Sequence Assembly

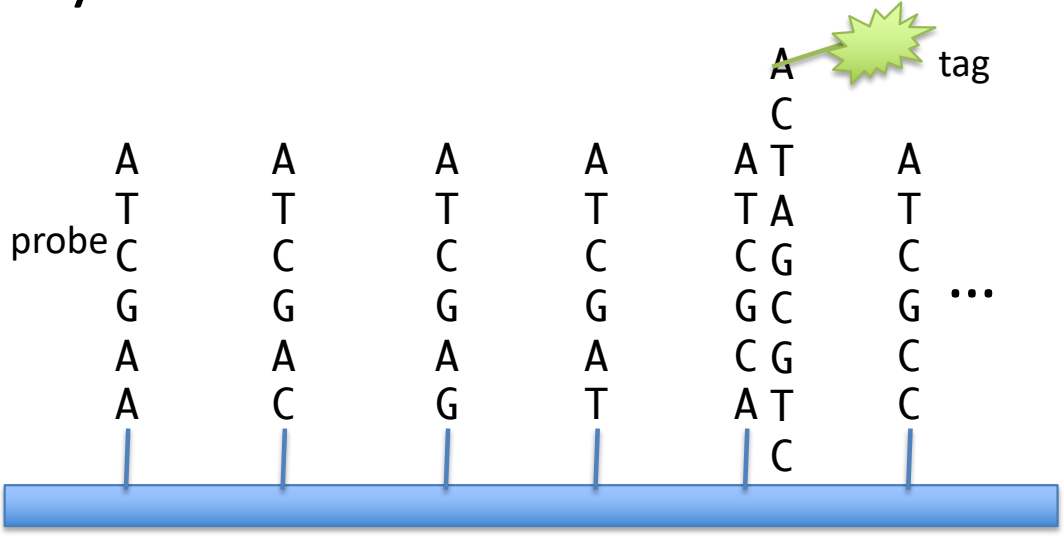
Spectral assembly

Outline

- Sequencing by hybridization (SBH) approach
- SBH data and computational task
- Eulerian paths and graphs
- Eulerian-cycle based algorithm for SBH
- SBH shortcomings in practice
- Spectral assembly with read data

Universal DNA arrays

- Array with all possible oligonucleotides (short DNA sequence) of a certain length as probes
- Sample is labeled and then washed over array
- Hybridization is detected from labels



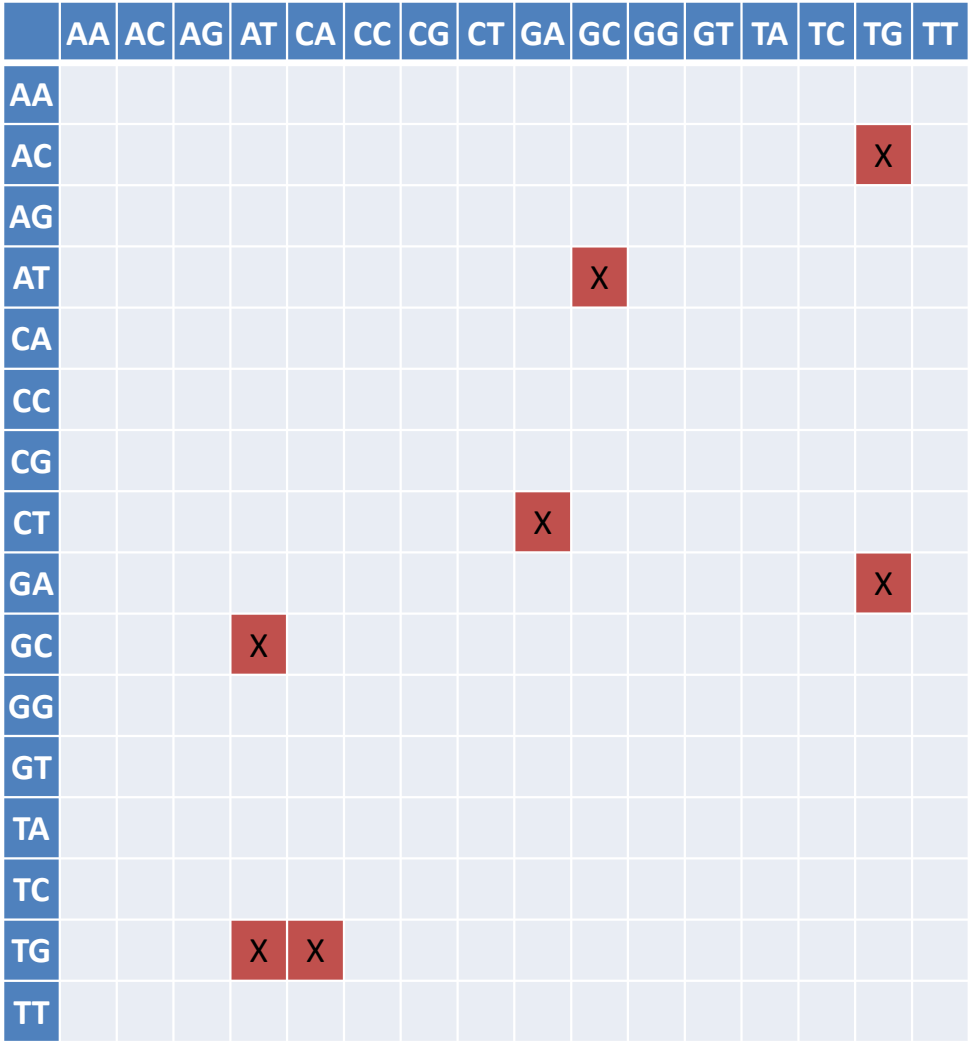
Sequencing by Hybridization (SBH)

