

## Computer Science and Biology(cont.)

Jun 11, 2007  
KangKang Yin

## fragment assembly

- ▶ how might you assemble the following fragments?

AATCTG  
CCGCAA  
ATCTGTAAATCCG  
CTGTAAAT

## example (continued)

- ▶ one possible arrangement:

AATCTG    CCGCAA  
          TGTAATCCG  
          CTGTAAAT

- ▶ this yields the strand(length 17)

AATCTGTAAATCCGCAA

## example (continued)

- ▶ another possible arrangement:

          AATCTG  
CCGCAA    TGTAATCCG  
          CTGTAAAT

- ▶ this yields the strand(length 18)

CCGCAATCTGTAAATCCG

## Philosophy of piecing

- ▶ Look for longest overlaps. Long sequences of repetition are unlikely to occur unless they are the overlaps introduced by the cleavage
- ▶ Choose the shortest assembled strand

## Developing the algorithm

Equivalent to the TSP

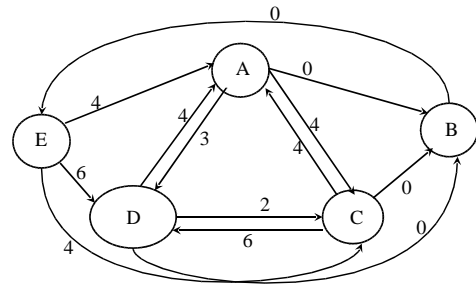


## TSP(Traveling Salesman Problem)

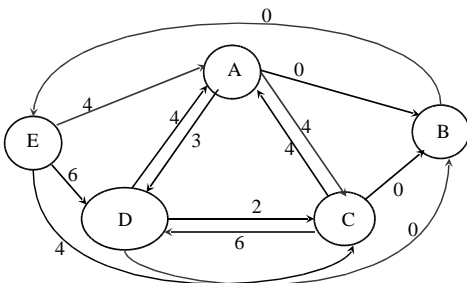
► Given a number of cities and the costs of traveling from any city to any other city, what is the cheapest round-trip route that visits each city exactly once and then returns to the starting city?

- weighted graph: edges in the graph are labeled with costs
- a *tour* : must visit each node exactly *once*, and end at its starting point
- find the *cheapest* tour in a given graph

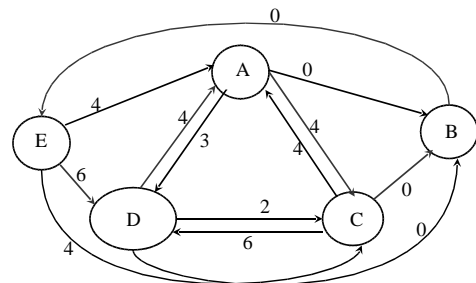
## TSP instance



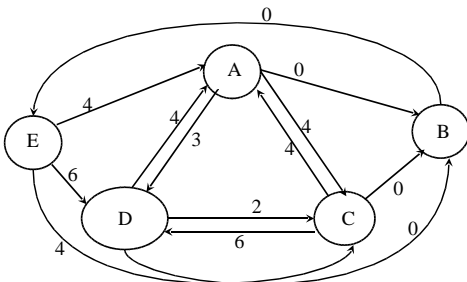
## one possible tour



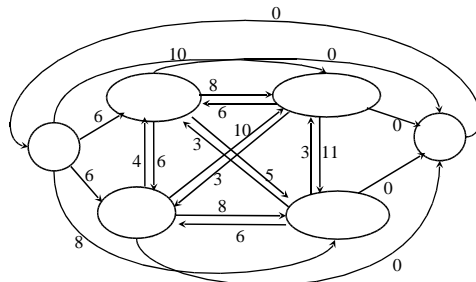
## another possible tour



## which is the cheapest tour?



## Another TSP Instance



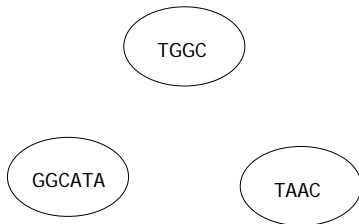
## NP-hard

- ▶ no-one has found an algorithm that will quickly solve the TSP on all graphs
- ▶ \$1M prize available for anyone who finds a fast algorithm for TSP, or proves that no good algorithm exists!
- ▶ still, decades of research have yielded algorithms that work reasonably well on many practical graphs

## back to fragment assembly

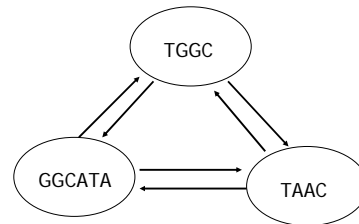
- ▶ how might you find the shortest assembled strand?  
AATCTG  
CCGCAA  
TGTAATCCG  
CTGTAAT
- ▶ TSP in disguise

