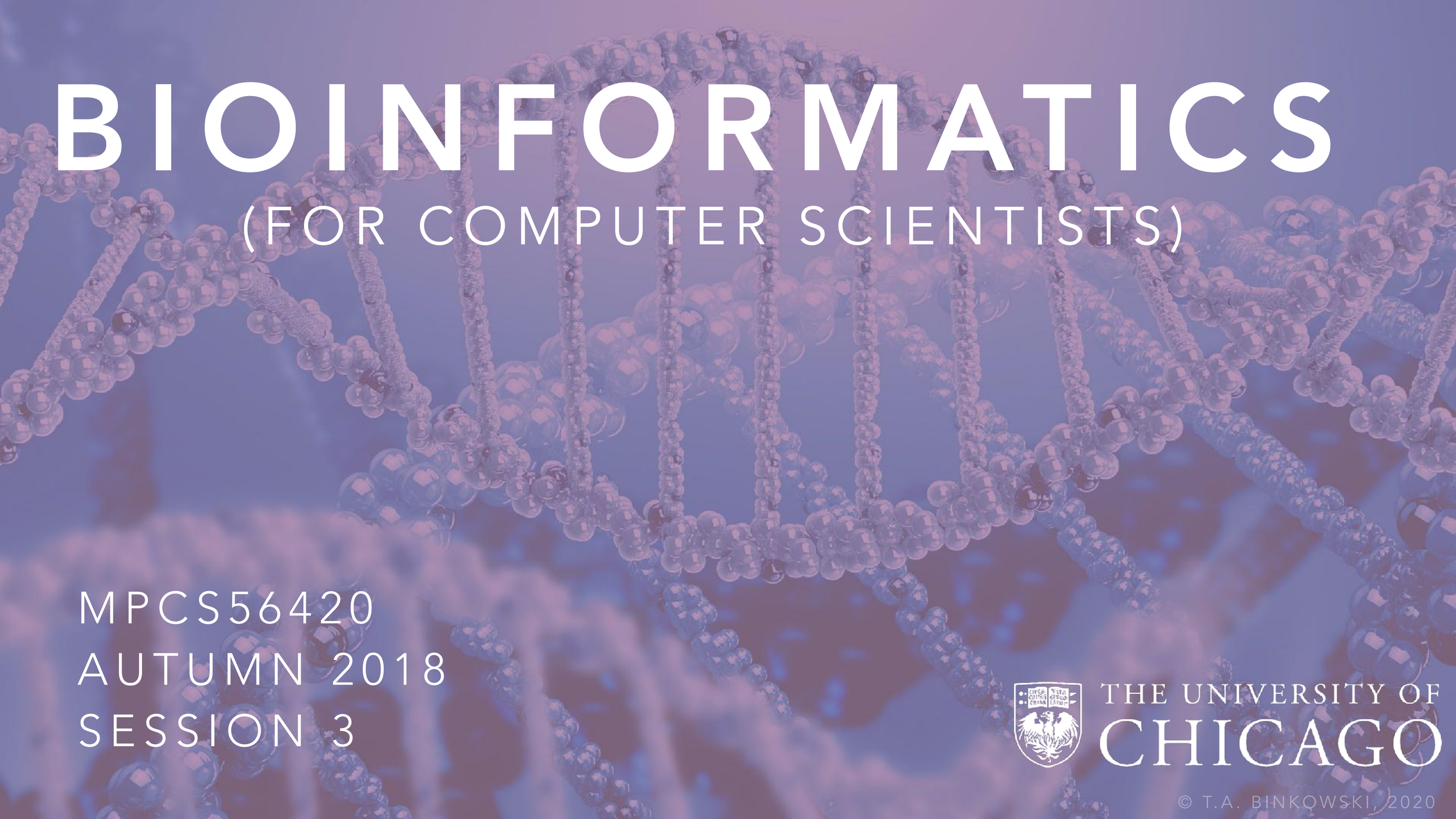


# BIOINFORMATICS

## (FOR COMPUTER SCIENTISTS)

MPCS56420
AUTUMN 2018
SESSION 3



THE UNIVERSITY OF CHICAGO


SESSION 3

# SEQUENCE ALIGNMENT

- Mutations, substitutions and evolutions
- Sequence similarity
- Pairwise sequence alignment
- Sequence alignment algorithms
- Substitution matrices
- Significance of alignments


ORF REVIEW

# TRANSCRIPTIONS & TRANSLATION

- Open reading frame (ORF)
  - A reading frame that contains a start codon and a stop codon, with multiple three-nucleotide codons in between
  - Hypothesis for correct reading frame from which to translate the DNA into protein
  - May contain introns (non-coding regions) in eukaryotes

- Coding Sequence (CDS)
  - The actual region of DNA that is translated to protein

