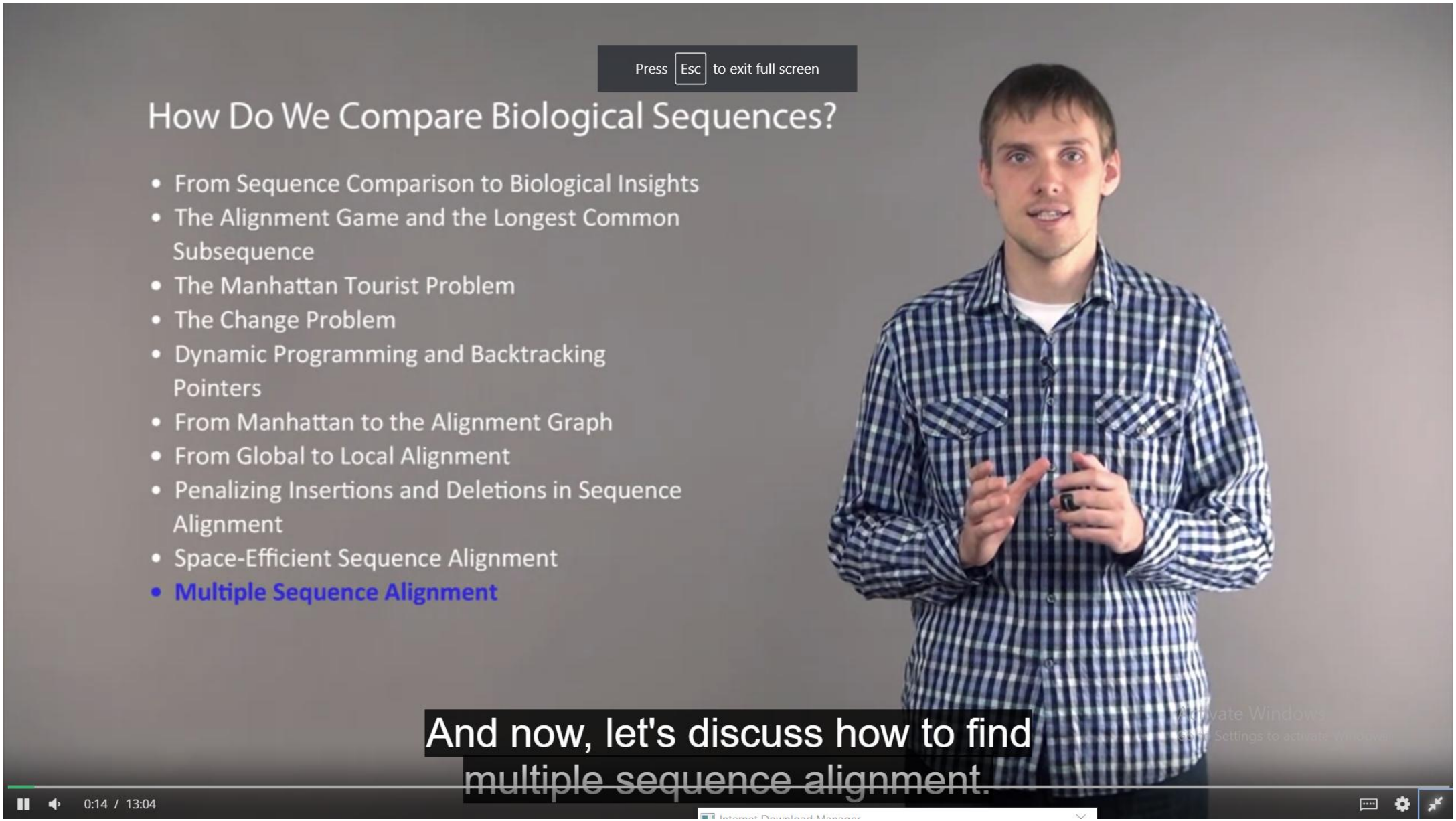

Press Esc to exit full screen
How Do We Compare Biological Sequences?
From Sequence Comparison to Biological Insights
The Alignment Game and the Longest Common Subsequence
The Manhattan Tourist Problem
The Change Problem
Dynamic Programming and Backtracking Pointers
From Manhattan to the Alignment Graph
From Global to Local Alignment
Penalizing Insertions and Deletions in Sequence Alignment
Space-Efficient Sequence Alignment
Multiple Sequence Alignment
And now, let's discuss how to find multiple sequence alignment.
0:14 / 13:04

From Pairwise to Multiple Alignment

A faint (and statistically borderline significant) similarity between two sequences becomes significant if it is present in many other sequences.

Multiple alignments can reveal subtle similarities that pairwise alignments fail to reveal.

Statistically significant similarities between two sequences becomes

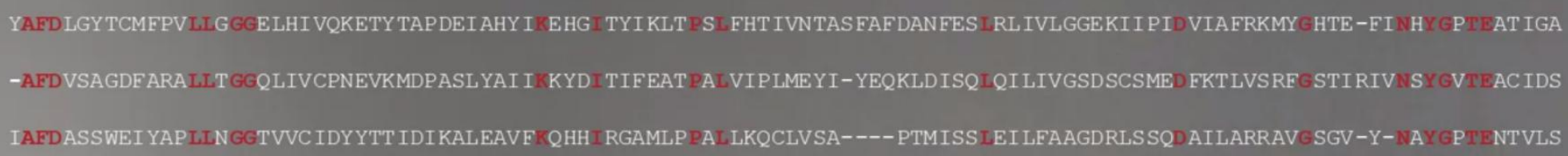
From Pairwise to Multiple Alignment

A faint (and statistically borderline significant) similarity between two sequences becomes significant if it is present in many other sequences.

Multiple alignments can reveal subtle similarities that pairwise alignments fail to reveal.

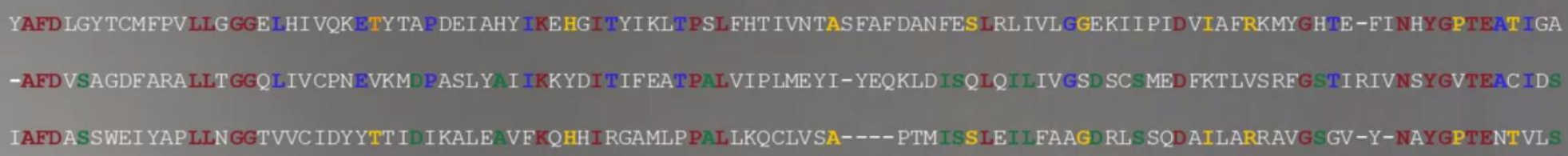
species and you find very conservative
domain in this species,

