Dynamic Programming
Edit Distance

(Class 17)

Dynamic Programming

- Dynamic programming is essentially *"recursion without repetition"*.
- Developing a dynamic programming algorithm generally involves two separate steps:
  - **Formulate problem recursively:** Write down a formula for the whole problem as a simple combination of answers to smaller subproblems.
  - **Build solution to recurrence from bottom up:** Write an algorithm that starts with base cases and works its way up to the final solution.

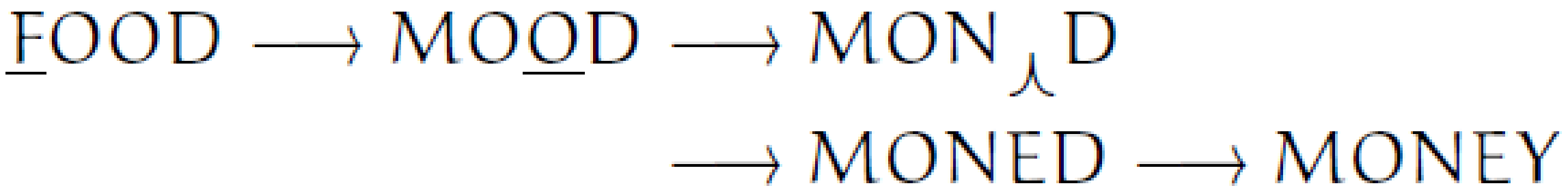
- Dynamic programming algorithms need to store the results of intermediate subproblems.
- This is often (but not always) done with some kind of table.
- We will now cover a number of examples of problems in which the solution is based on dynamic programming strategy.

Edit Distance

- Edit distance algorithm is a very good example of dynamic programming.
- The words “computer” and “commuter” are very similar, and a change of just one letter, will change the first word into the second.
- The word “sport” can be changed into “sort” by the deletion of the ‘p’, or equivalently, ‘sort’ can be changed into ‘sport’ by the insertion of ‘p’.

- The edit distance of two strings, s1 and s2, is defined as the minimum number of point mutations required to change s1 into s2.
- Where a point mutation is one of:
  - change a letter,
  - insert a letter or
  - delete a letter

- For example, the edit distance between FOOD and MONEY is at most 4:



- There are numerous applications of the Edit Distance algorithm.

Edit Distance: Applications

1. Spelling Correction
2. Plagiarism Detection
3. Computational Molecular Biology
4. Speech Recognition
5. Longest Common Subsequence (LCS)

1. Spelling Correction

- If a text contains a word that is not in the dictionary, a ‘close’ word, i.e., one with a small edit distance, may be suggested as a correction.
- Most word processing applications, such as Microsoft Word, have spelling checking and correction facility.
- When Word, for example, finds an incorrectly spelled word, it makes suggestions of possible replacements.

