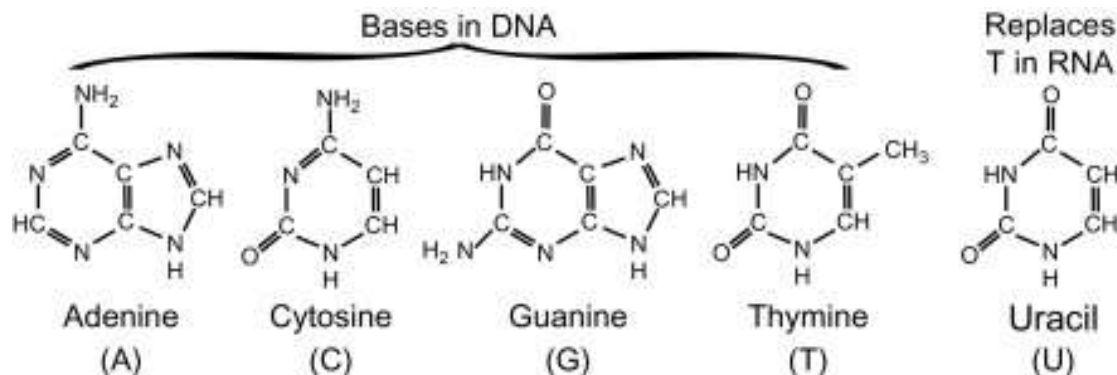


Genes and Sequences

Rupesh Nasre.
rupesh@cse.iitm.ac.in

Gene

- Four nucleotide bases in DNA: Adenine, Cytosine, Guanine, and Thymine.
- These bases pair specifically (A-T, C-G) to form the genetic code within the DNA double helix.
- Thus, a gene is simply a substring of A, C, G, T.



Gene Operations

- Match sequences (`strcmp`)
- Checking if a gene sequence exists (`strstr`)
- Finding subsequences separated by a delimiter (`strtok`)
 - How does `strtok` work?
- Concatenate two genes (`strcat`)
- Generate a random sequence
- Generalizations
 - Arbitrary set of characters
 - `int`, `struct`, arbitrary types?

Gene Operations in Shell

- Match sequences (`strcmp`, `grep`, `diff`)
- Checking if a gene sequence exists (`strstr`, `grep`)
- Finding subsequences separated by a delimiter (`strtok`, `grep`)
 - Replacing them will need `sed`
- Concatenate two genes (`strcat`, `>>`)
- Other types: Can be supported, but shell is best suited for the string type.

```
: diff gene1 gene2
: diff gene1 gene3
1c1
< TGCTCTGACCGTGTTGAGAA
---
> TGATCTGACCGTGTTGAGAA
: wc -l gene4
7
: grep ACC gene4 | wc -l
5
: grep ACC gene4
TGCTCTGACCGTGTTGAGAA
TGATCTGACCGTGTTGAGACC
TGCTCGGACCGTGTTGAGAA
TGCTCTGACCGTGTTGAGAA
TGCTCTGACCGTRACC
:
```

```
: # replace ACC by GGTTCC
: sed 's/ACC/GGTTCC/' gene4
TGCTCTGGGTTCCGTGTTGAGAA
TGATCTGGGTTCCGTGTTGAGACC
TGCTCGGGGTTCCGTGTTGAGAA
TGCTCTGACAGTGTTGAGAAGTC
TGCTCTGACGTTGTTGAGAAAAA
TGCTCTGGGTTCCGTGTTGAGAA
: sed 's/ACC/GGTTCC/g' gene4
TGCTCTGGGTTCCGTGTTGAGAA
TGATCTGGGTTCCGTGTTGAGGGTTCC
TGCTCGGGGTTCCGTGTTGAGAA
TGCTCTGACAGTGTTGAGAAGTC
TGCTCTGACGTTGTTGAGAAAAA
TGCTCTGGGTTCCGTGTTGAGAA
TGCTCTGGGTTCCGTRGGTTCC
: # gene5 = gene3 + gene4
: cat gene3 gene4 >gene5
OR
: cp gene3 gene5; cat gene4 >>gene5
```

Edit Distance

- Given two genes, edit distance is the minimum number of changes (mutations) required to convert one into another.
 - Changes are: update, delete, add (of a single char)
 - $\text{dist}(\text{ACGT}, \text{CCGT}) == 1$
 - $\text{dist}(\text{ACGT}, \text{AGT}) == 1$
 - $\text{dist}(\text{ACGT}, \text{AG}) == 2$
 - $\text{dist}(\text{AGGACTGA}, \text{GATTCAGGCCCC}) == ?$
- How do you solve this problem?

In this course...

Application	Programming	Tools
Puzzle of grass, goat, lion	Pointers, structures	Editor, visualization
Data maintenance	File I/O	I/O redirection, make
Item grouping	Linked Lists	Debugging
Gene sequencing	Recursion, strings	Memory profiling, ulimit
Closest match	Matrices	Compute profiling, plotting
Keep the recent	Arrays	Version control
Cryptography	Strings, bitwise operators	Library (ncurses)
Unit testing	Command-line	Testing
Web-page creation, analysis	HTML, Strings	wget, shell scripting
Game	Graphics	Library (graphics)
Basics	Python	--

Our goal: *to have fun with automation*