# TRANSCRIPTIONS & TRANSLATION

5′- CCGATGTCATAAGAC - 3′

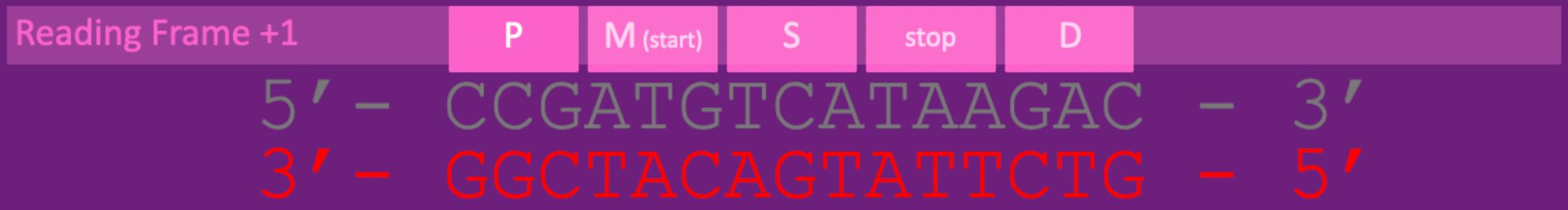
# TRANSCRIPTIONS & TRANSLATION

- Find the complement sequence

5′- CCGATGTCATAAGAC - 3′

3′- GGCTACAGTATTCTG - 5′

# TRANSCRIPTIONS & TRANSLATION

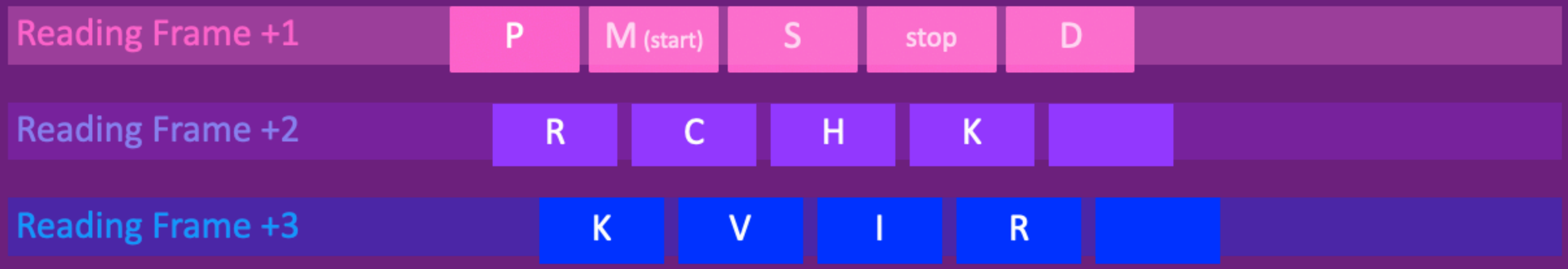


# TRANSCRIPTIONS & TRANSLATION

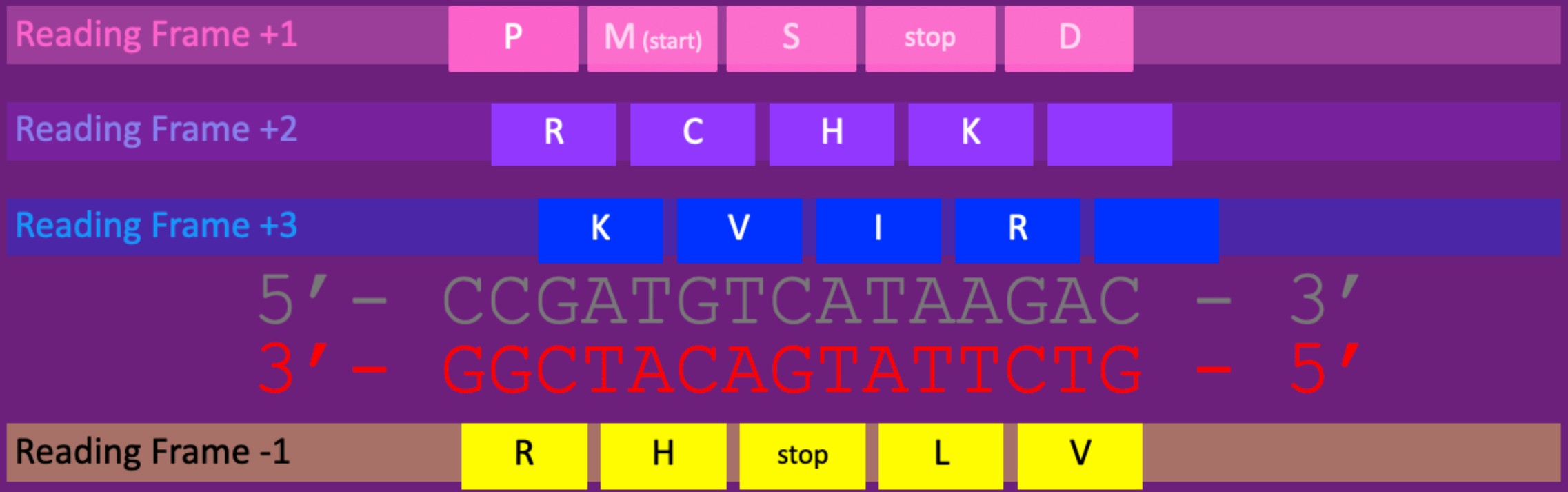

5′– CCGATGTCATAAGAC – 3′

3′– GGCTACAGTATTCTG – 5′

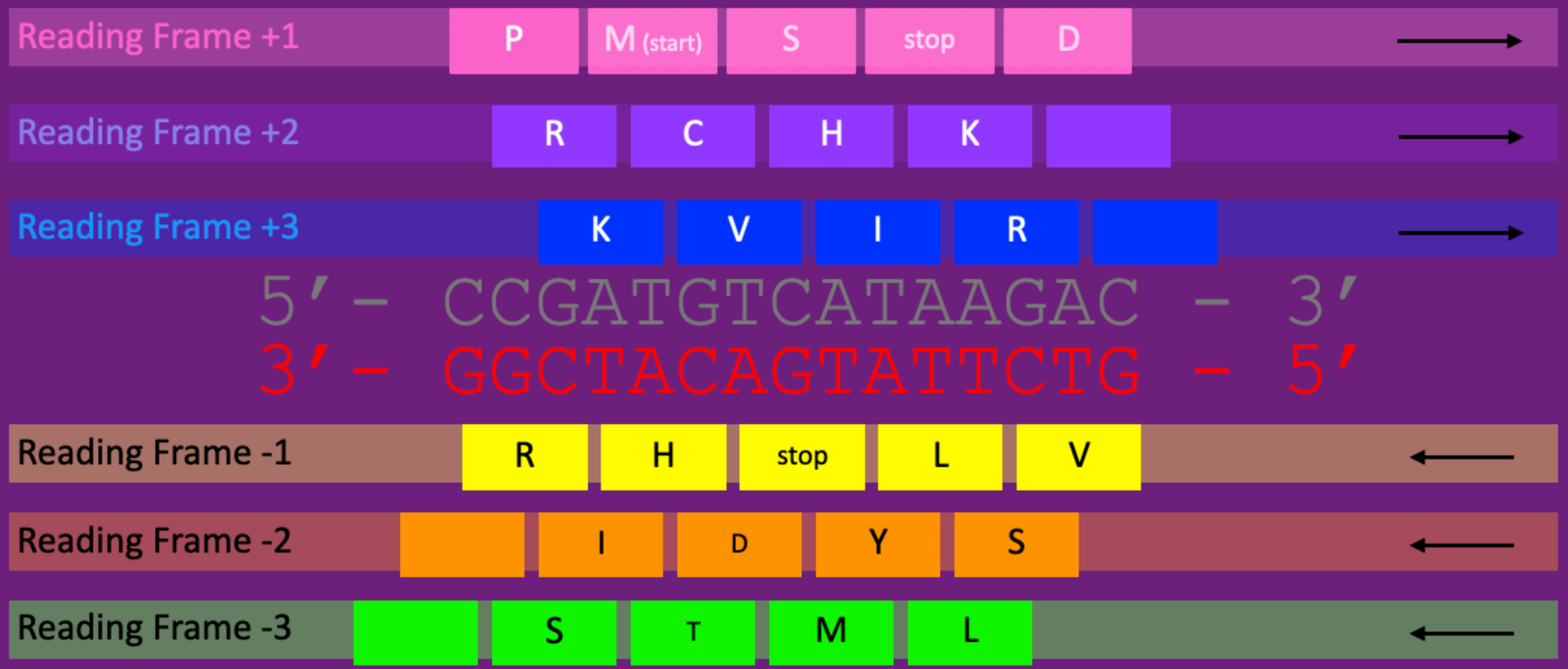
# TRANSCRIPTIONS & TRANSLATION



# TRANSCRIPTIONS & TRANSLATION

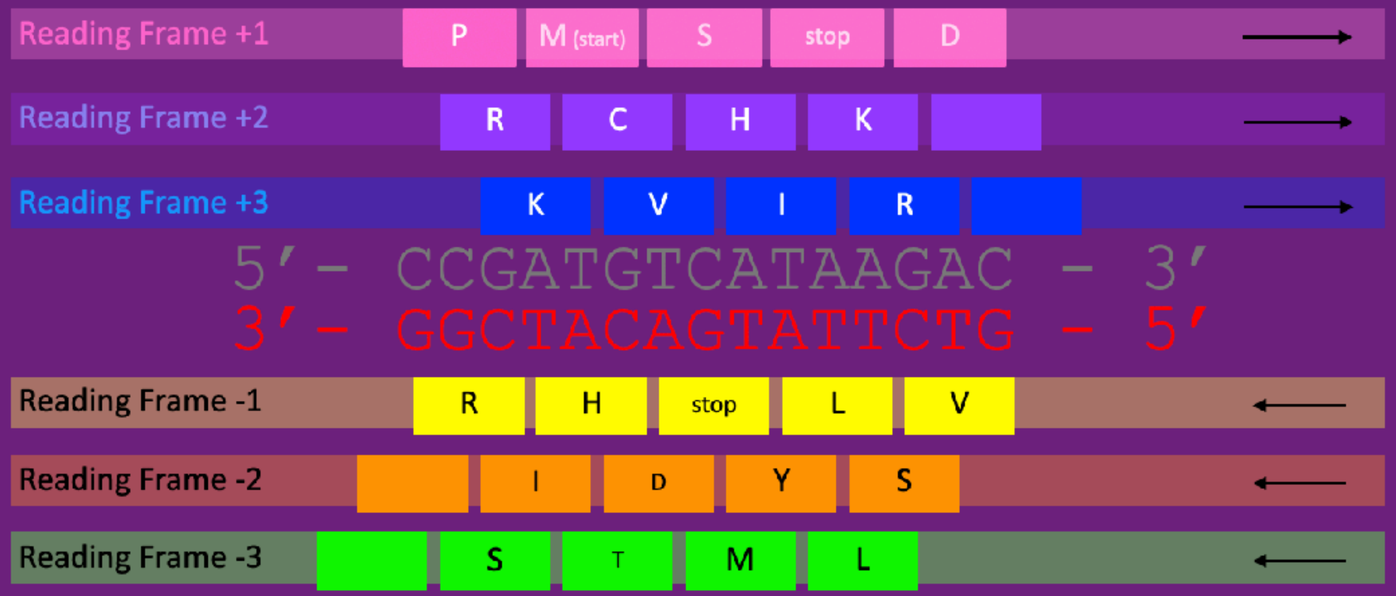


# TRANSCRIPTIONS & TRANSLATION



# TRANSCRIPTIONS & TRANSLATION

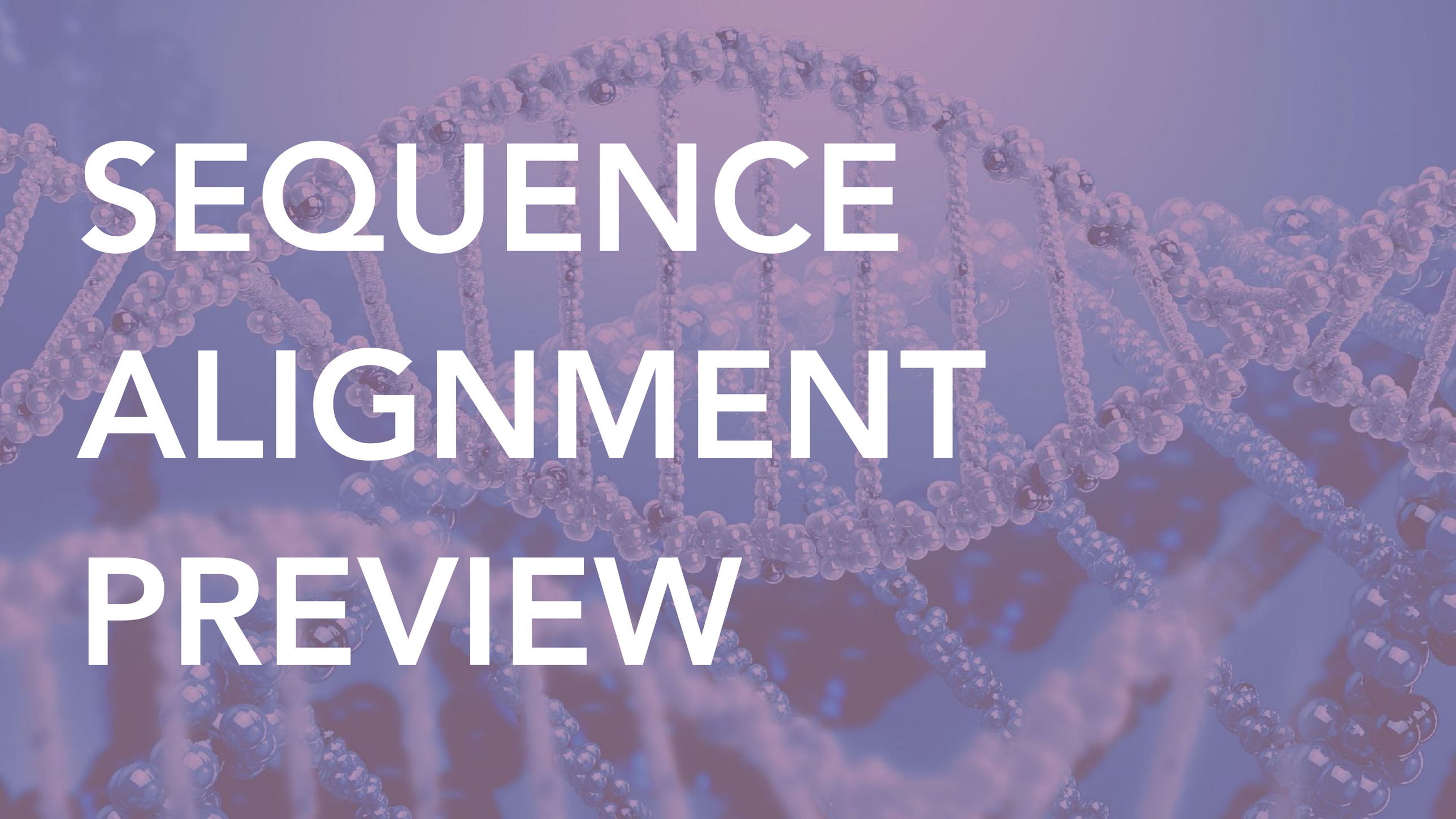
- Possible Sequences
  - PMS-stop-D
  - RCHK
  - KVIR
  - VL-stop-HR
  - SYDI
  - LMTS



# TRANSCRIPTIONS & TRANSLATION

- Possible Sequences
  - PMS-stop-D
  - RCHK
  - KVIR
  - VL-stop-HR
  - SYDI
  - LMTS