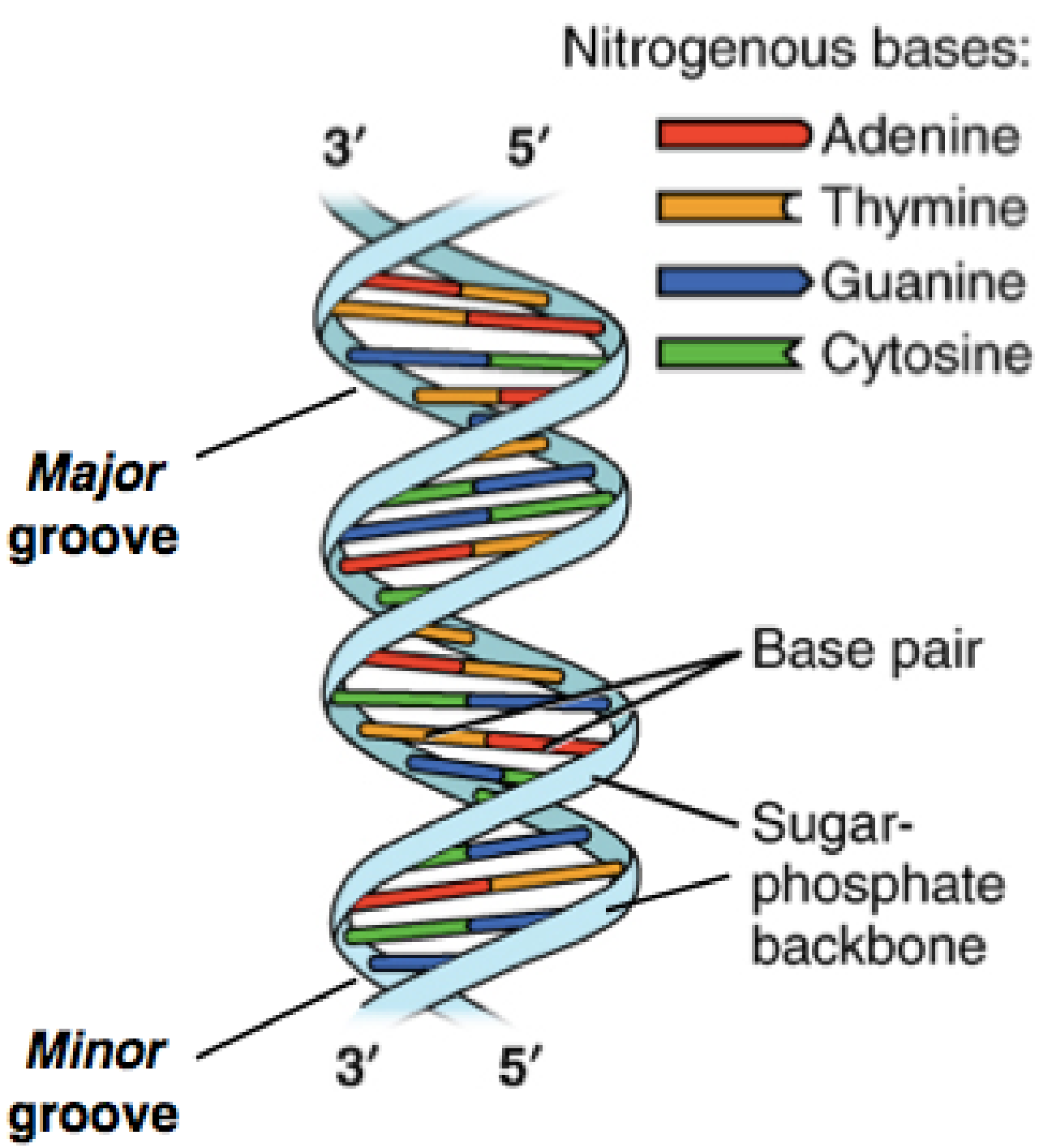
2. Plagiarism Detection

- If someone copies, say, a C program and makes a few changes here and there, for example, change variable names, add a comment of two, the edit distance between the source and copy may be small.
- The edit distance provides an indication of similarity that might be too close in some situations.

3. Computational Molecular Biology

- The monomer units of DNA are nucleotides, and the polymer is known as a “polynucleotide.”
- Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group.

- There are four different types of nucleotides found in DNA, differing only in the nitrogenous base.
- The four nucleotides are given one letter abbreviations as shorthand for the four bases.
  - A-adenine
  - G-guanine
  - C-cytosine
  - T-thymine


Nitrogenous bases:
Adenine
Thymine
Guanine
Cytosine
3′
5′
Major groove
Base pair
Sugar-phosphate backbone
Minor groove
3′
5′

Translation MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFE
SFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPE
NFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH

