

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

Reading Frame +1

P

M (start)

S

stop

D

5' – CCGATGTCATAAGAC – 3'
3' – GGCTACAGTATTCTG – 5'

TRANSCRIPTIONS & TRANSLATION

Reading Frame +1

P

M (start)

S

stop

D

Reading Frame +2

R

C

H

K

5' – CCGATGTCATAAGAC – 3'
3' – GGCTACAGTATTCTG – 5'

TRANSCRIPTIONS & TRANSLATION



5' – CCGATGTCATAAGAC – 3'
3' – GGCTACAGTATTCTG – 5'

TRANSCRIPTIONS & TRANSLATION

Reading Frame +1

P

M_(start)

S

stop

D

Reading Frame +2

R

C

H

K

Reading Frame +3

K

V

I

R

5' – CCGATGTCATAAGAC – 3'
3' – GGCTACAGTATTCTG – 5'

Reading Frame -1

R

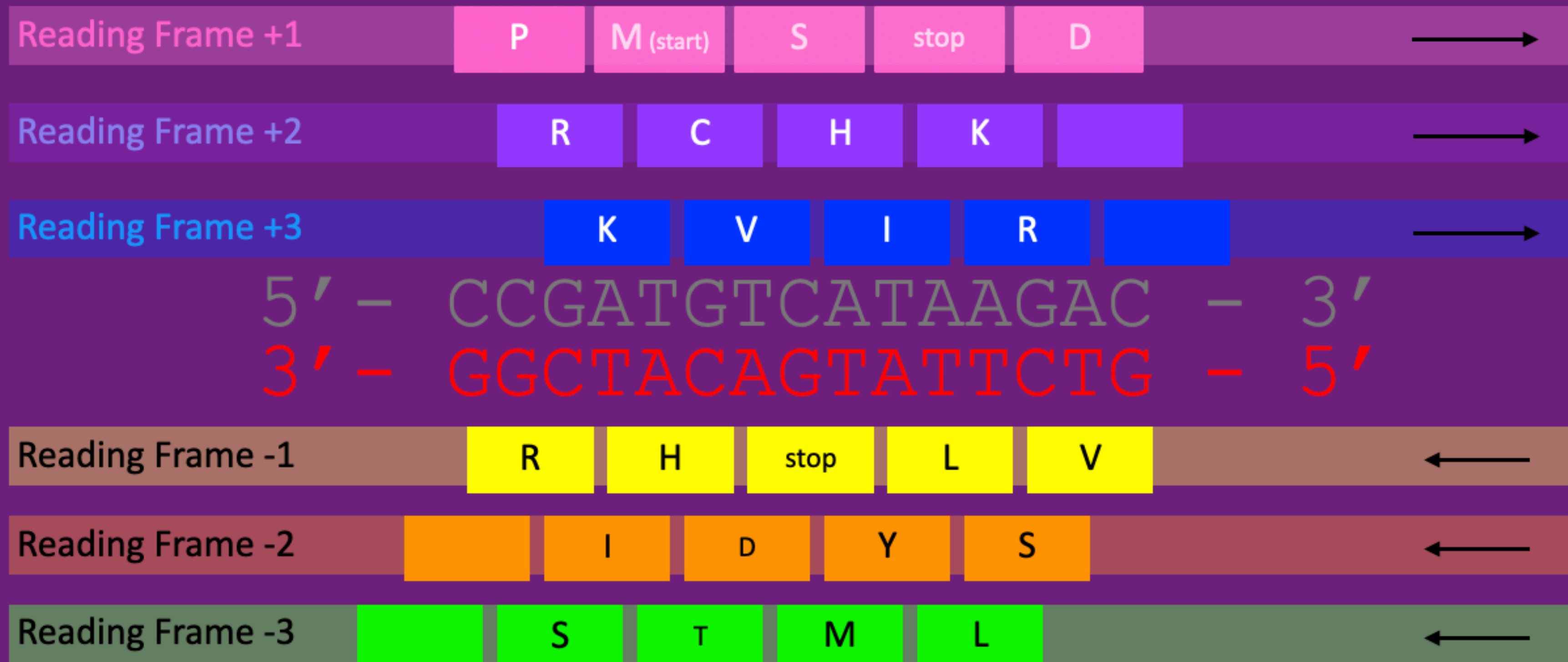
H

stop

L

V

TRANSCRIPTIONS & TRANSLATION