## Example



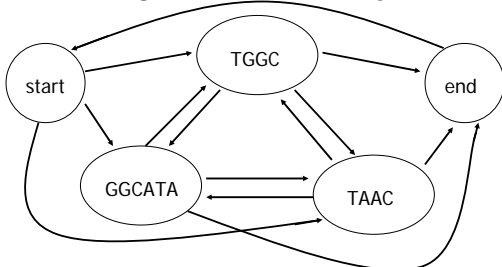
one node per fragment

## from fragment assembly to TSP



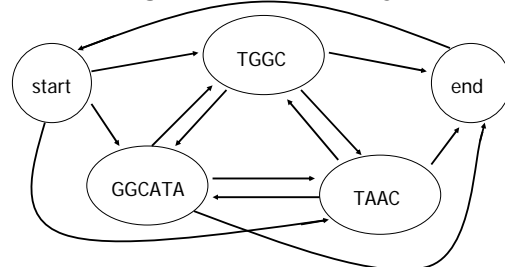
add a directed edge between each pair of nodes

## from fragment assembly to TSP

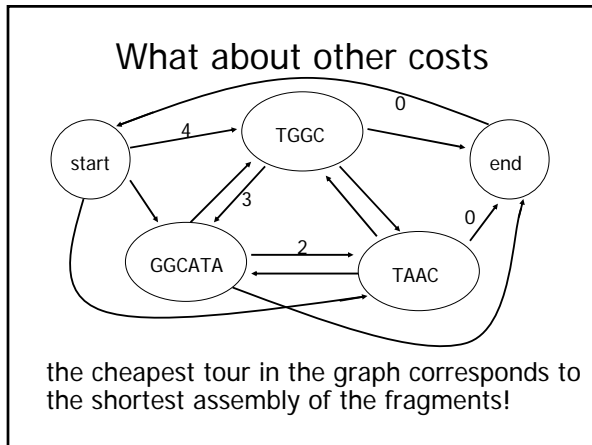
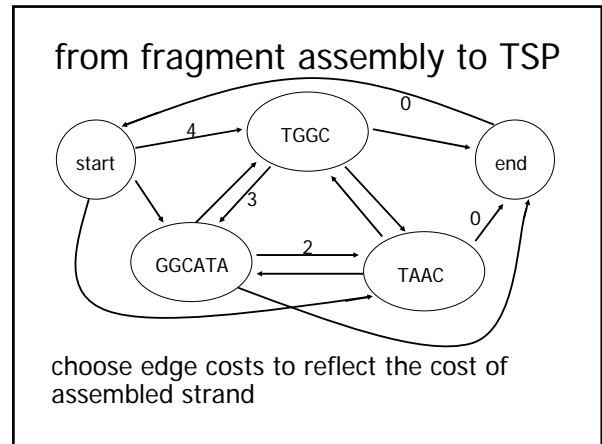
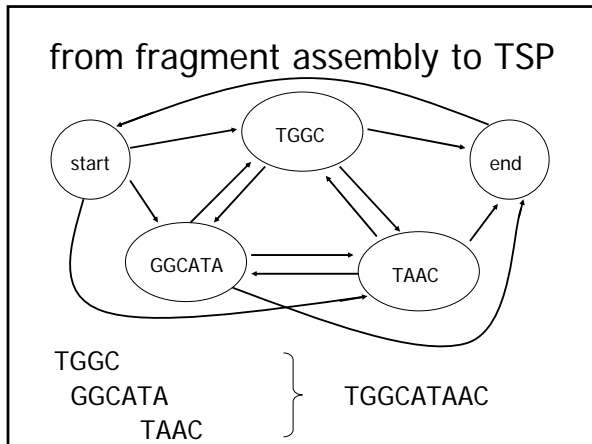


- ▶ a *start* node with an edge to all nodes
- ▶ an *end* node with an edge from all nodes
- ▶ An edge from *end* to *start*

## from fragment assembly to TSP



- ▶ a tour, starting at *start*, corresponds to an assembly of the fragments



### Now try this

► construct the TSP instance corresponding to the following fragment assembly instance

AATCTG  
CCGCAA  
TGTAATCCG  
CTGTAAT

### Summary: fragment assembly

- an important step in sequencing a genome
- can be solved by computer, using algorithms for a seemingly unrelated problem - the TSP

### Computational complexity: P vs. NP

- the TSP is an example of the famous so-called 'NP-hard' problems for which no fast algorithms (in polynomial time) are known
- Solving TSP
  - brute force search  $O(n!)$
  - dynamic programming
  - heuristic search
    - genetic algorithms
    - simulated annealing
    - Tabu search...