- Which one is an actual coding sequence?
  - PMS-stop-D
  - Coding sequence has to have a start (M) and stop codon with at least one amino acid in between


SEQUENCE
ALIGNMENT
PREVIEW

# Issues in Aligning Sequences

# ISSUES IN ALIGNING SEQUENCES


- What is sequence alignment?
  - Locating equivalent regions of two or more sequences to assess their overall similarity
  - Not necessarily “maximize”; sometimes small segments may be more informative

# ISSUES IN ALIGNING SEQUENCES

**A Compilation of f($n$, $m$) for $1 \leq n \leq 5, 10$, and $2 \leq m \leq 5$**

| $m$ | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| $n = 1$ | 3 | 13 | 75 | 541 |
| 2 | 13 | 409 | 23917 | 2244361 |
| 3 | 63 | 16081 | 10681263 | 14638756721 |
| 4 | 321 | 699121 | 5552351121 | 117629959485121 |
| 5 | 1683 | 32193253 | 3147728203035 | $1.05 \times 10^{18}$ |
| 10 | 8097453 | 9850349744182729 | $3.32 \times 10^{26}$ | $1.35 \times 10^{38}$ |

$f(m,n)$ : The total number of possible alignments between $\vec{a}$ and $\vec{b}$

- Given two sequences, the number of possible alignments is exponential

# ISSUES IN ALIGNING SEQUENCES

- Finding the “correct” alignment involves
  - Defining a scoring scheme
  - Finding an alignment with the best score
- Alignments of related sequences should give good scores compared with alignments of randomly chosen sequences
- The correct alignment of two related sequences should ideally be the one that gives the best score

IN PRACTICE, THE CORRECT ALIGNMENT DOES NOT NECESSARILY HAVE THE BEST SCORE, SINCE NO “PERFECT” SCORING SCHEME HAS BEEN DEVISED

# ISSUES IN ALIGNING SEQUENCES

## QUESTIONS IN SEQUENCE ALIGNMENT

- What type of alignment should be performed?
  - Align the entire sequence or part of it?
  - Two sequences or multiple sequences?