Alignment of Three A-domains

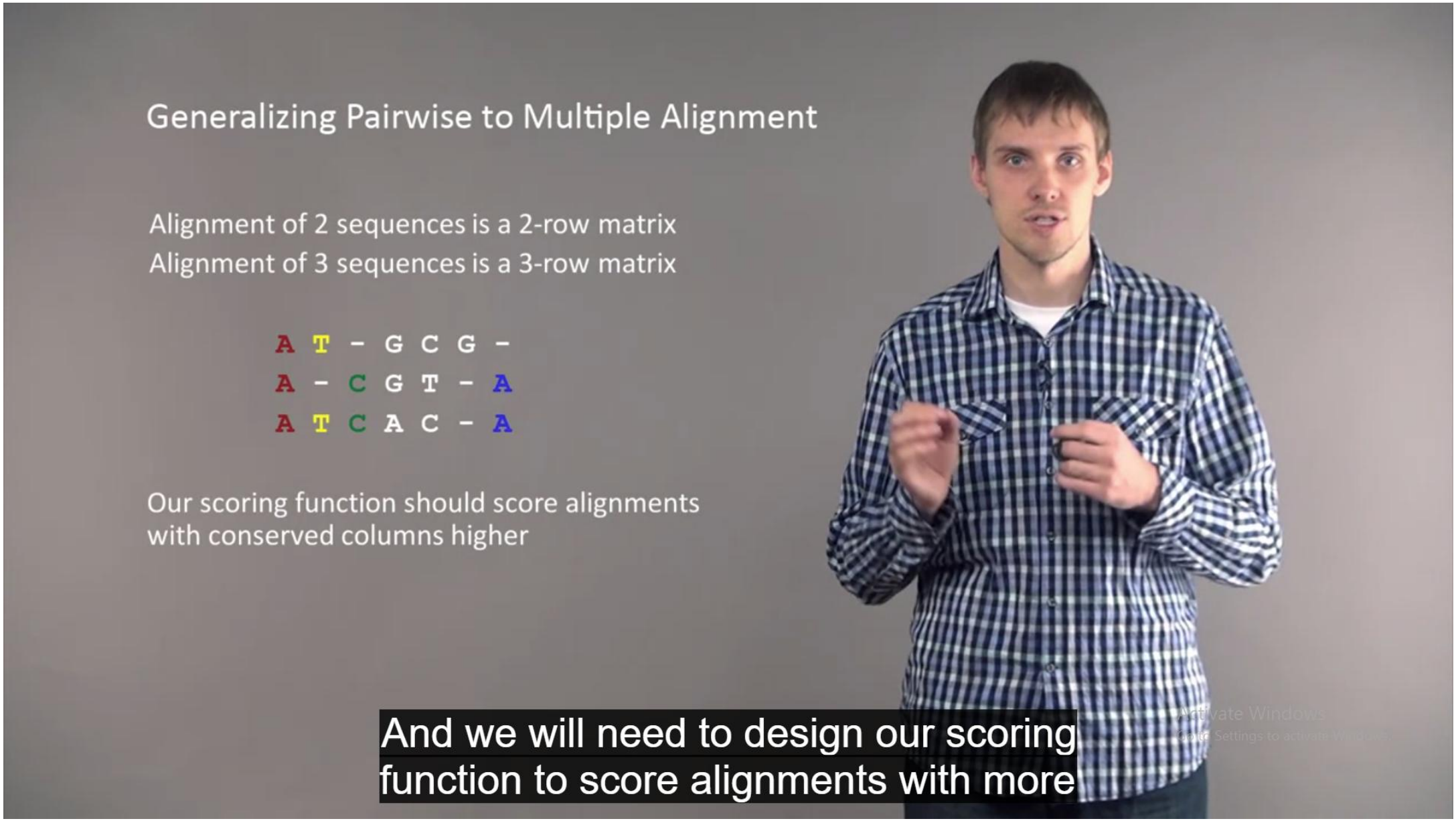


For example, when you align three A-domains,

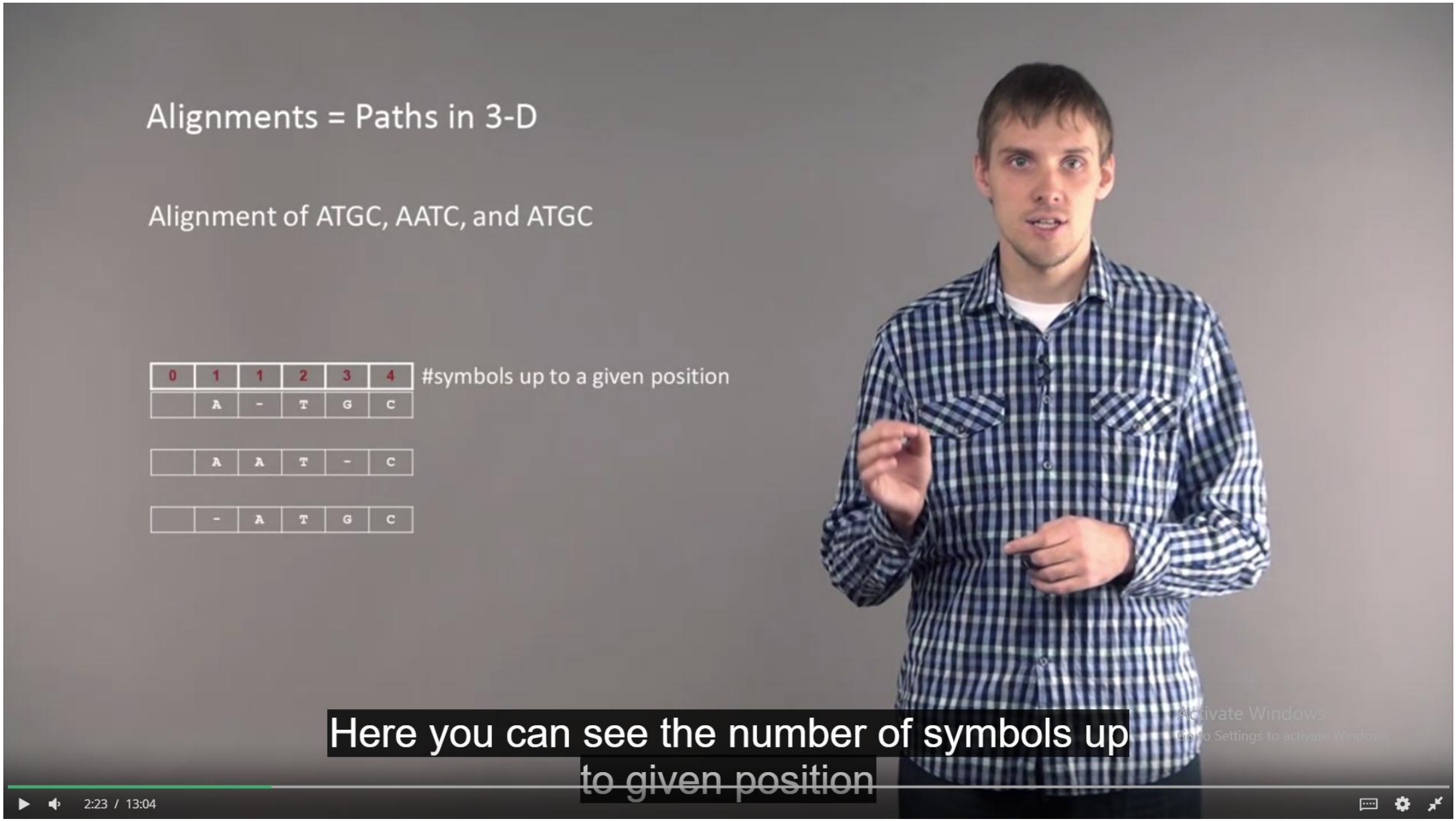
Alignment of Three A-domains