TRANSCRIPTIONS & TRANSLATION

- Possible Sequences

- PMS-stop-D

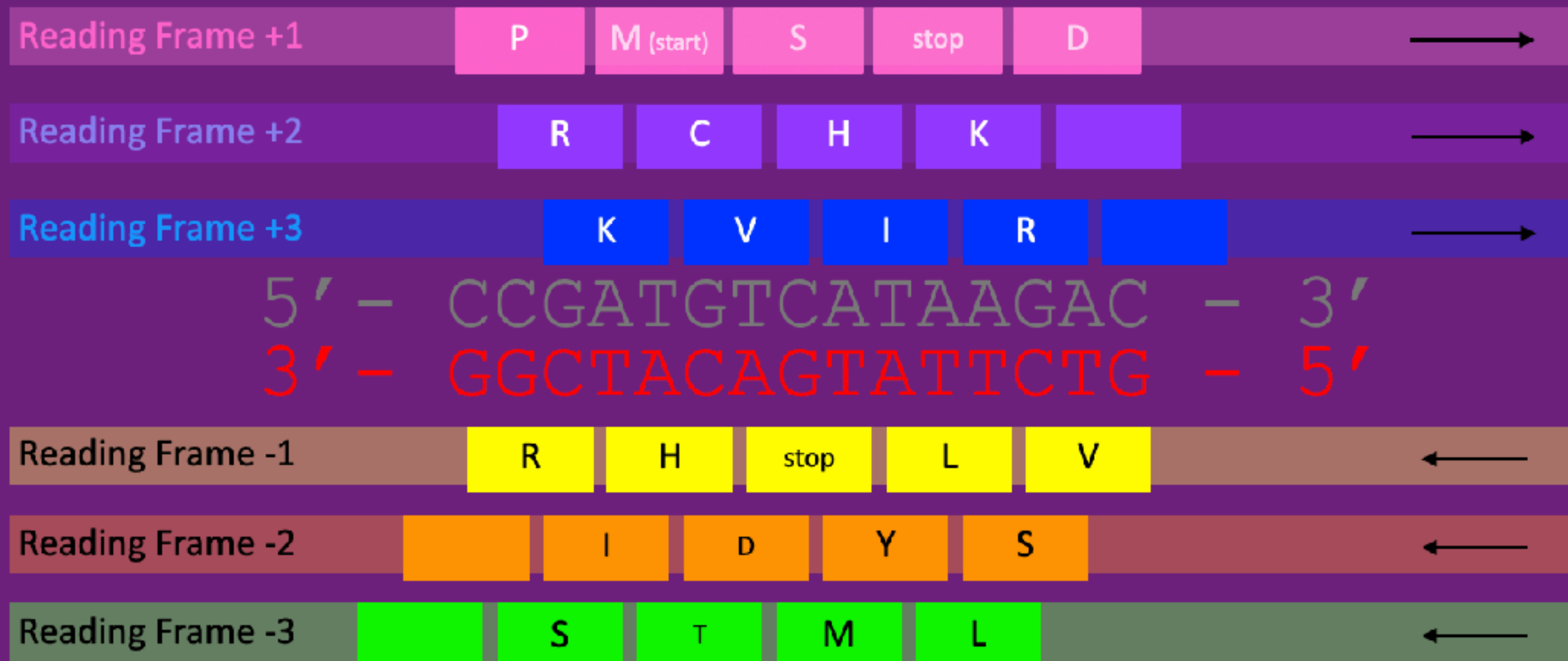
- RCHK

- KVIR

- VL-stop-HR

- SYDI

- LMTS



TRANSCRIPTIONS & TRANSLATION

- Possible Sequences

- PMS-stop-D

- RCHK

- KVIR

- VL-stop-HR

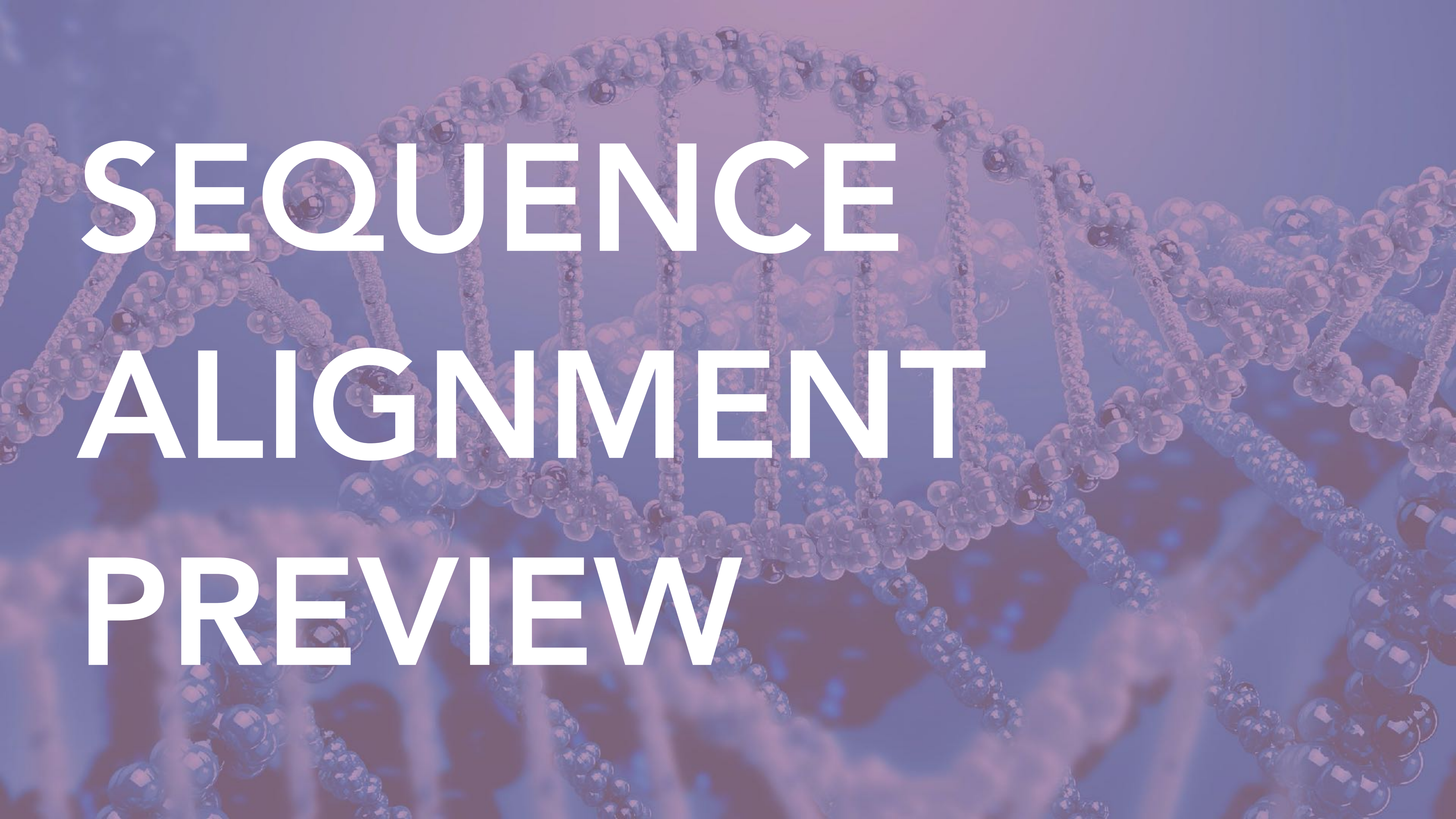
- SYDI

- LMTS

- Which one is an actual coding sequence?