# ISSUES IN ALIGNING SEQUENCES

## QUESTIONS IN SEQUENCE ALIGNMENT

- How to find the alignment?
  - Which search algorithms should be considered?
  - What are the tradeoffs?

# ISSUES IN ALIGNING SEQUENCES

## QUESTIONS IN SEQUENCE ALIGNMENT

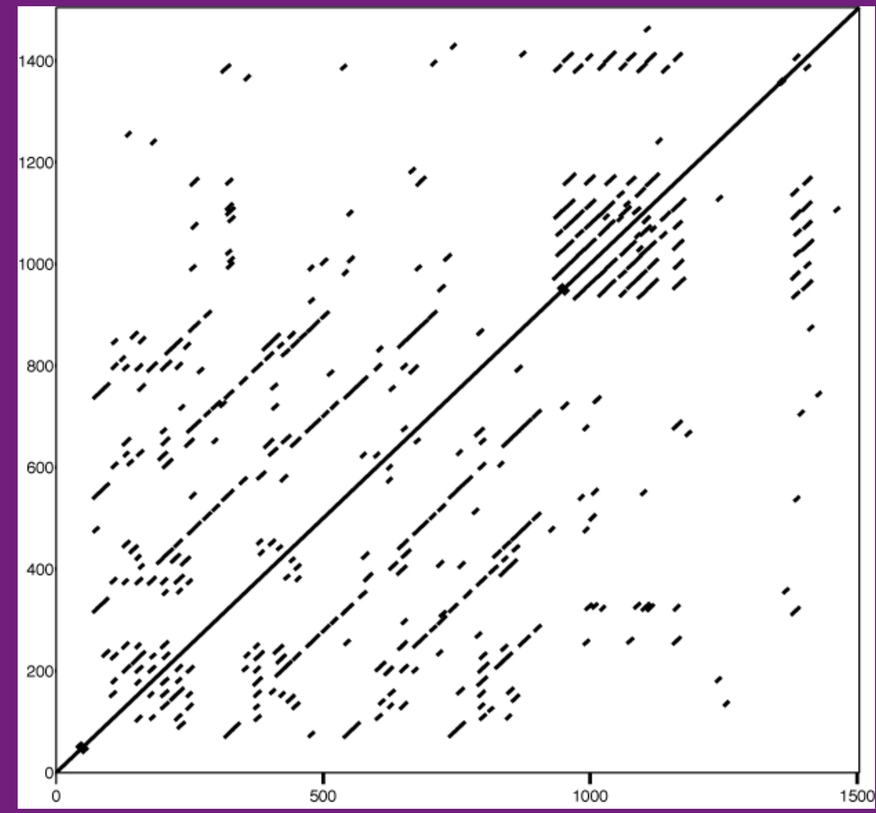
- Potential algorithms for sequence alignment
  - Visual
  - Dynamic Programming
  - Multiple sequence alignment
  - Database

QUICK VISUAL REPRESENTATION

SLOW, BUT WILL FIND OPTIMAL ALIGNMENT

MULTISTEP, EVOLUTIONARY ANALYSIS

FAST, BUT HEURISTIC



# ISSUES IN ALIGNING SEQUENCES

## QUESTIONS IN SEQUENCE ALIGNMENT

- How to score an alignment?
    - Sequences typically differ in length
    - Some characters (nucleotide or amino acid) are more substitutable than others

# ISSUES IN ALIGNING SEQUENCES

THIS SEQUENCE

THATISASEQUENCE

- Sequences of unequal length
  - Homologous genes often vary in sequence length and composition

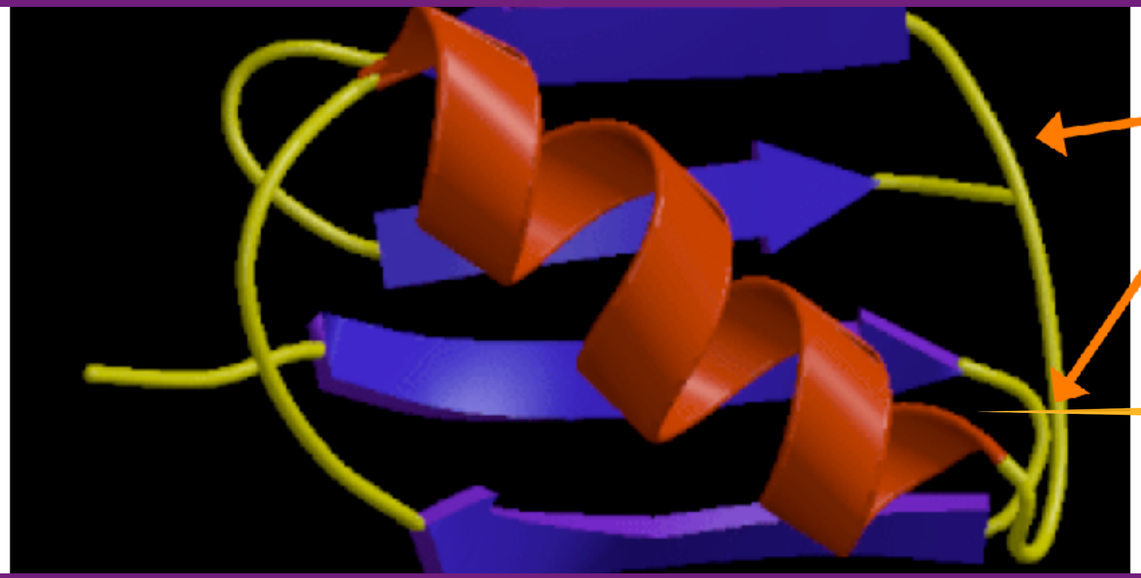
# ISSUES IN ALIGNING SEQUENCES

| | | |
|---|---|---|
| substitutions: | ACGA | AGGA |
| insertions: | ACGA | ACGGA |
| deletions: | ACGA | AGA |