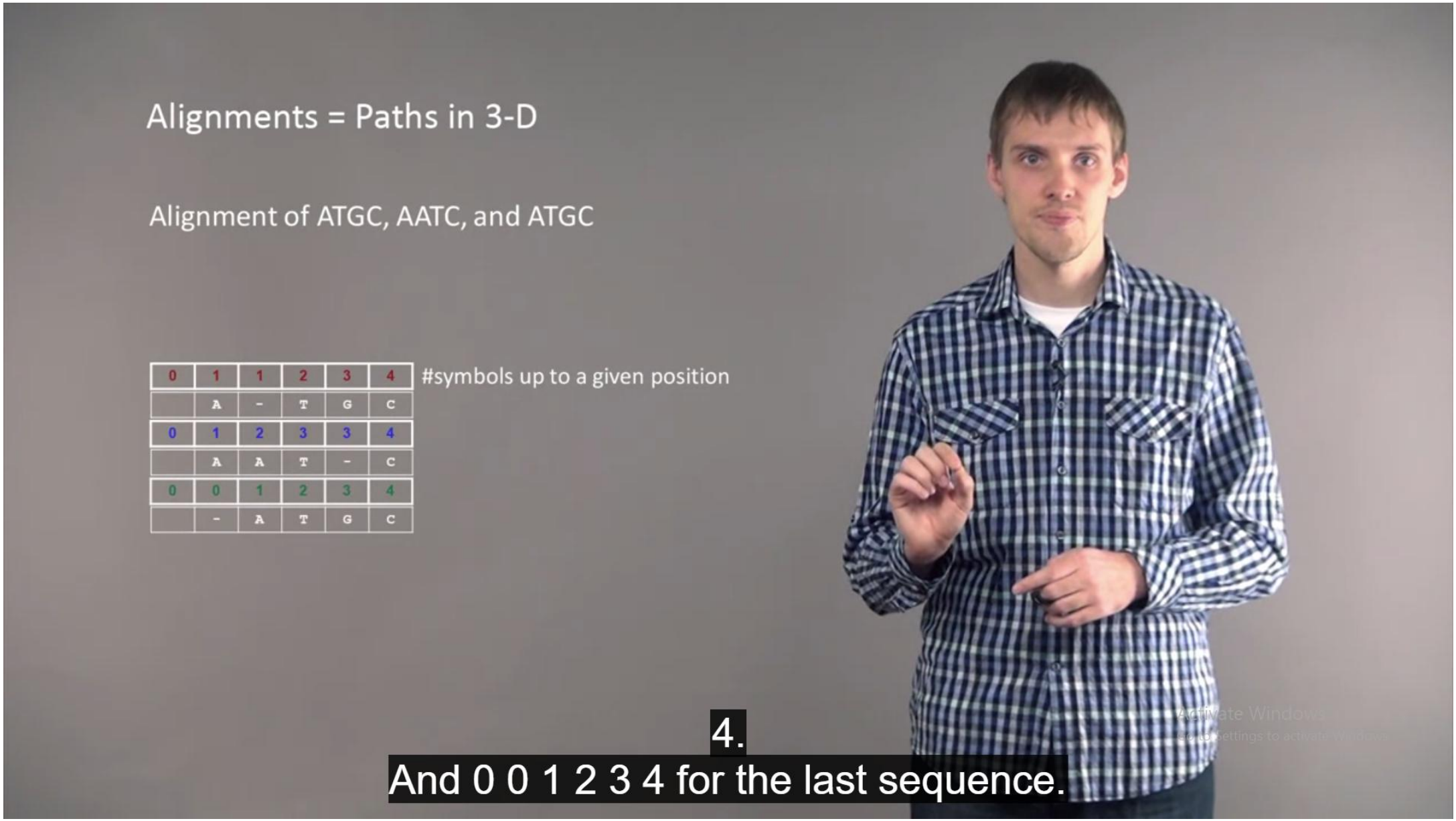


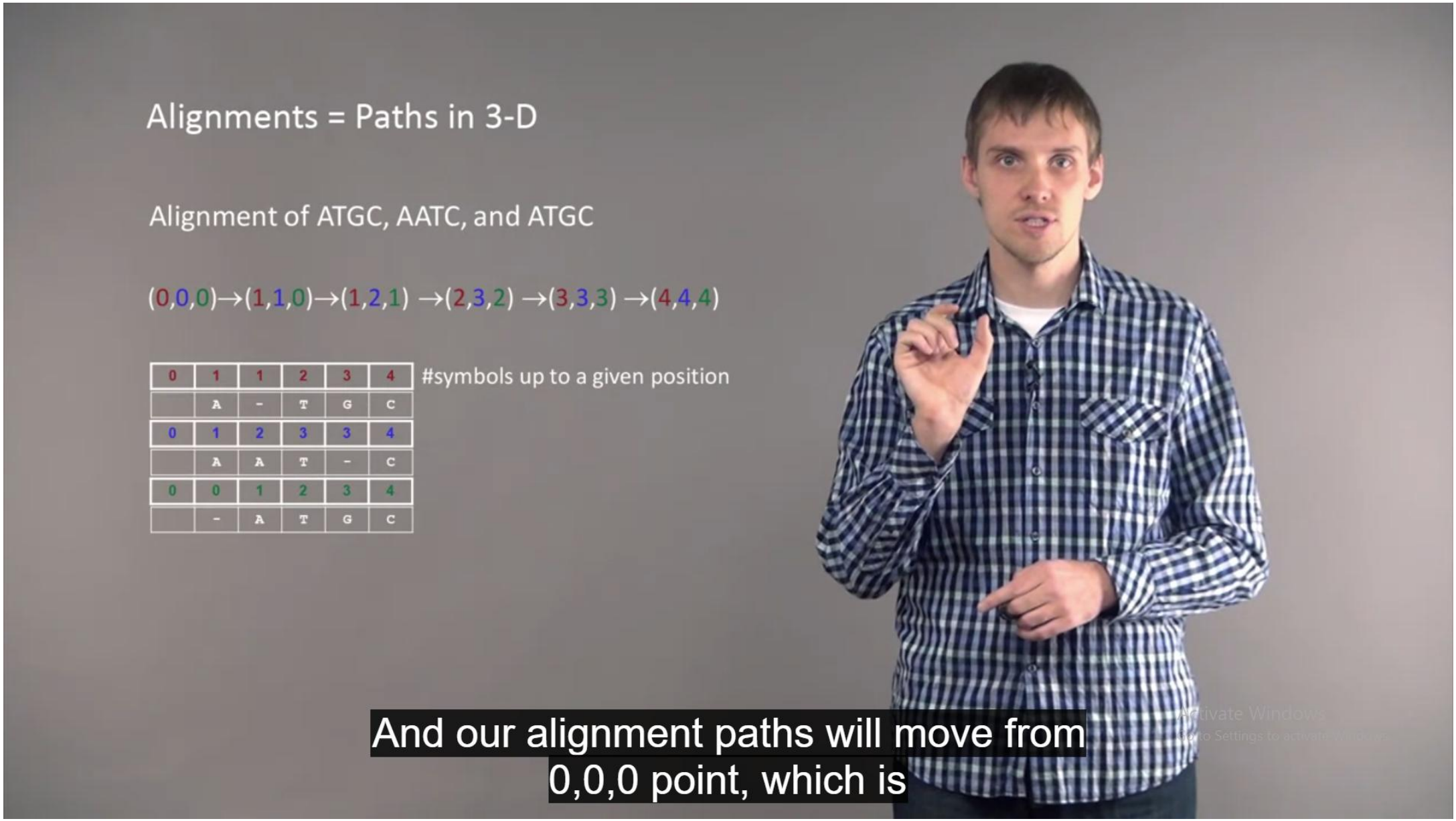
A-domains, but when you're trying to compare all of them at once, you will