- PMS-stop-D

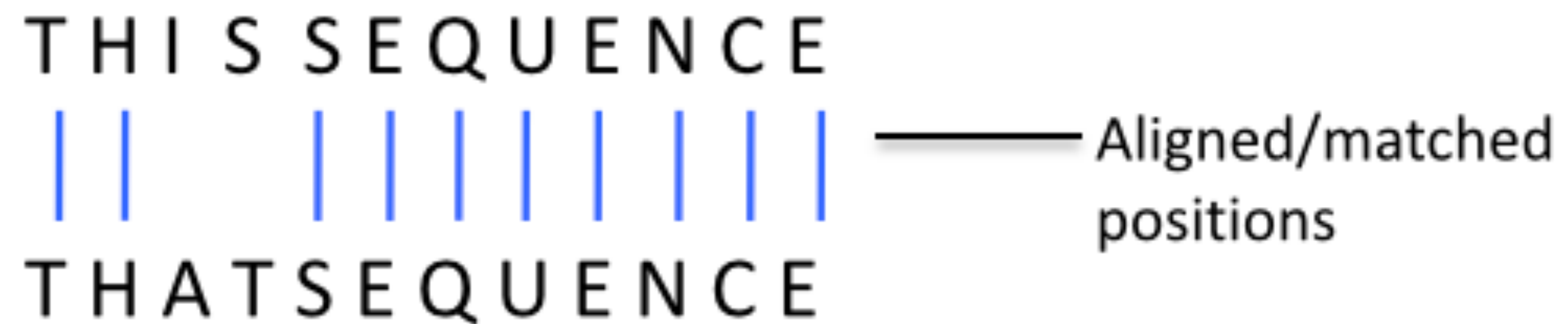
- Coding sequence has to have a start (M) and stop codon with at least one amino acid in between



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	1.05×10^{18}
10	8097453	9850349744182729	3.32×10^{26}	1.35×10^{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 optimal 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