- Mutations cause changes in sequences that may/may not have effect on biology
  - DNA
  - Protein

INTRODUCES GAPS IN ALIGNMENTS

# ISSUES IN ALIGNING SEQUENCES

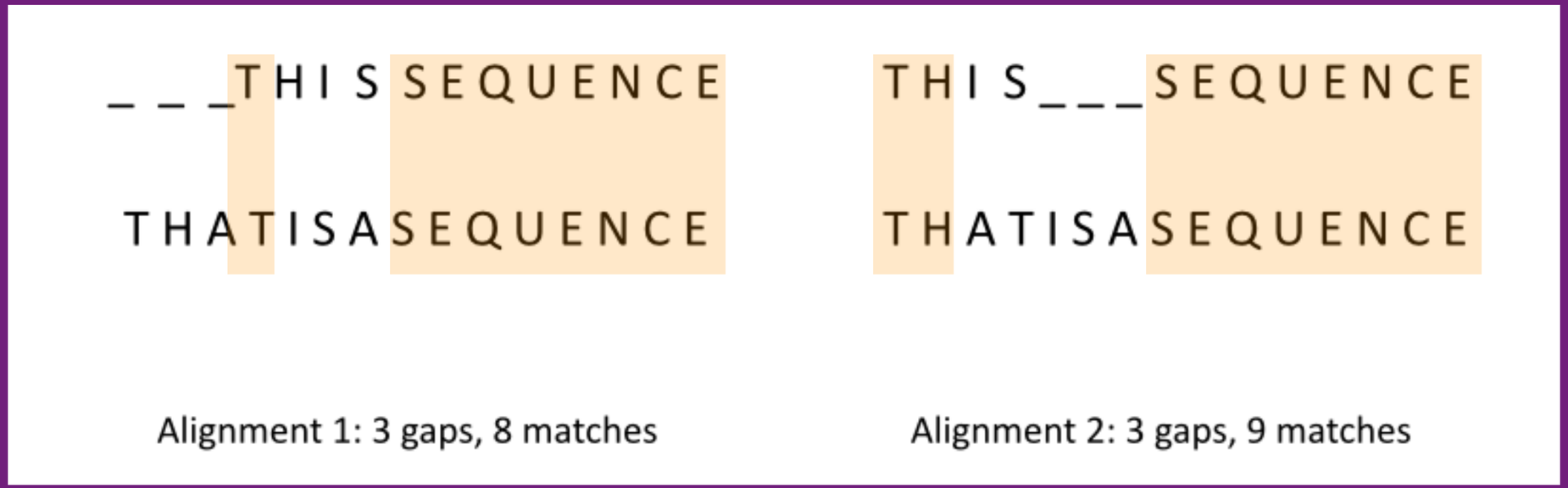


*loop structures*: insertions/deletions here not so significant

LOOP REGIONS ARE HIGHLY FLEXIBLE; NOT CONSERVED

- How is it that two “similar” sequences may have large insertions/deletions?
  - Some insertions and deletions may not significantly affect the structure of a protein

