

Dishan's Problem. A squirrel who is a biologist¹ wants to check whether a DNA sequence matches a pattern. The sequence contains only A, C, G, and T. The pattern can also contain two special symbols:

1. ? matches exactly one DNA character.
2. * matches any number of DNA characters, including zero.

For example,

Sequence	Pattern	Matches?
ACGTT	A?G*	Yes
ACT	A*T	Yes
AT	A?T	No
ACGT	*CG	No

Given a sequence S of length n and a pattern P of length m , determine whether they match in $O(nm)$ time.

Solution. We implement the following DP algorithm:

$\text{dp}(i, j)$: outputs true iff $S[1, i]$ matches $P[1, j]$.

$\text{dp}(i, j)$

1. If $i > 0$ and $j = 0$: output False
2. If $i = 0$ and $j = 0$: output True
3. Switch $P[j]$:
 - (a) Case $P[j] = ?$: Output $\text{dp}(i - 1, j - 1)$.
 - (b) Case $P[j] = *$: Output $\text{dp}(i - 1, j)$ or $\text{dp}(i, j - 1)$.
 - (c) Default: `//P[j] is ACTG`
Output $P[j] = S[i]$ and $\text{dp}(i - 1, j - 1)$.

¹This squirrel also happens to “[sic] be really fat” according to Dishan.

Dishan's Followup Problem. The same squirrel who is a biologist has found that none of its DNA sequences match the pattern. The biologist squirrel is now interested in applying gene editing software to make the sequences match the pattern.

Given a DNA sequence S of length n and a pattern P of length m , find the minimum number of edits needed to transform S into a string that matches P in $O(nm)$ time.²

Solution. We implement the following DP algorithm:

$\text{dp}(i, j)$: outputs minimum number of edits to make $S[1, i]$ matches $P[1, j]$.

$\text{dp}(i, j)$

1. If $i > 0$ and $j = 0$: output i
2. If $i = 0$ and $j = 0$: output 0
3. Switch $P[j]$:
 - (a) Case $P[j] = ?$: Output $\min\{\text{dp}(i-1, j-1), 1+\text{dp}(i-1, j), 1+\text{dp}(i, j-1)\}$.
 - (b) Case $P[j] = *$: Output $\text{dp}(i-1, j)$ or $\text{dp}(i, j-1)$.
 - (c) Default: *//P[j] is ACTG*
 - If $P[j] = S[i]$: Output $\min\{\text{dp}(i-1, j-1), 1+\text{dp}(i-1, j), 1+\text{dp}(i, j-1)\}$.
 - Else: Output $\min\{1 + \text{dp}(i-1, j-1), 1 + \text{dp}(i-1, j), 1 + \text{dp}(i, j-1)\}$.

²Each edit to S may either (1) insert one DNA character, (2) delete one DNA character, or (3) substitute one DNA character with another. As before, $?$ matches exactly one DNA character, and $*$ matches any number of DNA characters, including zero. The pattern must match the entire resulting sequence.