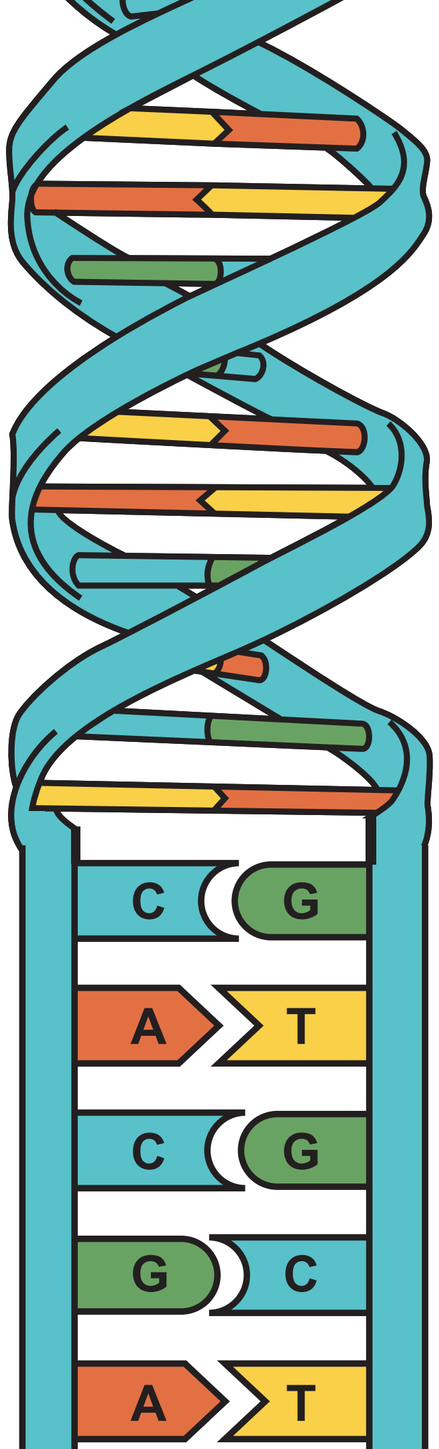


DNA SEQUENCING WITH FM-INDEX

(FULL-TEXT INDEX IN MINUTE SPACE)

TEAM-7 AMN AND RAYYAN

WHY THE FM-INDEX AND HOW IT IS APPLIED HERE.

WHERE IS DNA FOUND?

DNA IS IN EVERY CELL OF EVERY LIVING THING. IT IS FOUND WITHIN THE CHROMOSOMES OF THE CELL. CHROMOSOMES WORK TO BUILD PROTEINS AND ASSIST IN DUPLICATION OR DIVISION OF THE CELLS.

OUR PROGRAM FOCUSES ON SEARCHING PATTERNS THAT OCCUR WITHIN THE DNA SEQUENCE AND PERFORMING MUTATIONS.

WHAT DOES DNA LOOK LIKE?

ADENINE THYMINE

CYTOSINE GUANINE

THE BASES OF DNA PAIR WITH EACH OTHER IN A PREDICTABLE WAY.

