3 - Overlaps

Monday, 11 October 2021 11:05

WEEK 3

CLOSEST MATCH LOCATED IN FINAL ROW

HORIZONTAL & VERTICAL

+1 TO ACCOUNT FOR FINAL GAP

TRACEBACK - FOLLOWING MATRIX WHERE YOU STARTED

$O(p \times t)$ APPROXIMATE MATCHING

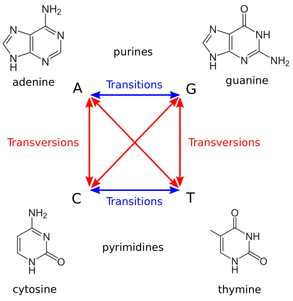
$O(t)$ EXACT MATCHING

DYNAMIC PROGRAMMING FOR EDIT DISTANCE

GLOBAL & LOCAL ALIGNMENT

IN OUT EDIT DISTANCE EVERY EDIT HAS THE SAME WEIGTH

WE CAN PENALIZE EDIT BASED ON HOW RARE THEY ARE



PROBABILISTICALLY TRANSV SHOULD BE x2 TRANSI

BUT IN REAL LIFE IT'S THE OTHER WAY AROUND

HUMAN SUBSTITUTION RATE 1/1000

SMALL GAP RATE 1/3000

INSTEAD OF +1 WE LOOK AT A $\sigma(x)$ FUNCTION
THAT GIVES THE CORRESPONDING PENALITY

REALLY LONG COMPUTATIONALLY

$100 \times 3{,}2 \cdot 10^{5}$ MATRIX!

WE NEED AN EFFICIENT ALGORITHM
WE CANNOT SPEND A YEAR ANALIZING
THE RESULT OF JUST 1 WEEK OF WORK

INDEXES

DON'T DEAL VERY WELL W/ MISMATCHES & GAPS

INDEX ARE BETTER SUITED FOR EXACT MATCH PROBLEM

FROM READS TO GENOME

ALIGNMENT # ASSEMBLY

IN ASSEMBLY YOU DON'T HAVE A REFERENCE GENOME

DE-NOVO SHOTGUN ASSEMBLY PROBLEM

↓ ↓

FROM SCRATCH

RANDOM BITS FROM GENOME

THIS PROBLEM IS MORE DIFFICULT AND COMPUTATIONALLY INTENSIVE

COVERAGE - AMOUNT OF REDUNDANT INFO

NOT AL SEQUENCES AGREE ON THE BASES

AVERAL COVERAGE WE HAVE TO KNOW LEN(GENOME)

IF THE 2 SEQUENCES ARE SIMILAR, IT COULD BE A HINT
THAT THE 2 READS MIGHT HAVE ORIGINATE FROM THE
SAME LOCATION IN THE GENOME

OVERLAP

IF SUFFIX & PREFIX ARE SIMILAR :
2 SEQUENCES CAN OVERLAP

DIFFERENCES IN BASES

- MEASUREMENT ERROR
- BASECALLER SOFTWARE ERROR

POLYPLOIDY

2 CHROMOSOMES ARE DIFFERENT

MORE COVERAGE LEADS TO MORE & LONGER OVERLAPS

OVERLAPS ARE THE GLUE THAT HELP US IN OUR ASSEMBLY

OVERLAPPING (DIRECT) GRAPH

NODE = CONCEPT
EDGE = RELATIONSHIP

THRESHOLD TO OVERLAP