SEQUENCE (5′ to 3′ sequence of the coding strand)

```
   1 ccctgtggag ccacacccta gggttggcca atctactccc aggagcaggg agggcaggag
  61 ccagggctgg gcataaaagt cagggcagag ccatctattg cttacatttg cttctgacac
 121 aactgtgttc actagcaacc tcaaacagac accatggtgc acctgactcc tgaggagaag
 181 tctgccgtta ctgccctgtg gggcaaggtg aacgtggatg aagttggtgg tgaggccctg
 241 ggcaggttgg tatcaaggtt acaagacagg tttaaggaga ccaatagaaa ctgggcatgt
 301 ggagacagag aagactcttg ggtttctgat aggcactgac tctctctgcc tattggtcta
 361 ttttcccacc cttaggctgc tggtggtcta cccttggacc cagaggttct ttgagtcctt
 421 tggggatctg tccactcctg atgctgttat gggcaaccct aaggtgaagg ctcatggcaa
 481 gaaagtgctc ggtgcctta gtgatggcct ggctcacctg gacaacctca agggcacctt
 541 tgccacactg agtgagctgc actgtgacaa gctgcacgtg gatcctgaga acttcagggt
 601 gagtctatgg gacccttgat gttttctttc cccttctttt ctatggttaa gttcatgtca
 661 taggaagggg agaagtaaca gggtacagtt tagaatggga aacagacgaa tgattgcatc
 721 agtgtggaag tctcaggatc gttttagttt cttttatttg ctgttcataa caattgtttt
 781 cttttgttta attcttgctt tctttttttt tcttctccgc aatttttact attatactta
 841 atgccttaac attgtgtata acaaaaggaa atatctctga gatacattaa gtaacttaaa
 901 aaaaaacttt acacagtctg cctagtacat tactatttgg aatatatgtg tgcttatttg
 961 catattcata atctccctac tttatttttct tttatttttа attgatacat aatcattata
1021 catatttatg ggttaaagtg taatgtttta atatgtgtac acatattgac caaatcaggg
1081 taattttgca tttgtaattt taaaaaatgc tttcttcttt taatatactt ttttgtttat
1141 cttatttcta atactttccc taatctcttt ctttcagggc aataatgata caatgtatca
1201 tgcctctttg caccattcta aagaataaca gtgataattt ctgggttaag gcaatagcaa
1261 tatttctgca tataaatatt tctgcatata aattgtaact gatgtaagag gtttcatatt
1321 gctaatagca gctacaatcc agctaccatt ctgcttttat tttatggttg ggataaggct
1381 ggattattct gagtccaagc taggcccttt tgctaatcat gttcatacct cttatcttcc
1441 tcccacagct cctgggcaac gtgctggtct gtgtgctggc ccatcacttt ggcaaagaat
1501 tcaccccacc agtgcaggct gcctatcaga aagtggtggc tggtgtggct aatgccctgg
1561 cccacaagta tcactaagct cgctttcttg ctgtccaatt tctattaaag gttcctttgt
1621 tccctaagtc caactactaa actgggggat attatgaagg gccttgagca tctggattct
1681 gcctaataaa aaacatttat tttcattgca atgatgtatt taaattattt ctgaatattt
1741 tactaaaaag ggaatgtggg aggtcagtgc atttaaaaca taaagaaatg atgagctgtt
1801 caaaccttgg gaaaatacac tatatcttaa actccatgaa agaaggtgag gctgcaacca
1861 gctaatgcac attggcaaca gcccctgatg cctatgcctt attcatccct cagaaaagga
```

- This image was an example of the DNA coding sample.
- If we are given two such samples and we have to find the edit distance between these two DNA coding samples.
- The edit distance like algorithms is used to compute a distance between DNA sequences (strings over A,C,G,T, or protein sequences (over an alphabet of 20 amino acids), for various purposes, e.g.:
  - to find genes or proteins that may have shared functions or properties.
  - to infer family relationships and evolutionary trees over different organisms.

4. Speech Recognition

- Algorithms similar to those for the edit-distance problem are used in some speech recognition systems.
- Find a close match between a new utterance and one in a library of classified utterances.