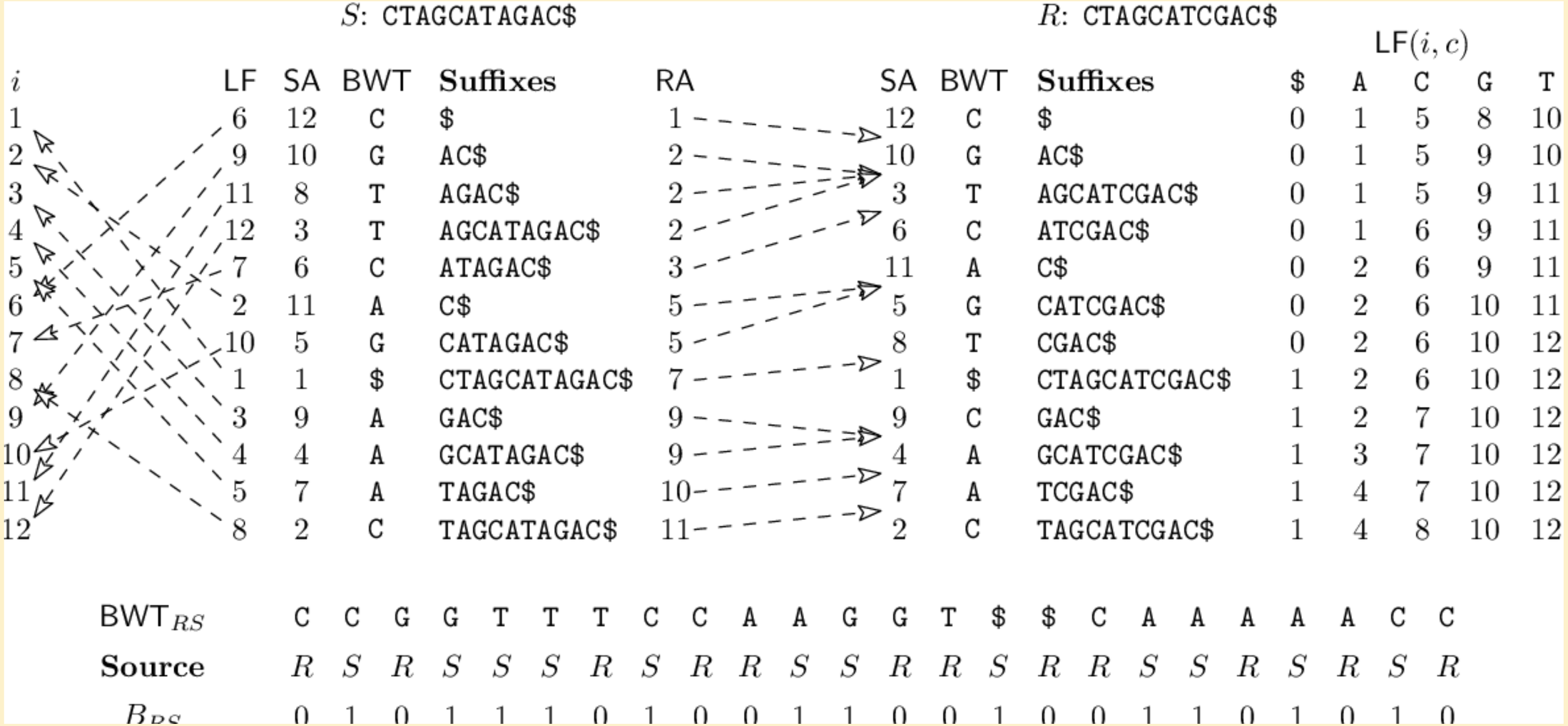
A ALWAYS PAIRS WITH T

C ALWAYS PAIRS WITH G

HOW DOES DNA WORK?

THE 4 LETTERS OF DNA MAKE UP CODONS. THESE CHEMICALS ARE REPEATED IN VARIOUS ORDERS OVER AND OVER. **THESE CODONS WHICH ARE READ IN SETS OF THREE MAKE UP AMINO ACIDS WHICH MAKE UP GENES**. THESE GENES TELL CELLS HOW TO MAKE A PROTEIN THAT CONTROLS EVERYTHING IN THE CELL.

FM-INDEX MODE OF FUNCTIONALITY



S: CTAGCATAGAC$

i	LF	SA	BWT	Suffixes
1	6	12	C	$
2	9	10	G	AC$
3	11	8	T	AGAC$
4	12	3	T	AGCATAGAC$
5	7	6	C	ATAGAC$
6	2	11	A	C$
7	10	5	G	CATAGAC$
8	1	1	$	CTAGCATAGAC$
9	3	9	A	GAC$
10	4	4	A	GCATAGAC$
11	5	7	A	TAGAC$
12	8	2	C	TAGCATAGAC$

RA

1
2
2
2
3
5
5
7
9
9
10
11

R: CTAGCATCGAC$

SA	BWT	Suffixes
12	C	$
10	G	AC$
3	T	AGCATCGAC$
6	C	ATCGAC$
11	A	C$
5	G	CATCGAC$
8	T	CGAC$
1	$	CTAGCATCGAC$
9	C	GAC$
4	A	GCATCGAC$
7	A	TCGAC$
2	C	TAGCATCGAC$

LF(i, c)

$	A	C	G	T
0	1	5	8	10
0	1	5	9	10
0	1	5	9	11
0	1	6	9	11
0	2	6	9	11
0	2	6	10	11
0	2	6	10	12
1	2	6	10	12
1	2	7	10	12
1	3	7	10	12
1	4	7	10	12
1	4	8	10	12

BWT_{RS}	C	C	G	G	T	T	T	C	C	A	A	G	G	T	$	$	C	A	A	A	A	A	C	C
Source	R	S	R	S	S	S	R	S	R	R	S	S	R	R	S	R	R	S	S	R	S	R	S	R
B_{RS}	0	1	0	1	1	1	0	1	0	0	1	1	0	0	1	0	0	1	1	0	1	0	1	0

IMAGE FROM SEMANTIC SCHOLAR

OUR PROGRAM USES THE FM-INDEX AND ITS BACKWARD SEARCH FEATURE ALONGSIDE THE SUFFIX ARRAY, COUNT TABLES, AND OCCURRENCE RANKS TO LOCATE EXACTLY WHERE A QUERIED PATTERN OCCURS. THIS QUERIED PATTERN'S INDICES IN THE LARGER SEQUENCE ARE RETURNED AND THEN USED FOR MUTATIONS.

THE FM INDEX'S EFFICIENCY IS MORE APPARENT AT A LARGER SCALE WHERE COMPLETE SEQUENCES ARE USED. FOR OUR IMPLEMENTATION WE HAVE LIMITED IT TO SEQUENCES NO LONGER THAN 243 BASES