Bioinformatics Feature Extraction

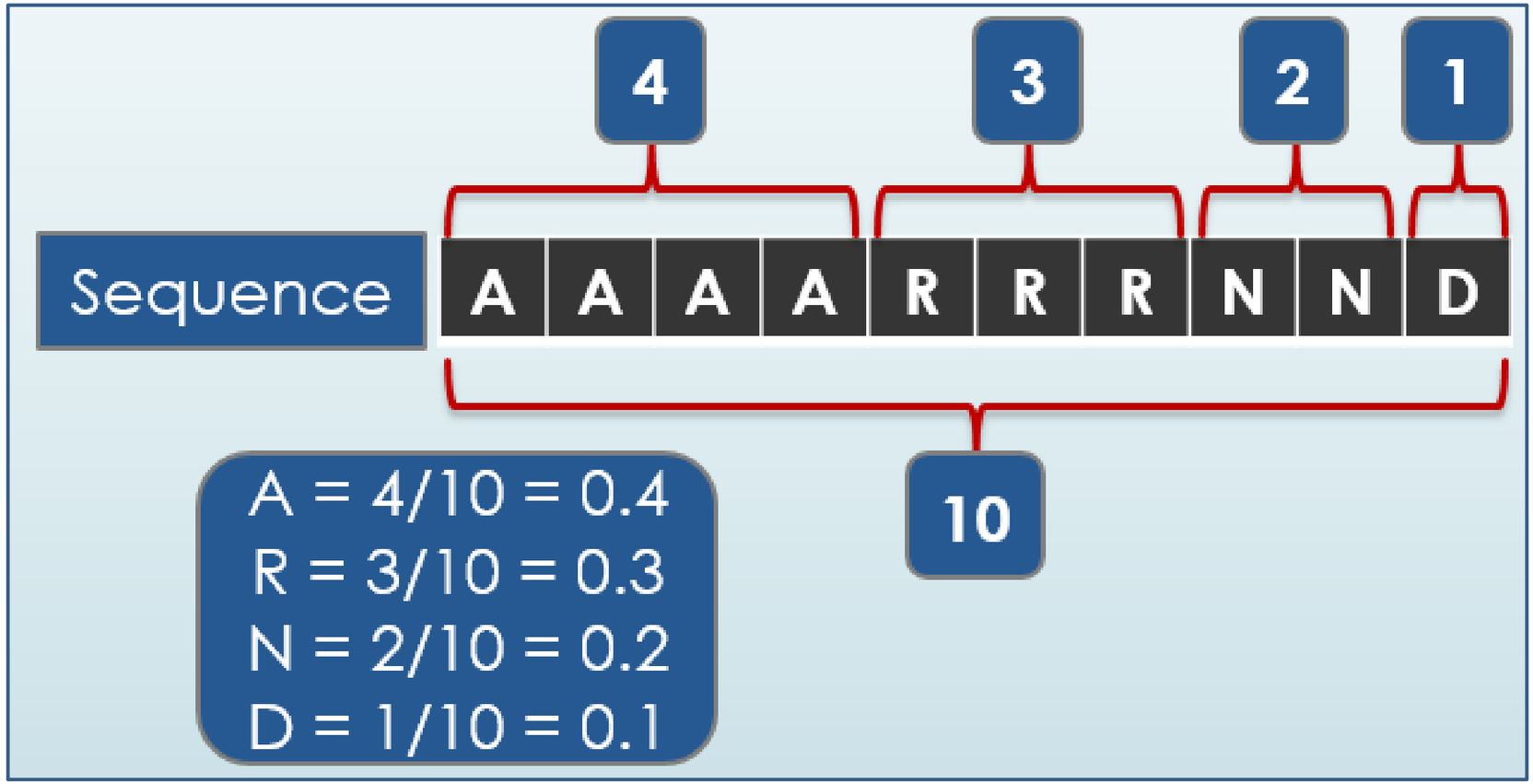
Introduction to different features in bioinformatics

Amino Acid Occurrence

- Amino acid occurrence is the number of amino acids of each type present in a protein.
- For example, the T4 lysozyme has 164 residues, and the amino acid occurrence is the information about each of the 20 amino acid residues in this protein, i.e., Ala: 15, Asp: 10, Cys: 2, etc.

Feature Extraction

- Amino Acid Composition



Feature Extraction

- Dipeptide Pair Composition (DPC)