5. Longest Common Subsequence (LCS)

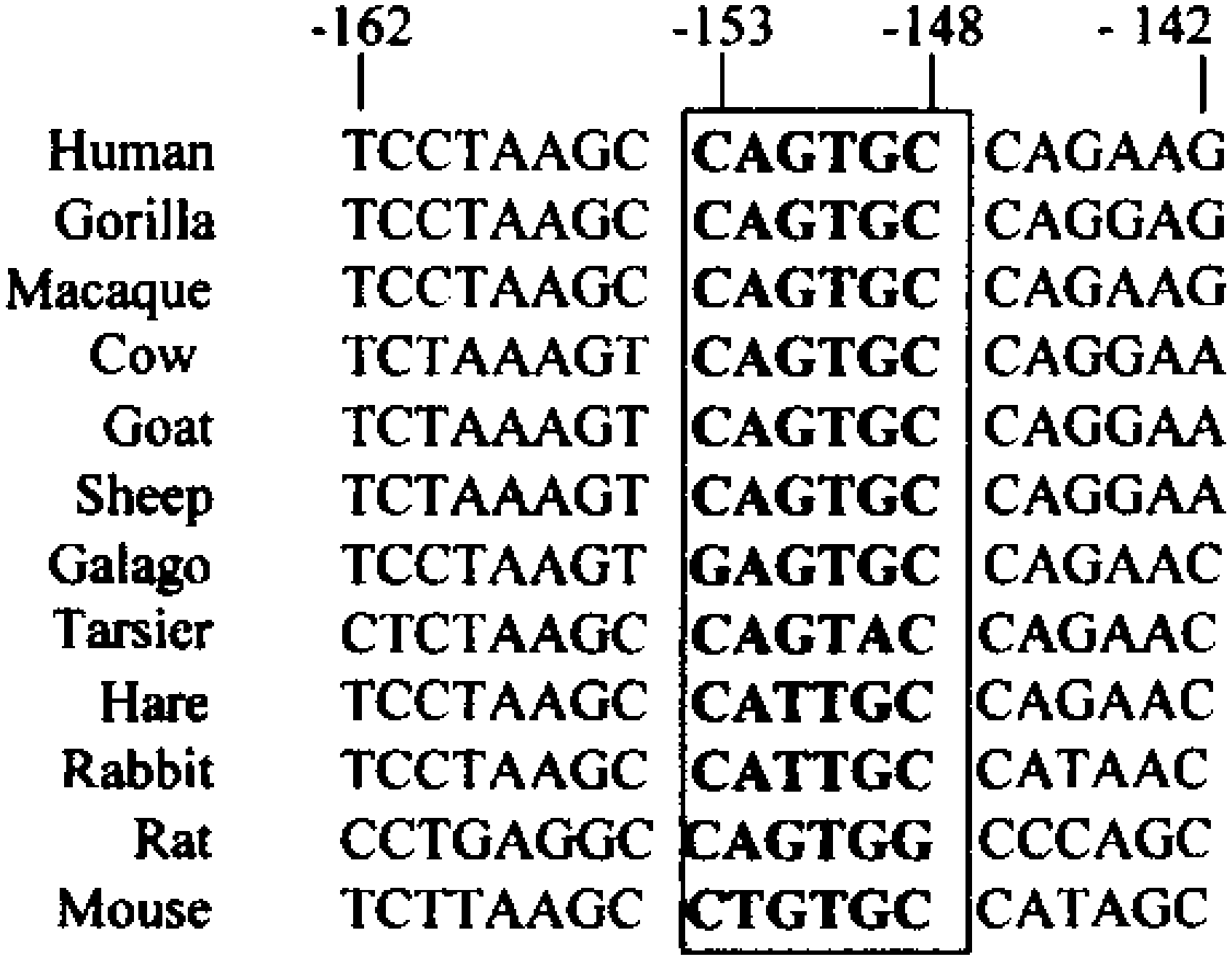
- This is another example where we don’t change the strings but only want to find the common subsequence.
- It is finding the similarity between two sequences.
- For example, we have these two sequences of nucleotides:

$$S1 = ACCGGTCGAGTGCGCGGAAGCCGGCCGAA$$
$$S2 = GTCGTTCGGAATGCCGTTGCTCTGTAAA$$

- We want to find two things:
  - Is whole S2 is present in S2?
  - Find the subsequence that is common in both S1 and S2.

- The common sequence in both S1 and S2 is:

$$S3 = GTCGTCGGAAGCCGGCCGAA$$


-162
-153
-148
- 142
Human TCCTAAGC CAGTGC CAGAAG
Gorilla TCCTAAGC CAGTGC CAGGAG
Macaque TCCTAAGC CAGTGC CAGAAG
Cow TCTAAAGT CAGTGC CAGGAA
Goat TCTAAAGT CAGTGC CAGGAA
Sheep TCTAAAGT CAGTGC CAGGAA
Galago TCCTAAGT GAGTGC CAGAAC
Tarsier CTCTAAGC CAGTAC CAGAAC
Hare TCCTAAGC CATTGC CAGAAC
Rabbit TCCTAAGC CATTGC CATAAC
Rat CCTGAGGC CAGTGG CCCAGC
Mouse TCTTAAGC CTGTGC CATAGC

Edit Distance Dynamic Programming Formulation

- How can we know if a problem can be solved using dynamic programming?
- We cannot use any formula or mathematical technique to understand that a particular problem can be solved using dynamic programming.
- It is like an art and gained with experience.

Edit Distance Algorithm

- A better way to display this editing process is to place the words above the other:

S	*D*	*I*	*M*	*D*	*M*
M	A	_	T	H	S
A	_	R	T	_	S

- The first word has a gap for every insertion (I) and the second word has a gap for every deletion (D).
- Columns with two different characters correspond to substitutions (S).
- Matches or Maintain (M) do not count.

Edit Transcript

- It is defined as a string over the alphabet M, S, I, D that describes a transformation of one string into another.
- For example:

$$\begin{array}{cccccccc} S & & D & & I & M & D & M \\ 1+ & & 1+ & & 1+ & 0+ & 1+ & 0+ \end{array} = 4$$

- In general, it is not easy to determine the optimal edit distance.
- For example, the distance between ALGORITHM and ALTRUISTIC is at most 6.

A	L	G	O	R	_	I	_	T	H	M
A	L	_	T	R	U	I	S	T	I	C