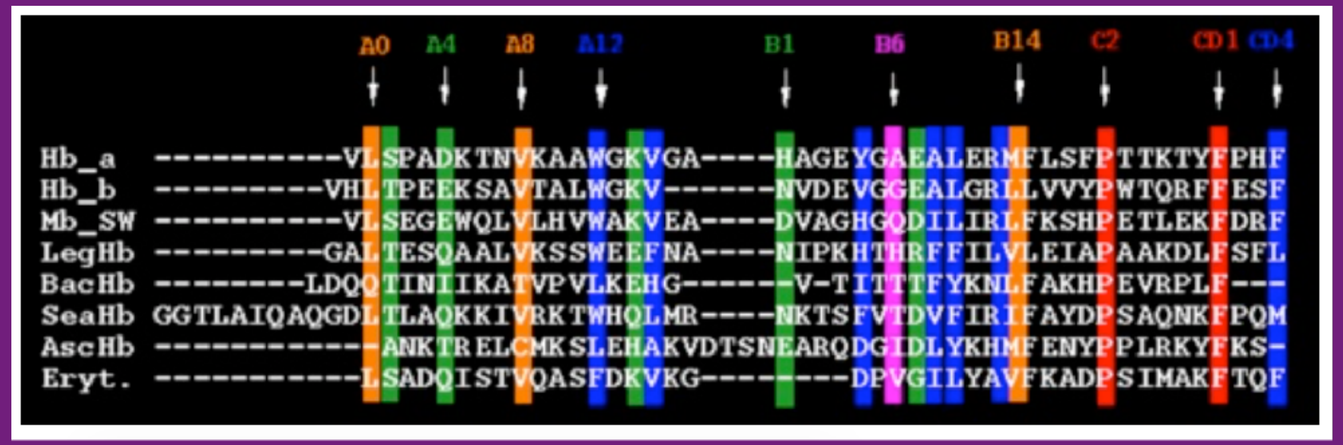
# ISSUES IN ALIGNING SEQUENCES



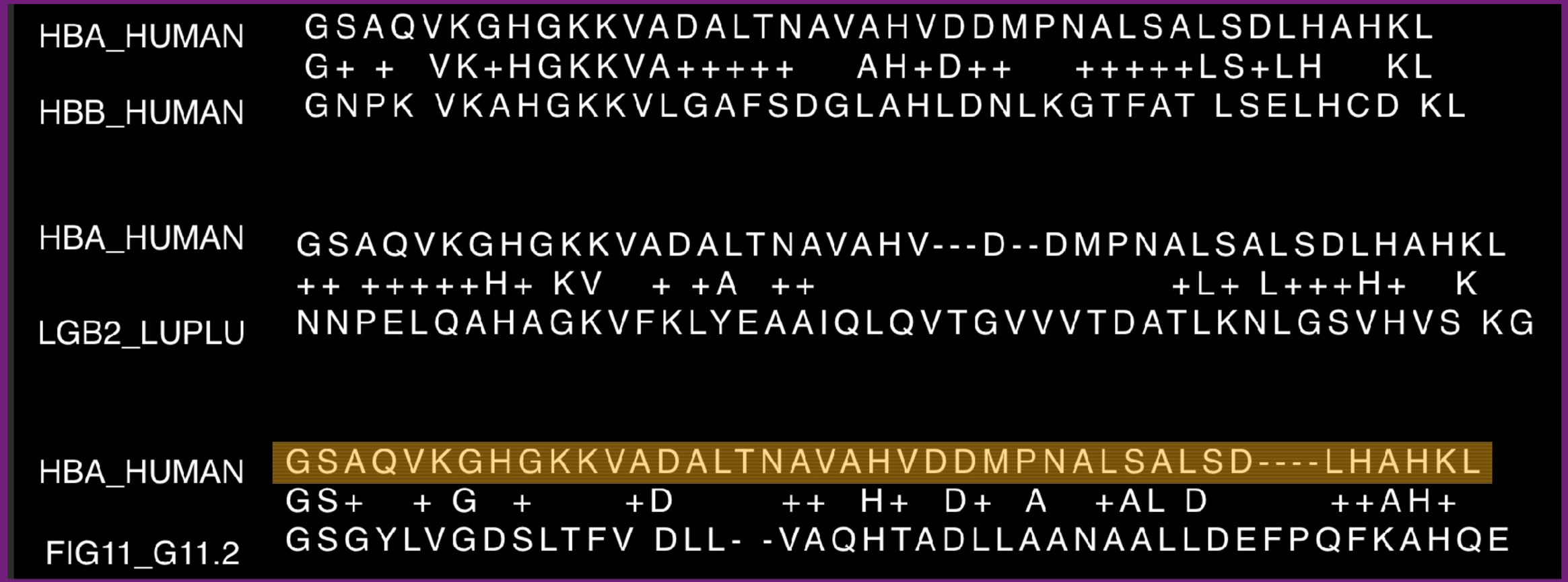

- Incorporating gaps while aligning sequences
  - Which is better?

# ISSUES IN ALIGNING SEQUENCES

- Aligned sequences (with gaps) shows conservation across evolutionarily related proteins
  - Import residues are conserved in sequence
  - Are structurally important