QUICK VISUAL
REPRESENTATION

- Dynamic Programming

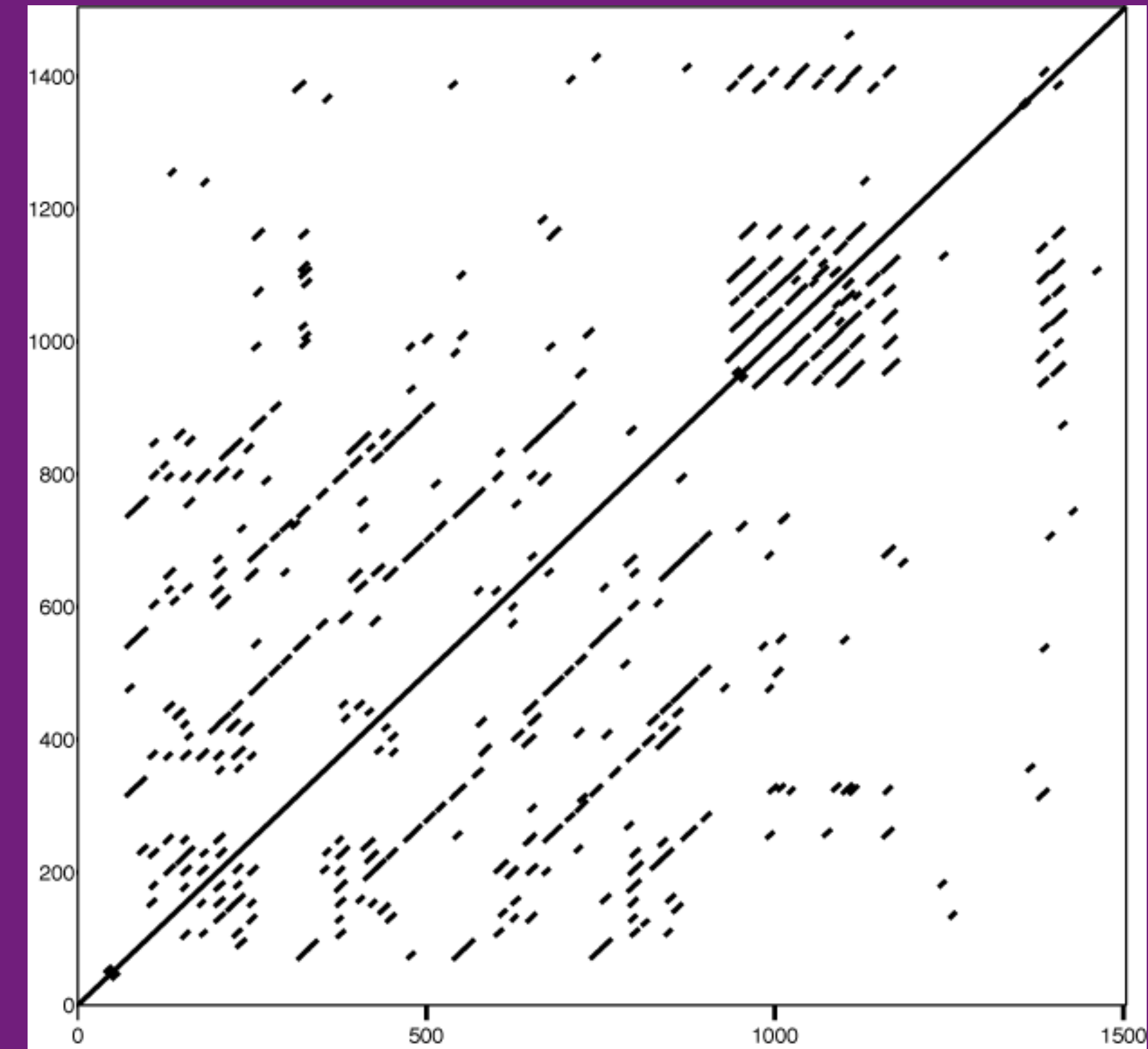
SLOW, BUT WILL
FIND OPTIMAL
ALIGNMENT

- Multiple sequence alignment

MULTISTEP,
EVOLUTIONARY
ANALYSIS

- Database

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

THAT IS A SEQUENCE

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

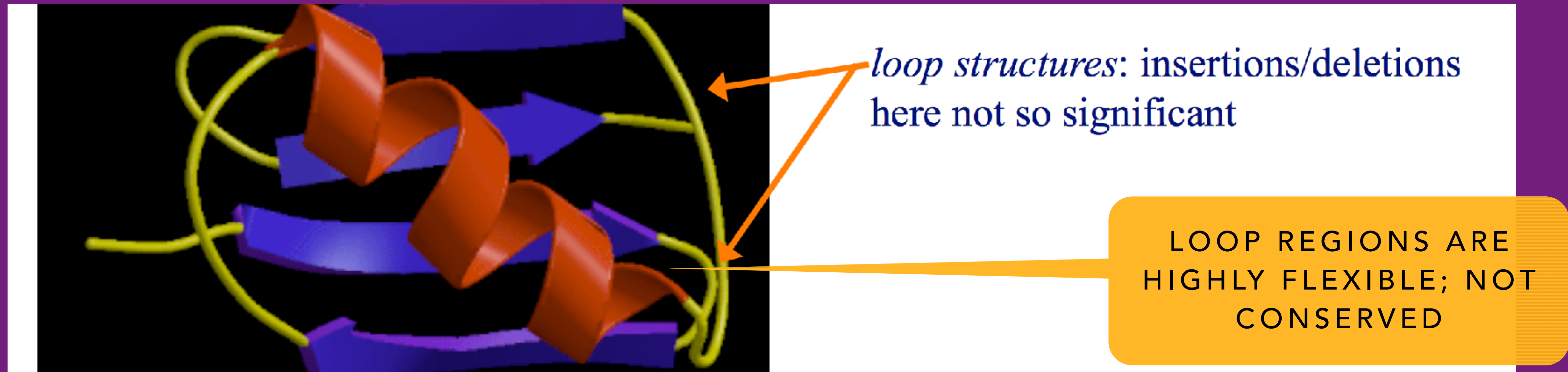
ISSUES IN ALIGNING SEQUENCES

substitutions:	ACGA	AGGA
insertions:	ACGA	ACGA
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