Generalizing Pairwise to Multiple Alignment
Alignment of 2 sequences is a 2-row matrix
Alignment of 3 sequences is a 3-row matrix
A T - G C G -
A - C G T - A
A T C A C - A
Our scoring function should score alignments with conserved columns higher
And we will need to design our scoring function to score alignments with more

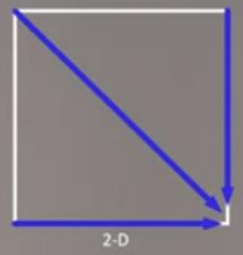

Alignments = Paths in 3-D
Alignment of ATGC, AATC, and ATGC
A - T G C
A A T - C
- A T G C
our pairwise alignment was a path in a two dimensional Manhattan grid.

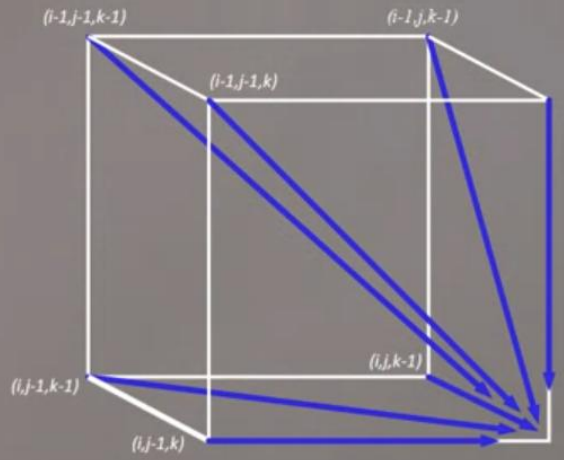

Alignments = Paths in 3-D
Alignment of ATGC, AATC, and ATGC
0 1 1 2 3 4
#symbols up to a given position
A - T G C
A A T - C
- A T G C
Here you can see the number of symbols up to given position
2:23 / 13:04


Alignments = Paths in 3-D
Alignment of ATGC, AATC, and ATGC
0 1 1 2 3 4
A - T G C
0 1 2 3 3 4
A A T - C
0 0 1 2 3 4
- A T G C
#symbols up to a given position
4.
And 0 0 1 2 3 4 for the last sequence.


Alignments = Paths in 3-D
Alignment of ATGC, AATC, and ATGC
(0,0,0)→(1,1,0)→(1,2,1) →(2,3,2) →(3,3,3) →(4,4,4)
0 1 1 2 3 4
A - T G C
0 1 2 3 3 4
A A T - C
0 0 1 2 3 4
- A T G C
#symbols up to a given position
And our alignment paths will move from
0,0,0 point, which is

2-D Alignment Cell versus 3-D Alignment Cell





And two dimensional alignment

Multiple Alignment: Dynamic Programming

$$s_{i,j,k} = \max \begin{cases} s_{i-1,j-1,k-1} + \delta(v_i, w_j, u_k) \\ s_{i-1,j-1,k} + \delta(v_i, w_j, -) \\ s_{i-1,j,k-1} + \delta(v_i, -, u_k) \\ s_{i,j-1,k-1} + \delta(-, w_j, u_k) \\ s_{i-1,j,k} + \delta(v_i, -, -) \\ s_{i,j-1,k} + \delta(-, w_j, -) \\ s_{i,j,k-1} + \delta(-, -, u_k) \end{cases}$$

