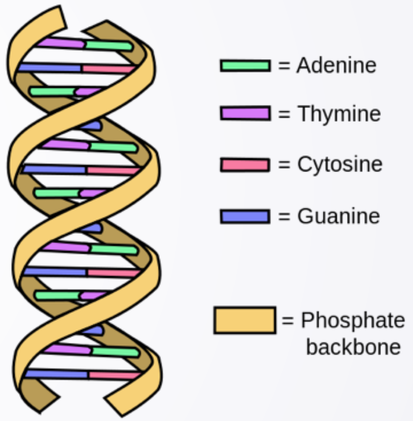




First Four Steps

1. Work Example By Hand
2. Write Down What You Did
3. Find Patterns
4. Check By Hand

- **CTACGATGCCTGATGA**
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
  - Use specific strand to reason
- Find index of start codon ATG: 5
- Stop codon TGA after 7: 13
- Multiple of 3? NO
- Steps change with different DNA strands?



Duke UNIVERSITY