4 - De-novo Assembly

Monday, 11 October 2021 11:05

SHORTEST COMMON SUPERSTRING PROBLEM (SCS)

GIVEN A COLLECTION OF STRINGS, FIND SHORTEST STRING THAT CONTAINS THEM ALL

SIMILAR TO RECONSTRUCTING GENOME FROM READS

THIS PROBLEM IS NP COMPLETE

WE DON'T HAVE EFFICIENT ALGORITHMS FOR BIG DATA

PROCESS:

TRY ALL ORDERINGS (PERMUTATIONS)

GIVE 2 STRINGS TOGETHER ACCORDING TO MAX OVERLAP

FIND THE SHORTEST GLUED-STRING

N! ALL ORDERINGS
TOO MUCH INFO TO PROCESS INTRACTABLE

GREEDY SCS

FAST BUT SOLUTION IS NOT GUARANTEED TO BE THE OPTIMAL ONE

MAKES DECISION TO OPTIMIZE ONLY ONE PARAMETER

GREEDY IS FAST BUT NOT ALWAYS CORRECT
THE OTHER ALGO IS NOT FEASABLE

- FIND 2 STRINGS W BEST OVERLAP
- REPLACE 2 STRINGS W/ OVERLAPPED ONE
- DO THIS UNTIL ONE REMAINS (SCS)

LAWS OF ASSEMBLY

1 SAME SUFFIX & PREFIX = POSSIBLE OVERLAP

2 MORE COVERAGE = MORE & BETTER OVERLAPS

3 REPEATS MAKE ASSEMBLY DIFFICULT

SCS DOES NOT WORK IF THE GENOME IS REPETITIVE

SCS CLUSTERS THE REPETITIVE PART INTO FEWER COPIES

AMBIGUITY SOMETIME IT IS NOT POSSIBLE TO RECONSTRUCT THE GENOME

REPEATS MAKE ASSEMBLY DIFFICULT

45% OF HUMAN DNA IS REPETITIVE (TRANSPOSONS)

SCS WAS OUR FIRST ATTEMP TO FORMULATE A COMPUTATIONAL PROBLEM THAT WHEN SOLVED IT CAN ALLOW US TO ASSEMBLY THE GENOME FROM SCRATCH

DE BRUIJN GRAPHS & EULERIAN WALKS

DIRECTED GRAPH ⇆
EDGES HAVE A DIRECTION (MULTIGRAPH)

ASSUMPTIONS

SEQ READS CONSIST OF K-MER
LEFT & RIGHT (K-1) -MER

GENOME RECONSTRUCTION IS OBTAINED BY A SIMPLE WALK ACROSS THE GRAPH

CROSSING ALL EDGES EXACTLY ONCE

NOTE NOT ALL GRAPHS ARE EULERIAN

WHEN EULERIAN WALK GOES WRONG

DE BRUIJN GRAPH SOLVE THE PROBLEMS OF REPETITION
WE CAN NOW RECONSTRUCT THE ORIGINAL GENOME

EULERIAN WALK GIVES YOU YOUR INITIAL GENOME

IF REPETITIVE GENOME WE CAN HAVE MORE THAN ONE VALID EULERIAN WALK

ONLY ONE WALK CORRESPONDS TO ORIGINAL
THE OTHERS ARE INCORRECT RESHUFFLES

IF YOU DECREASE THE K-MER LENGTH YOU INCREASE THE CHANCE TO BE AFFECTED BY REPEATS

SMALLER K ↘
MORE K-MER IN GENOME ↘
MORE REPEATS ↘
MORE EULERIAN WALKS ↓

ASSUMPTION ABOUT SEQ DATA

IF THERE ARE SOME ERRORS IN THE SEQUENCING THE RESULTING GRAPH WILL NOT BE EULERIAN

ASSEMBLERS IN PRACTICE

DE BRUIJN IS THE STANDARD

SCS IS A PLANED FORMULATION OF THE ASSEMBLY PROBLEM, BUT STILL A USEFUL ONE

WHAT MAKE THE ASSEMBLY MESSY

- SEQUENCING ERRORS
THEY CREATE DEAD ENDS IN GRAPH

- REDUNDANT NODES
SOME OVERLAPS MAY IMPLY ANOTHER

- POLYPLOIDY
NATURAL VARIANCE; YOUR COPIES OF THE SAME CHROMOSOME MAY BE DIFFERENT

EVERY ASSEMBLY IS FRAGMENTED

EVEN THE MOST STUDIED (HUMAN REF-GENOME) HAS HOLES IN IT

FUTURE ?

WE CAN OVERCOME III LAW BY MAKING THE READS LONGER

IT IS LIKE MAKING THE PUZZLE PIECES BIGGER
THE OVERALL PUZZLE WILL BE EASIER

LONGER READS PREVENT COLLAPSING

THEY ANCHOR REPETITIVE-SEQ TO THEY SURROUNDING NON-REPETITIVE CONTEXT

HOW TO GET LONGER READS ?

NEW SEQUENCING TECH

EX. PAIR-END REA

SEQUENCING 1 MOLECULE AT A TIME

COMPUTER SCIENCE & LIFE-SCIENCE

COMPUTATIONAL GENOMICS
COMPUTATIONAL BIOLOGY
GENOMICS DATA SCIENCE

ANALYZING HIGH THROUPUT DATA

COMPUTER SCIENTISTS CAN BE LEADERS IN LIFE SCIENCE AS WELL

LOADS OF PEOPLE IN THE 'HUMAN-GENOME-PROJECT' WERE COMPUTER SCIENTISTS

SHARED ABSTRACTIONS

HOW CAN I TURN A BIOLOGICAL PROBLEM INTO A PLAIN STRING PROBLEM ?

THANKS

JACOB PATT
IRA GOODING
JESSICA CROWL
JEFF LEAK
BRIAN CATHO
ROGER TANG