# NONMETRIC METHODS Recognition with Strings

Pattern Recognition

CSE 555/655

#### Attribute Lists

- Nominal Data
  - No natural notion of similarity or ordering
  - E.g., a fruit =(red,shiny,sweet,small)
- Distance between vectors cannot be measured
  - Nearest-neighbor uses distance
  - Neural network uses similarity of outputs

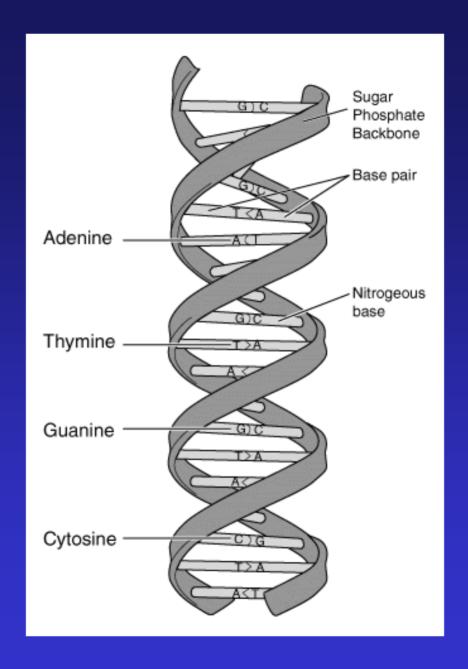
# Strings

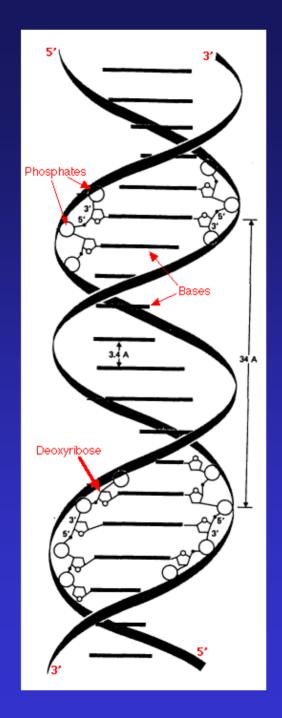
- Patterns are ordered sequences, or variable length strings of discrete items
- Examples
  - Sequence of letters in an English word
  - In DNA, bases in a gene sequence
  - In JOCR, brush strokes recognized by neural network, then string matching for OCR
- Move away from pdfs and metrics to syntactic methods

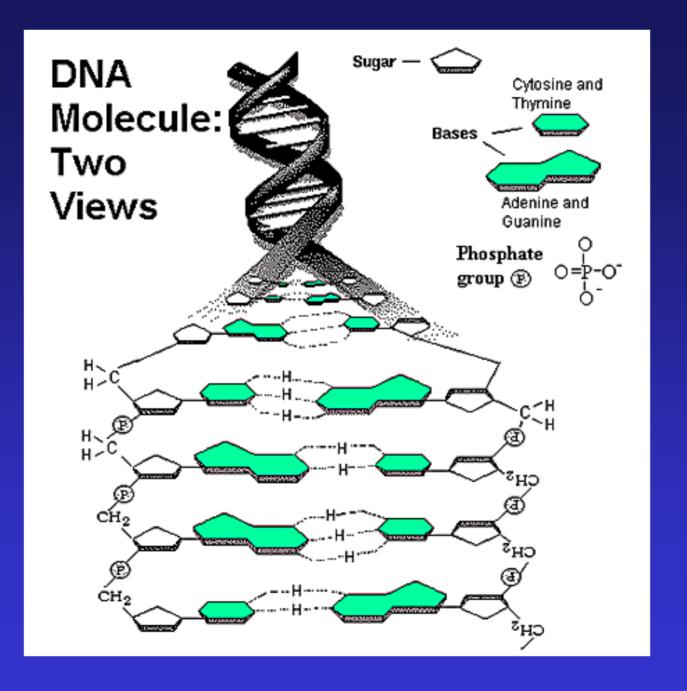
#### Gene Sequence

- DNA bases in gene sequence
- E.g., AGCTTCGAATC
- Letters stand for nucleic acids
  - A: adenine
  - G: guanine
  - C: cytosine
  - T: thymine











### String Pattern Classifiction

- String elements are nominal
  - Characters, letters, or symbols
- No obvious notion of distance between strings
- Strings need not be of same length
- Strings are not vectors, yet we will use notation, x = AGCTTC

# String Terminology

- Patterns, Strings, templates, words
- A long string is denoted text
- Any continuous string that is part of x is a substring, segment, or factor of x
  - E.g., "GCT" is a factor of "AGCTTC"

# String Recognition Problems

- Many computational problems with strings
- Problems of greatest importance:
- 1. String Matching: Given x and text, determine whether x is a factor of text and if so, where it appears
- 2. Edit Distance: Given two strings x and y compute minimum number of basic operations—insertions, deletions and substitutions—to transform x to y

### String Problems

3. String Matching with errors: Given **x** and *text*, find the locations in *text* where the cost or distance to any factor of **x** is minimal

### String Problems

4. String Matching with "Don't-care symbol": Same as basic string matching, but with special symbol-- \* which can match any other symbol., e.g.,

Given DNA segment x = AGCCG\*\*\*\*\*GACTG

Motifs:

Important for coding protein

Inert: no function

Is *text* in the class of sequences that could yield the particular protein