$\delta(x, y, z)$ is an entry in the 3-D scoring matrix

And here is dynamic programming equations for

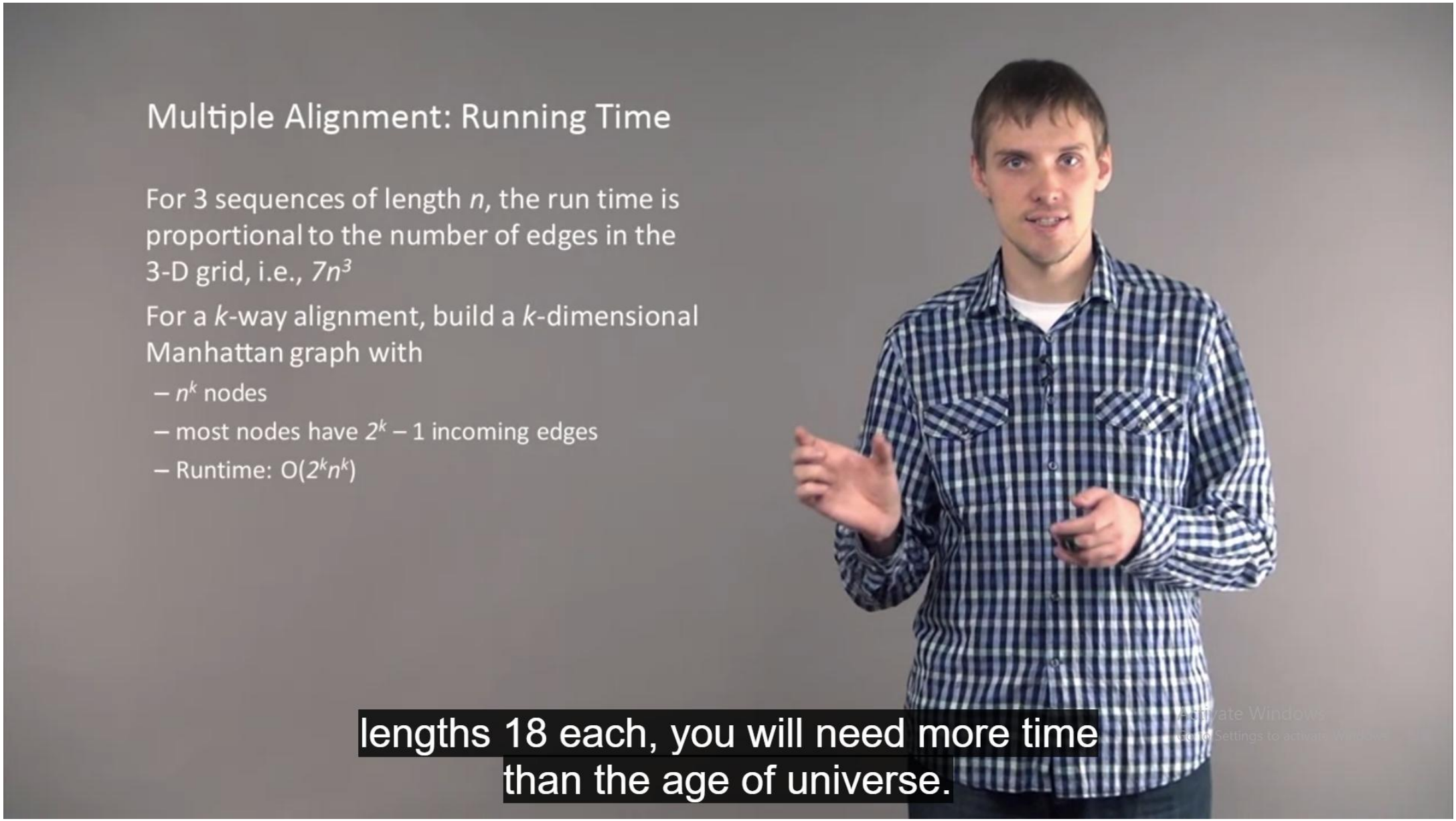
Multiple Alignment: Running Time

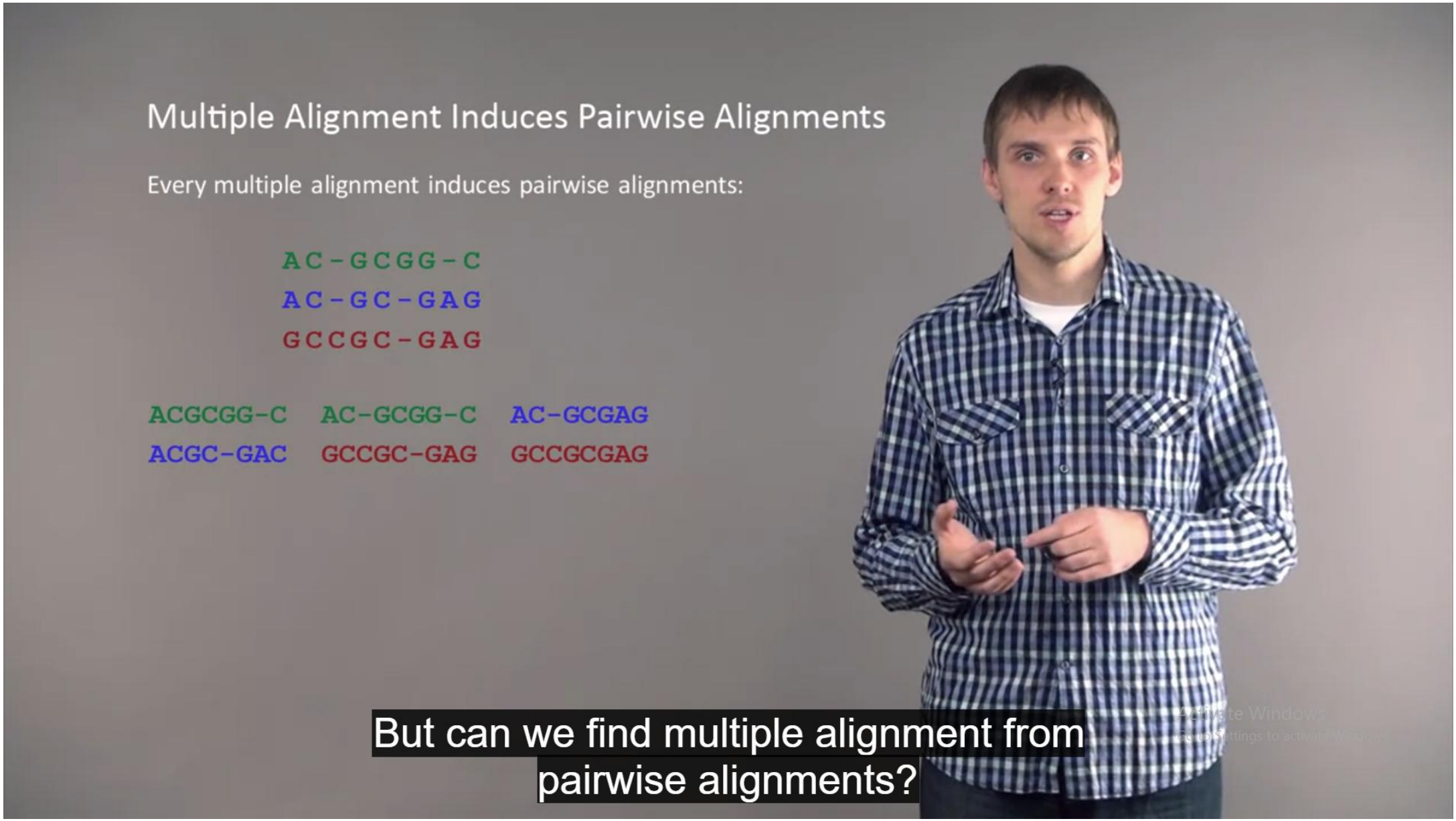
For 3 sequences of length n, the run time is proportional to the number of edges in the 3-D grid, i.e., $7n^3$

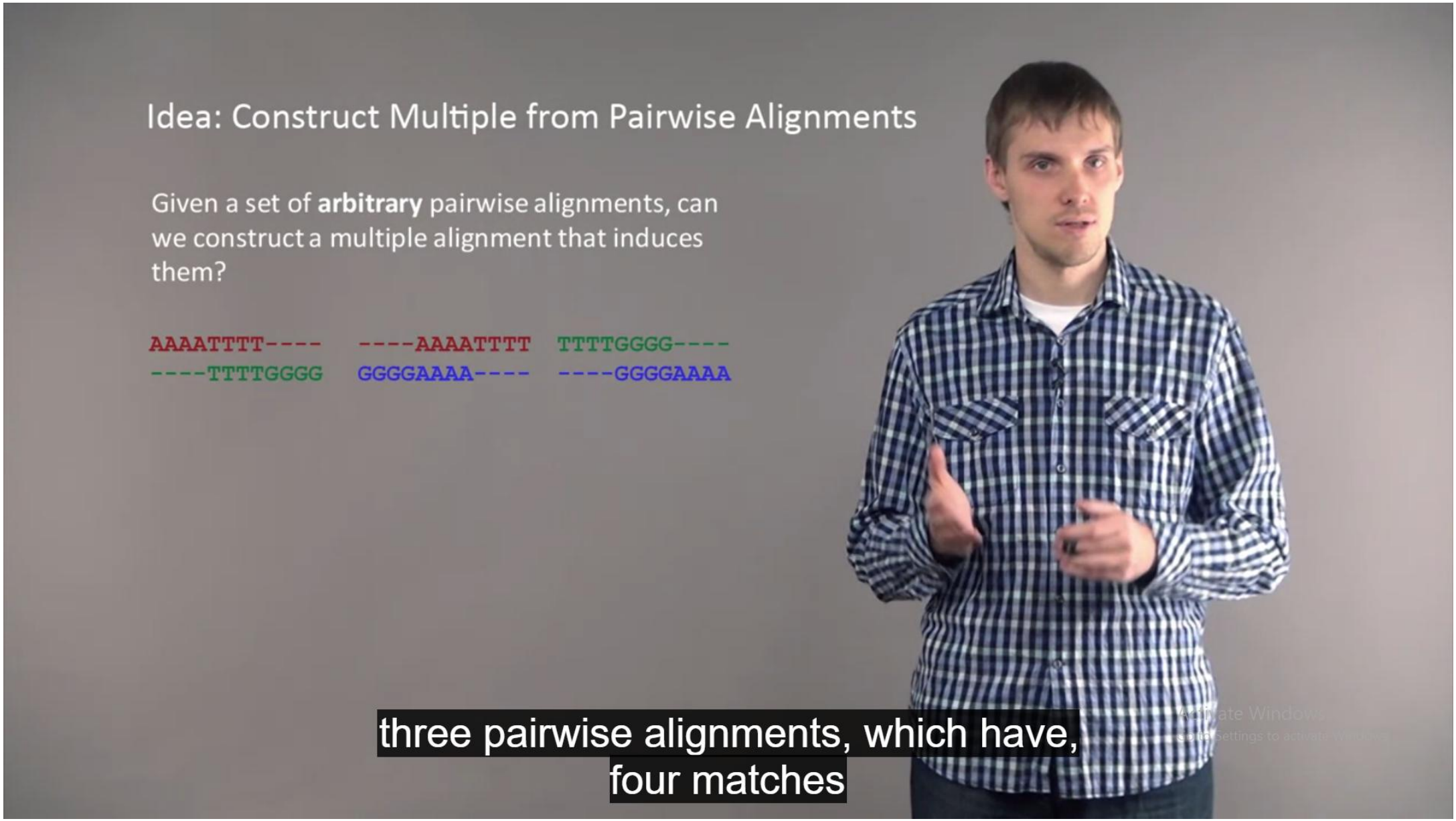
For a k-way alignment, build a k-dimensional Manhattan graph with

