Cairo University
Faculty of Computers and Artificial Intelligence
Computer Science Department

Biological sequence analysis
BIO456

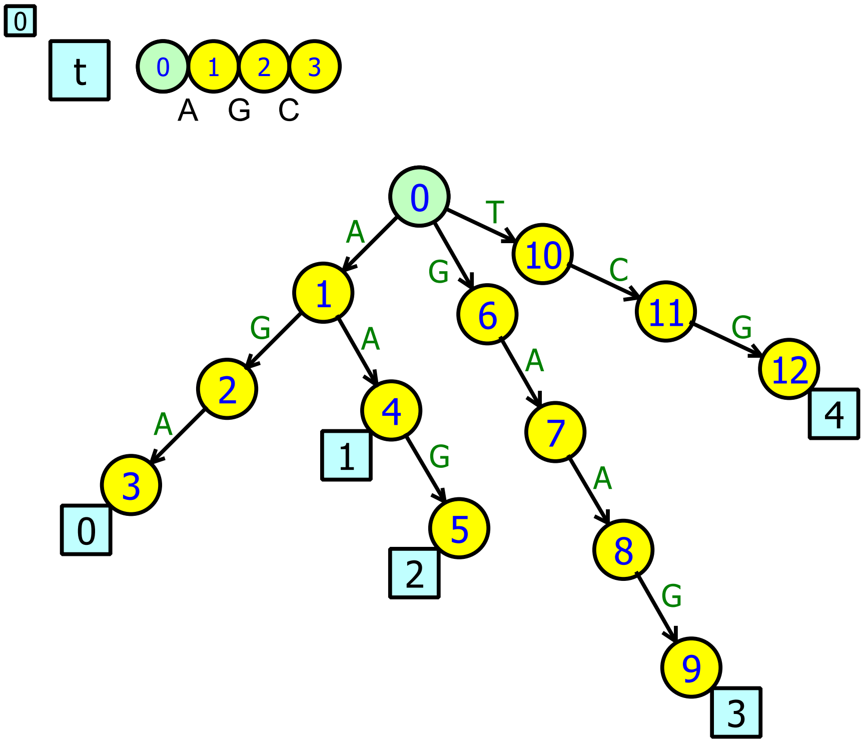
Approximate matching

Dr. Amin Allam

1 Hamming distance

In several cases, we are given an input pattern and we need to find similar patterns from a database of patterns. If we are not able to find exact-matching patterns, we need to find approximate-matching patterns. One measure of approximate matching is hamming distance: Two patterns have a hamming distance $= d$ if they have equal number of characters and a character-by-character comparison results in d unequal characters. For example, the hamming distance between `AGC` and `ACG` is d=2, while the hamming distance between `AGC` and `TGC` is d=1. The hamming distance between `AGC` and `AG` is not defined since they have different lengths.

First, we build a trie of the database patterns, and then we use recursion to backtrack over all possibilities as shown in the following example where the input pattern is `AGC` and the database patterns are { `AGA`, `AA`, `AAG`, `GAAG`, `TCG` } and we need to find all patterns with hamming distance≤ 2:



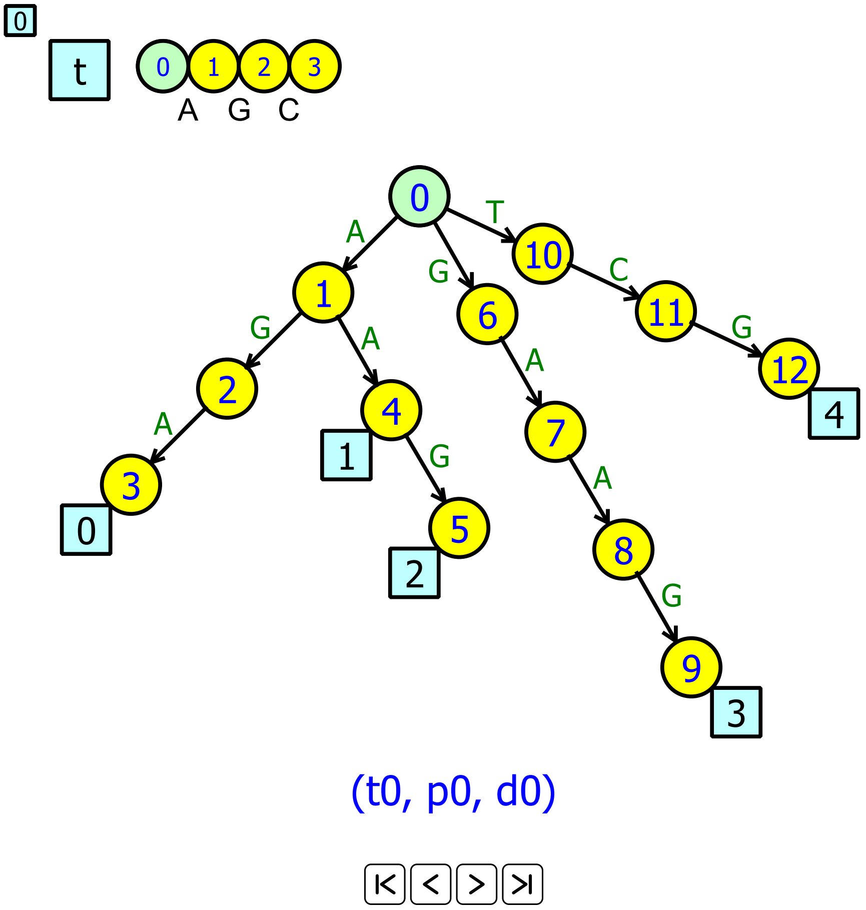
(t0, p0, d0)

|< < > >|

2 Edit distance

Another measure of approximate matching is edit distance: Two patterns have an edit distance $= d$ if the minimum number of insertion, deletion, and/or substitution operations required to convert one pattern to the other one $= d$. For example, the edit distance between AGC and AC is d=1 (One deletion operation of G from AGC will convert it to AC), while the edit distance between AGC and GCT is d=2 (Deletion of A from AGC then insertion of T) and the edit distance between AGC and GGC is d=1 (Substitution of the first character).

We can also use recursion to backtrack over all possibilities as shown in the following example where the input pattern is AGC and the database patterns are $\{$ AGA, AA, AAG, GAAG, TCG $\}$ and we need to find all patterns with edit distance≤ 1:


0
t
0
1
2
3
A G C
0
A
T
1
G
6
10
C
11
G
12
4
G
A
2
4
1
A
A
3
0
7
G
5
2
A
8
G
9
3
(t0, p0, d0)
|<
<
>
>|