# ISSUES IN ALIGNING SEQUENCES



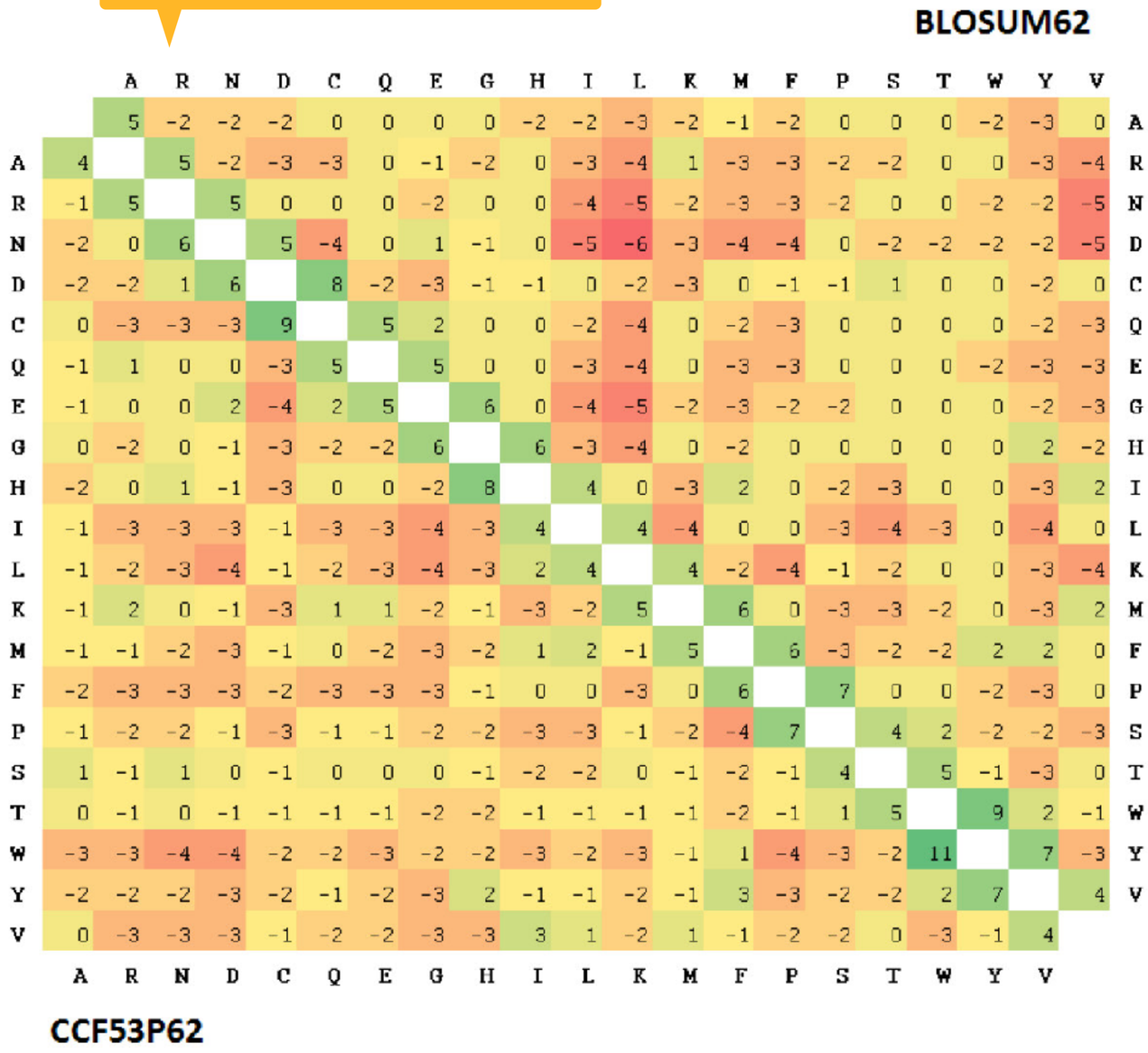

- Why do we need principled approaches to sequence alignment?
- Which of these alignments is not significant?

Example from Durbin & Eddy

# ISSUES IN ALIGNING SEQUENCES

- Components of scoring alignments
  - Percent identity
  - Gap penalty functions
  - Substitution matrices of amino acids
    - Matches may not be identical
    - Specialized matrices

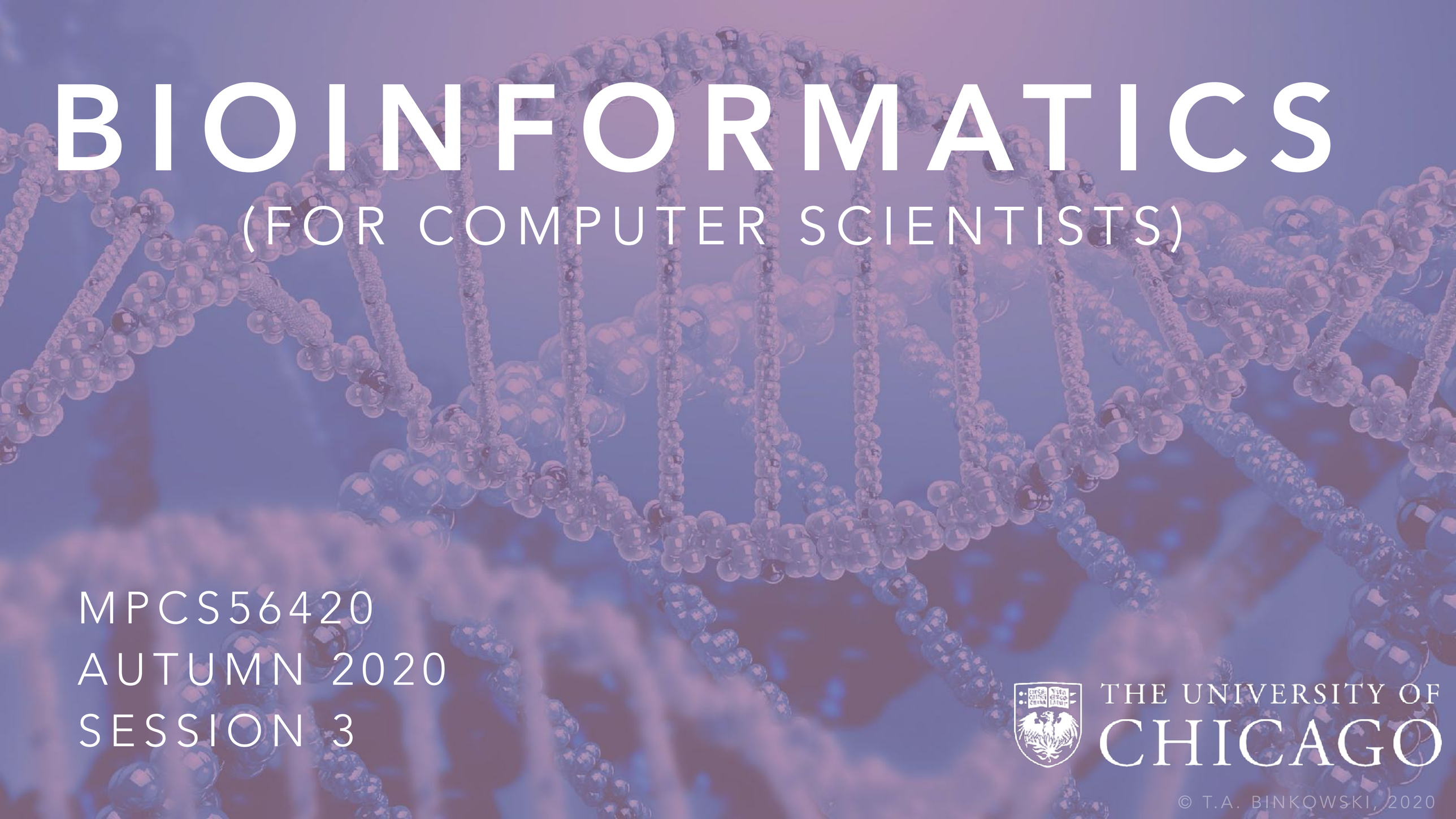
SUBSTITUTION MATRIX



# ISSUES IN ALIGNING SEQUENCES

## QUESTIONS IN SEQUENCE ALIGNMENT

- How to tell if the alignment is biologically meaningful/significant?
  - Assess how likely the alignment could have happened by random chance
  - Compare to known outcome



# BIOINFORMATICS

## (FOR COMPUTER SCIENTISTS)

MPCS56420
AUTUMN 2020
SESSION 3

THE UNIVERSITY OF CHICAGO

# SESSION 3B