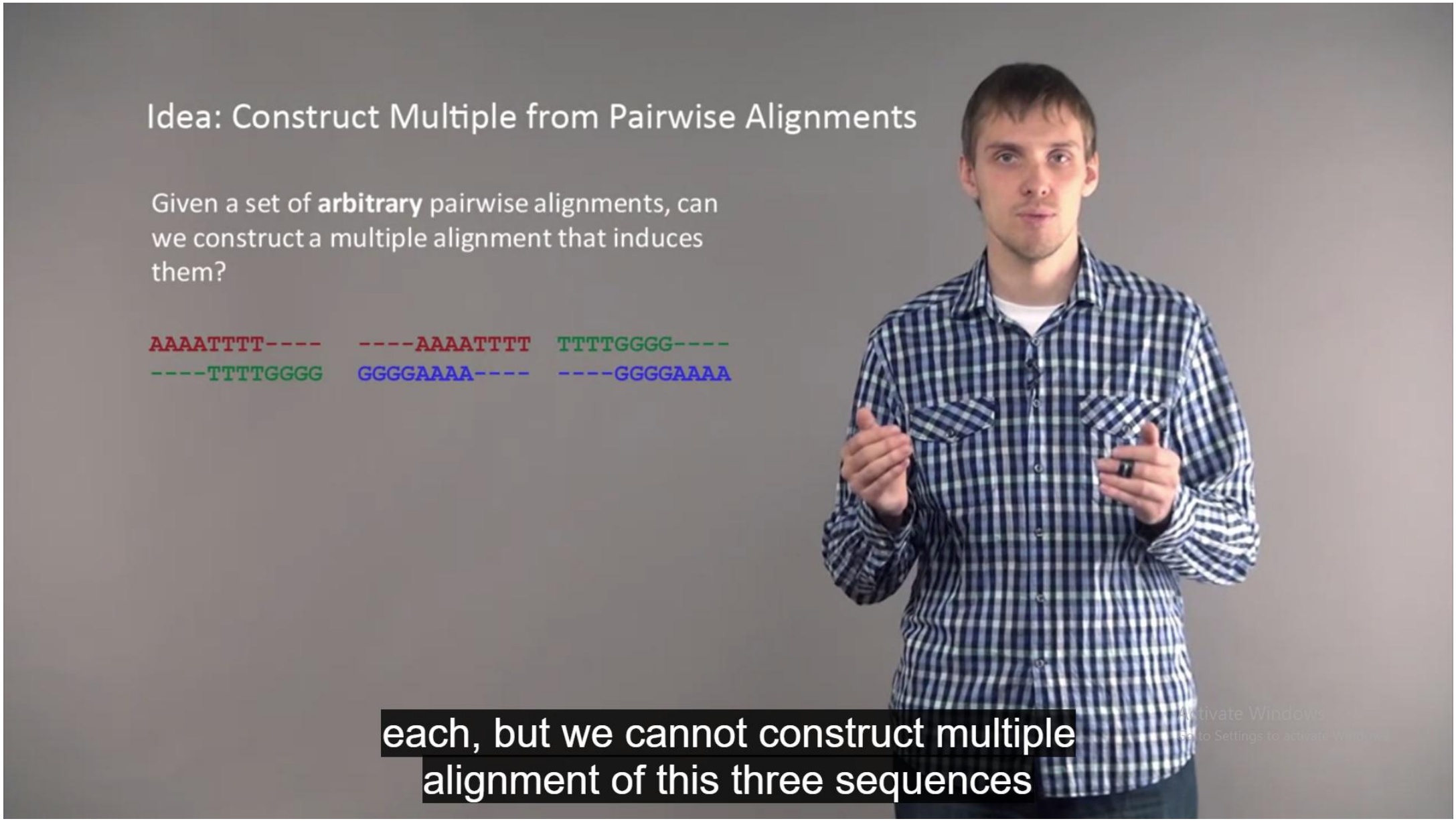
- n^k nodes
- most nodes have $2^k - 1$ incoming edges
- Runtime: $O(2^k n^k)$

Similar as before, but the number of edges now is much larger.