AA = 3/10 = 0.3

AR = 1/10 = 0.1

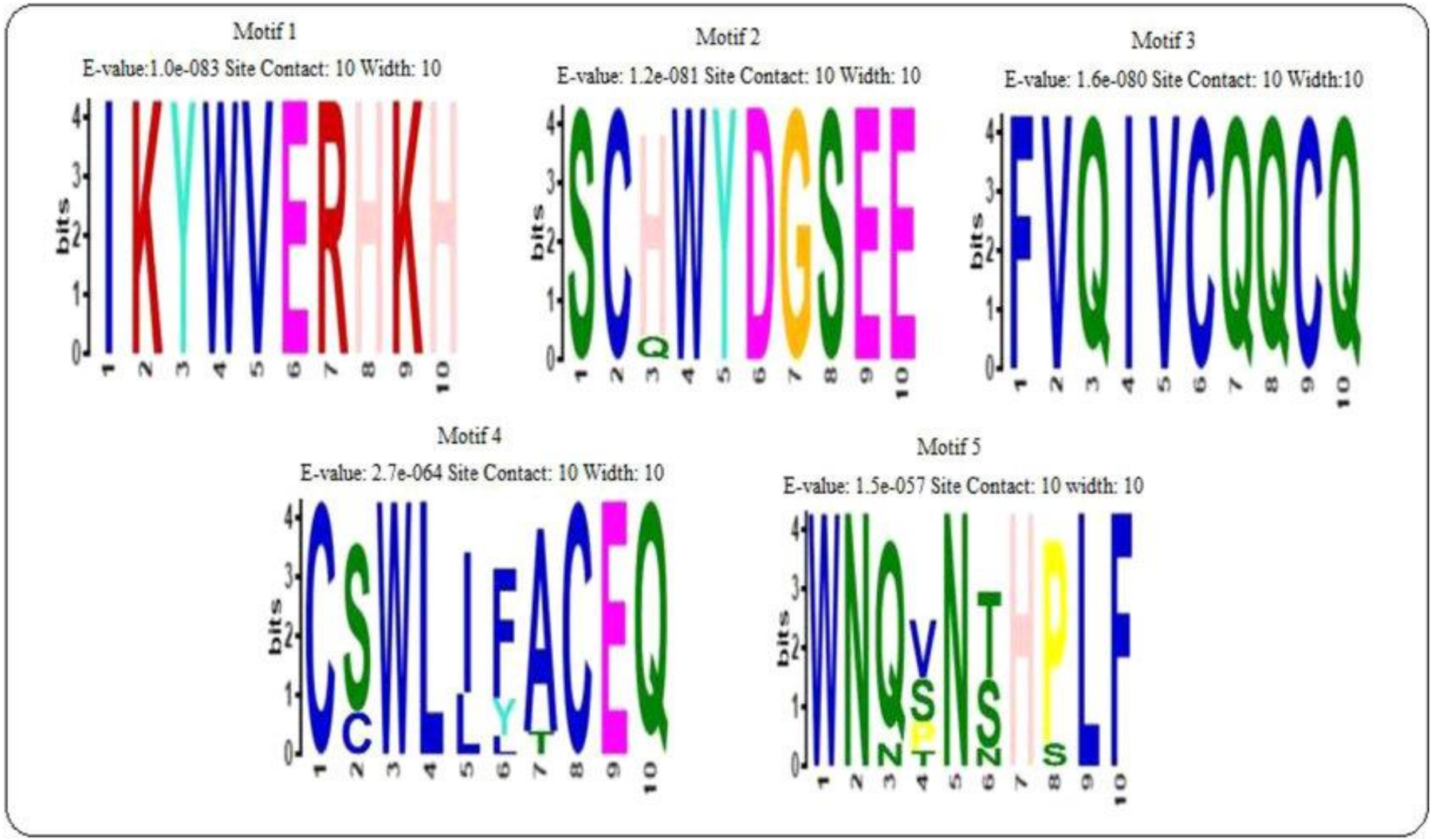
RR = 2/10 = 0.2

RN = 1/10 = 0.1

NN = 1/10 = 0.1

ND = 1/10 = 0.1

Motif Features



Moindi, A. et al. 2018. Expression of odorant co-receptor Orco in tissues and development stages of Glossina morsitans morsitans, Glossina fuscipies fuscipies and Glossina pallidipies. Scientific African. 1, (2018), e00011.

Feature Extraction

- Position Scoring Specific Matrix (PSSM)

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

A	-4	-9	-13	-15	-9	-10	-11	-16	-12	7	9	-6	7	0	-13	-9	-1	-9	-4	-9

A	-0.4	-0.9	-1.3	-1.5	-0.9	-1	-1.1	-1.6	-1.2	-0.7	-0.9	-0.6	0.7	0	-1.3	-0.9	-0.1	-0.9	-0.4	-0.9

A	0.40131234	0.289050497	.....	0.40131234	0.289050497

Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 1999, *292* (2), 195– 202

DNA Sequence Features

- DNA sequences are represented as the occurrence frequencies of k neighboring nucleic acids

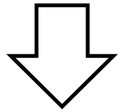
$$f(r,s) = \frac{N_{rs}}{N-1} \quad \mathrm{r,s} = 1,2,\ldots,16.$$

- N_{rs} is the number of dipeptide represented by deoxyribonucleic acid r and type s

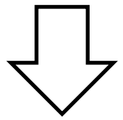
Noble WS, Kuehn S, Thurman R, Yu M, Stamatoyannopoulos J (2005). "Predicting the in vivo signature of human gene regulatory sequences." Bioinformatics, 21(suppl 1), i338–i343

2-mer

AATTCATGCGTCCGGACTTCTGCCTCGAGCCGCCGTACACTGGG
CCCTGCAAAGCTC



AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
3	3	2	2	3	6	5	6	2	7	3	2	1	5	4	2



reverse = TRUE

AA	AC	AG	AT	CA	CC	CG	GA	GC	TA
5	5	8	2	7	9	5	7	7	1