- 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

— — — THIS SEQUENCE
THAT IS A SEQUENCE

Alignment 1: 3 gaps, 8 matches

THIS — — — SEQUENCE
THAT IS A SEQUENCE

Alignment 2: 3 gaps, 9 matches

- 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

	A0	A4	A8	A12		B1	B6	B14	C2	CD1	CD4
	↓	↓	↓	↓		↓	↓	↓	↓	↓	↓
Hb_a	-----VL	SPADK	TNVKAA	WGKVGA	----	HAGEY	GAEAL	ERMFL	SFPTT	TKTY	FPHF
Hb_b	-----VHL	TPEEK	SAVTAL	WGKV	----	NVDEV	GGEAL	GRLLV	VYPWT	QRFF	ESF
Mb_SW	-----VL	SEGEW	QLVLHV	WAKVEA	----	DVAGH	GQDIL	IRLFK	SHPET	LEKF	DRF
LegHb	-----GAL	TESQA	ALVKSS	WEEFNA	----	NIPKH	THRFF	FILVLE	IAPAA	KDLF	SFL
BacHb	-----LDQ	TINI	IKA	TVPLKEHG	----	V-TIT	TTYK	NLFAKH	PEVR	PLF	---
SeaHb	GGTLAI	QAQGD	LTLAOK	KIVRKT	WHQLMR	----	NKTSF	VTDVF	IRIFAY	DP	SAQN
AscHb	-----	ANKTR	ELCMK	SLEHAK	VDT	SNEAR	QDGI	DLYKHM	FENYP	PLRKY	FKS
Eryt.	-----L	SADQI	STVQ	ASFDK	VKG	----	DPVG	ILYAV	FKADP	SIMAK	FTQF



ISSUES IN ALIGNING SEQUENCES

HBA_HUMAN	GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
	G+ + VK+HGKKVA++++ AH+D++ +++++LS+LH KL
HBB_HUMAN	GNPK VKAHGKKVLGAFSDGLAHLNKGTFAT LSELHCD KL
HBA_HUMAN	GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
	++ +++++H+ KV + +A ++ +L+ L+++H+ K
LGB2_LUPLU	NNPELQAHAGKVFKEYEAAIQLQVTGVVVTDATLKNLGSVHVS KG
HBA_HUMAN	GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
	GS+ + G + +D ++ H+ D+ A +ALD ++AH+
FIG11_G11.2	GSGYLVGDSLTFV DLL- -VAQHTADLLAANAALLDEFPQFKAHQE

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

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

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
A		5	-2	-2	-2	0	0	0	-2	-2	-3	-2	-1	-2	0	0	0	-2	-3	0	A
R	4		5	-2	-3	-3	0	-1	-2	0	-3	-4	1	-3	-3	-2	-2	0	0	-3	R
N	-1	5		5	0	0	0	-2	0	0	-4	-5	-2	-3	-3	-2	0	0	-2	-2	N
D	-2	0	6		5	-4	0	1	-1	0	-5	-6	-3	-4	-4	0	-2	-2	-2	-2	D
C	-2	-2	1	6		8	-2	-3	-1	-1	0	-2	-3	0	-1	-1	1	0	0	-2	C
Q	0	-3	-3	-3	9		5	2	0	0	-2	-4	0	-2	-3	0	0	0	0	-2	Q
E	-1	1	0	0	-3	5		5	0	0	-3	-4	0	-3	-3	0	0	0	-2	-3	E
G	-1	0	0	2	-4	2	5		6	0	-4	-5	-2	-3	-2	-2	0	0	0	-2	G
H	0	-2	0	-1	-3	-2	-2	6		6	-3	-4	0	-2	0	0	0	0	2	-2	H
I	-2	0	1	-1	-3	0	0	-2	8		4	0	-3	2	0	-2	-3	0	0	-3	I
L	-1	-3	-3	-3	-1	-3	-3	-4	-3	4		4	-4	0	0	-3	-4	-3	0	-4	L
K	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4		4	-2	-4	-1	-2	0	0	-3	K
M	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5		6	0	-3	-3	-2	0	-3	M
F	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5		6	-3	-2	-2	2	2	F
P	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6		7	0	0	-2	-3	P
S	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7		4	2	-2	-3	S
T	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4		5	-1	-3	T
W	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		9	2	W
Y	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		7	Y
V	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		V
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	

CCF53P62

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