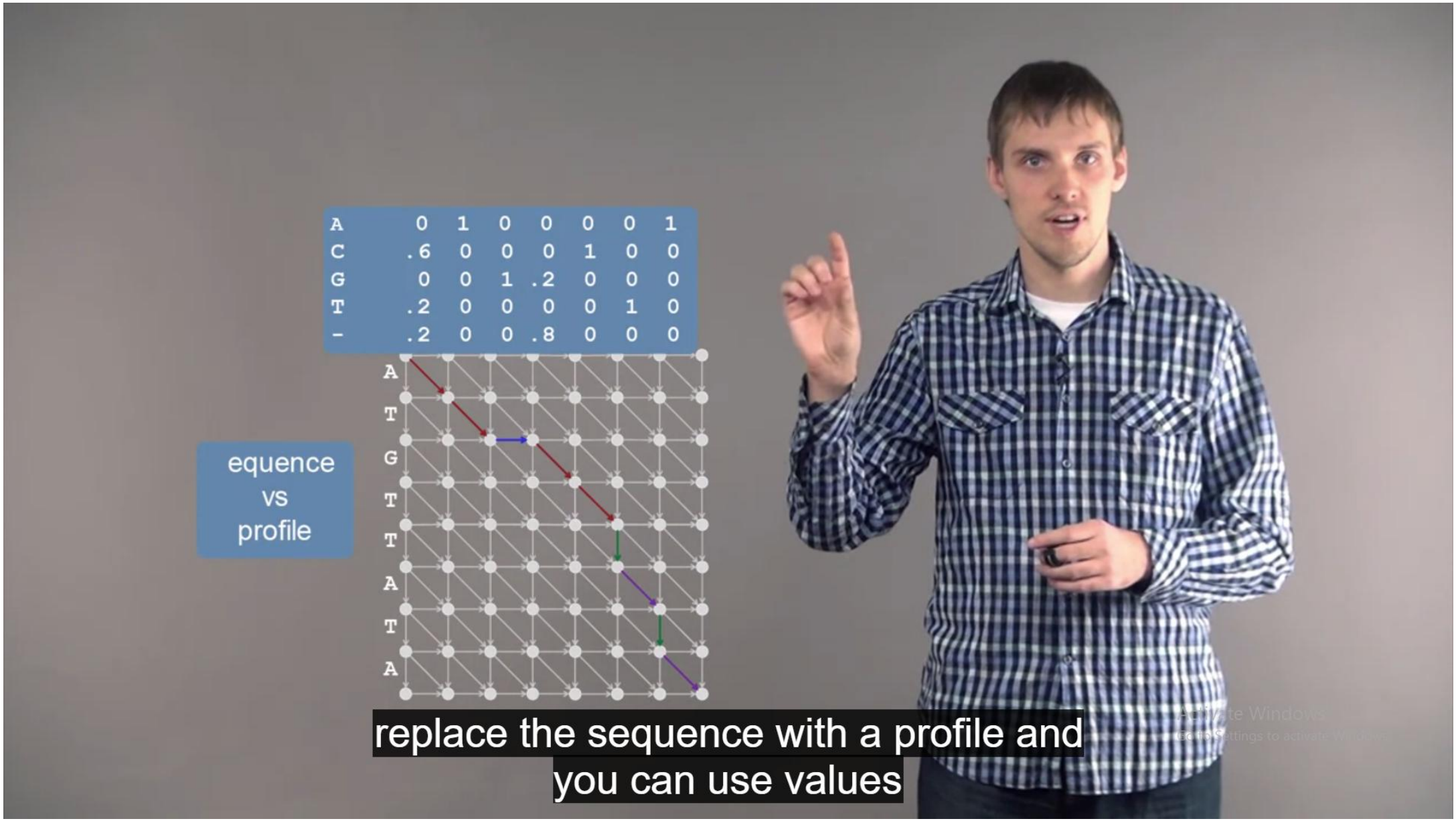

Multiple Alignment: Running Time
For 3 sequences of length n, the run time is proportional to the number of edges in the 3-D grid, i.e., $7n^3$
For a k-way alignment, build a k-dimensional Manhattan graph with
– n^k nodes
– most nodes have $2^k - 1$ incoming edges
– Runtime: $O(2^k n^k)$
lengths 18 each, you will need more time than the age of universe.


Multiple Alignment Induces Pairwise Alignments
Every multiple alignment induces pairwise alignments:
AC-GCGG-C
AC-GC-GAG
GCCGC-GAG
ACGCGG-C
ACGC-GAC
AC-GCGG-C
GCCGC-GAG
AC-GCGAG
GCCGCGAG
But can we find multiple alignment from pairwise alignments?


Idea: Construct Multiple from Pairwise Alignments
Given a set of **arbitrary** pairwise alignments, can we construct a multiple alignment that induces them?
AAAATTTT----
----AAAATTTT
TTTTGGGG----
----TTTTGGGG
GGGGAAAA----
----GGGGAAAA
three pairwise alignments, which have, four matches


Idea: Construct Multiple from Pairwise Alignments
Given a set of arbitrary pairwise alignments, can we construct a multiple alignment that induces them?
AAAATTTT----
----AAAATTTT
TTTTGGGG----
----TTTTGGGG
GGGGAAAA----
----GGGGAAAA
each, but we cannot construct multiple alignment of this three sequences


Aligning Profile Against Profile
In the past we were aligning a sequence against a sequence
– Can we align a sequence against a profile?
– Can we align a profile against a profile?
- A G G C T A T C A C C T G
T A G - C T A C C A - - - G
C A G - C T A C C A - - - G
C A G - C T A T C A C - G G
C A G - C T A T C G C - G G
A 0 1 0 0 0 0 1 0 0 .8 0 0 0 0
C .6 0 0 0 1 0 0 .4 1 0 .6 .2 0 0
G 0 0 1 .2 0 0 0 0 0 .2 0 0 .4 1
T .2 0 0 0 0 1 0 .6 0 0 0 0 .2 0
- .2 0 0 .8 0 0 0 0 0 0 .4 .8 .4 0
But can we align sequence against profile?
Or can we align profile against profile?

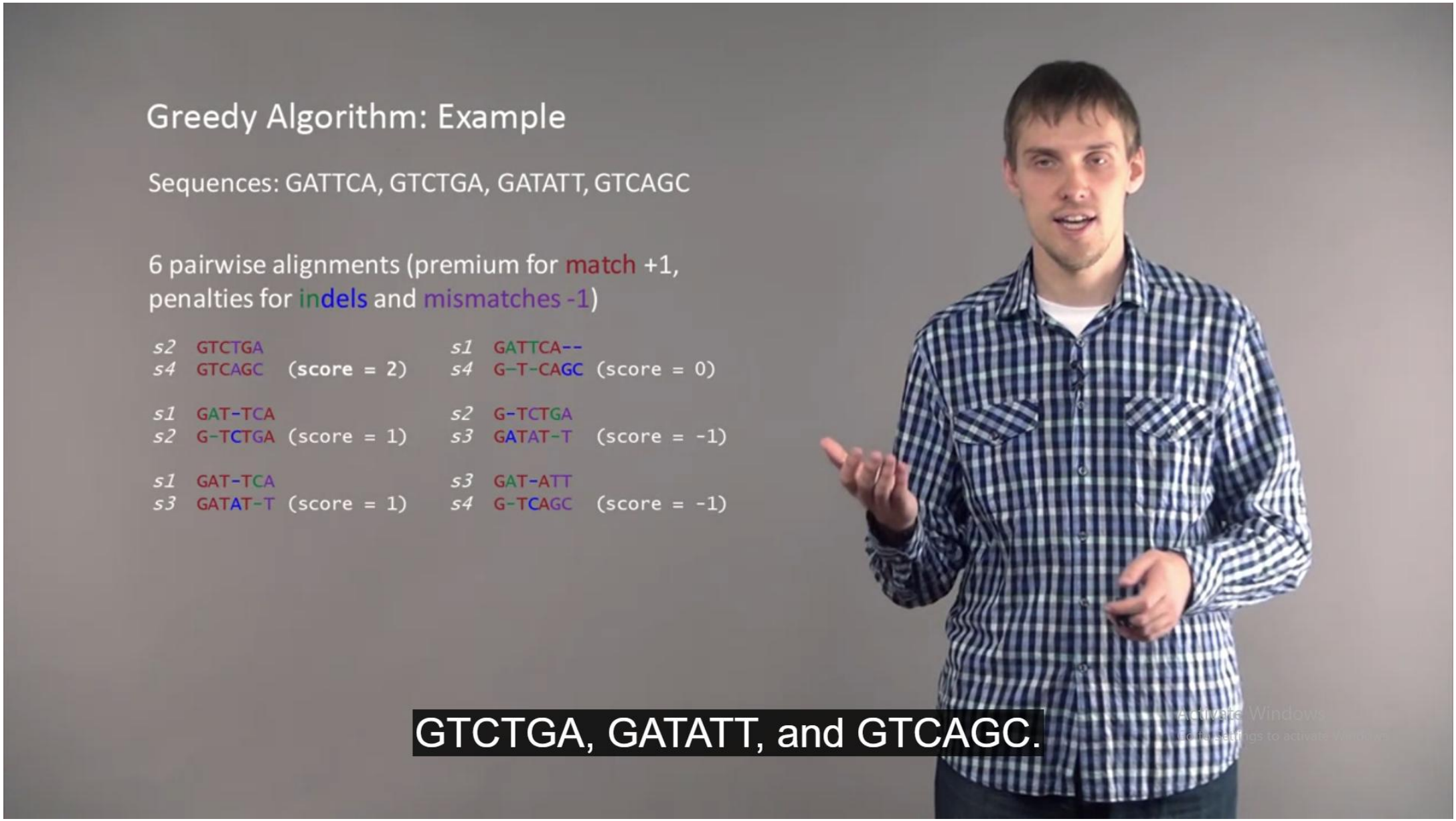

A 0 1 0 0 0 0 1
C .6 0 0 0 1 0 0
G 0 0 1 .2 0 0 0
T .2 0 0 0 0 1 0
- .2 0 0 .8 0 0 0
A
T
G
T
T
A
T
A
equence
vs
profile
replace the sequence with a profile and
you can use values

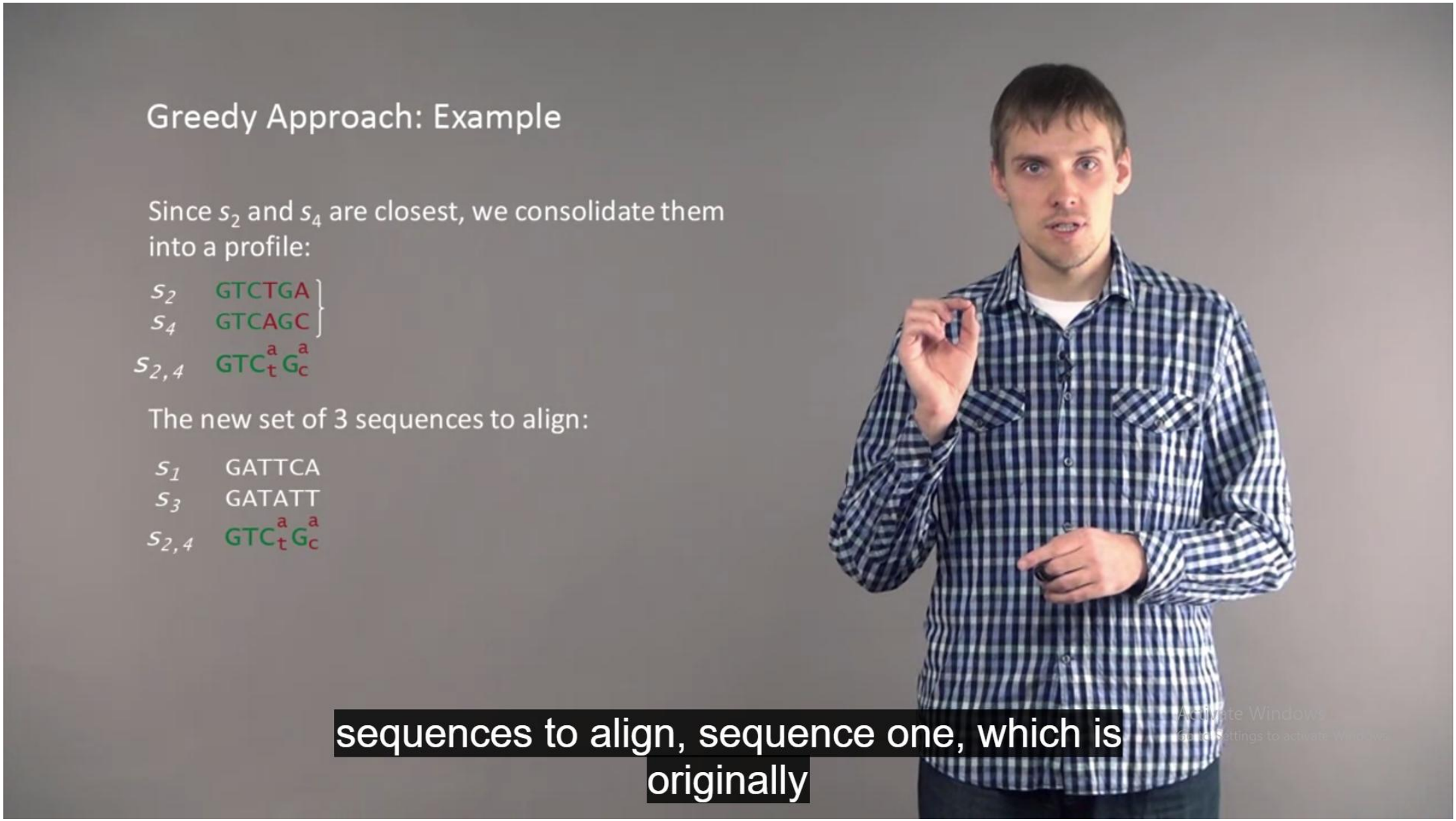
Greedy Multiple Alignment Algorithms

Choose the most similar sequences and combine them into a profile, thereby reducing *k*-way alignment to *(k-1)*-way alignment of *(k – 2)* sequences and 1 profile

Iterate…

to find multiple alignment.
We can construct simple greedy algorithm.


Greedy Algorithm: Example
Sequences: GATTCA, GTCTGA, GATATT, GTCAGC
6 pairwise alignments (premium for match +1, penalties for indels and mismatches -1)
s2 GTCTGA
s4 GTCAGC (score = 2)
s1 GATTCA--
s4 G-T-CAGC (score = 0)
s1 GAT-TCA
s2 G-TCTGA (score = 1)
s2 G-TCTGA
s3 GATAT-T (score = -1)
s1 GAT-TCA
s3 GATAT-T (score = 1)
s3 GAT-ATT
s4 G-TCAGC (score = -1)
GTCTGA, GATATT, and GTCAGC.


Greedy Approach: Example
Since s_2 and s_4 are closest, we consolidate them into a profile:
s_2 GTCTGA
s_4 GTCAGC
$s_{2,4}$ GTC$^{a}_{t}$G$^{a}_{c}$
The new set of 3 sequences to align:
s_1 GATTCA
s_3 GATATT
$s_{2,4}$ GTC$^{a}_{t}$G$^{a}_{c}$
sequences to align, sequence one, which is originally

We Learned a Lot about Alignment but…

Can you find all similarities shared by human, mouse, and rat genomes?

Can you rapidly find the best alignments between the human genome and millions of short reads?

Are alignment algorithms with quadratic running time practical when we analyze entire genomes?