- SBH array has probes for all possible *k*-mers
- For a given DNA sample, array tells us whether each *k*-mer is *PRESENT* or *ABSENT* in the sample
- The set of all *k*-mers present in a string *s* is called its *spectrum*
- Example:
  - s = ACTGATGCAT
  - *spectrum(s, 3)* = {ACT, ATG, CAT, CTG, GAT, GCA, TGA, TGC}

Example SBH Array

Sample:
ACTGATGCAT

Spectrum (k=4):
{ACTG, ATGC,
CTGA,GATG,
GCAT,TGAT,
TGCA}



SBH Problem

- Given: *k* and a set *S* of *k*-mers
- Do: Find a string *s*, such that *spectrum(s,k) = S*

{ACT, ATG, CAT, CTG, GAT, GCA, TGA, TGC}

?

SBH task vs. shortest superstring

- SBH task looks similar to the shortest superstring problem
- Similarities
  - Input is a set of strings, output is a single string
  - All input strings must be substring of solution
- Differences
  - All input strings in SBH have length = k
    - Input strings can have variable lengths for shortest superstring
  - All length k substrings of SBH output must be in input
    - Shortest superstring output can have substrings that are not in input

SBH as Eulerian path

- Could use Hamiltonian path approach, but not useful due to *NP*-completeness
- Instead, use *Eulerian* path approach
- *Eulerian path*: A path through a graph that traverses every edge exactly once
- Construct graph with all *(k-1)-mers* as vertices
- For each *k*-mer in spectrum, add edge from vertex representing *first k-1* characters to vertex representing *last k-1* characters

Properties of Eulerian graphs

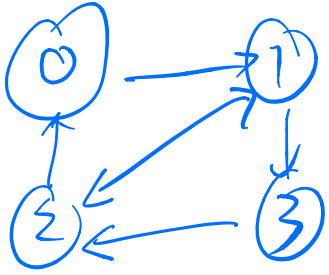
- It will be easier to consider *Eulerian cycles*: Eulerian paths that form a cycle
- Graphs that have an *Eulerian cycle* are simply called *Eulerian*
- ***Theorem:*** A connected directed graph is *Eulerian* if and only if each of its vertices are *balanced*
- A vertex *v* is *balanced* if *indegree(v) = outdegree(v)*
- There is a polynomial-time algorithm for finding Eulerian cycles!

Seven Bridges of Königsberg

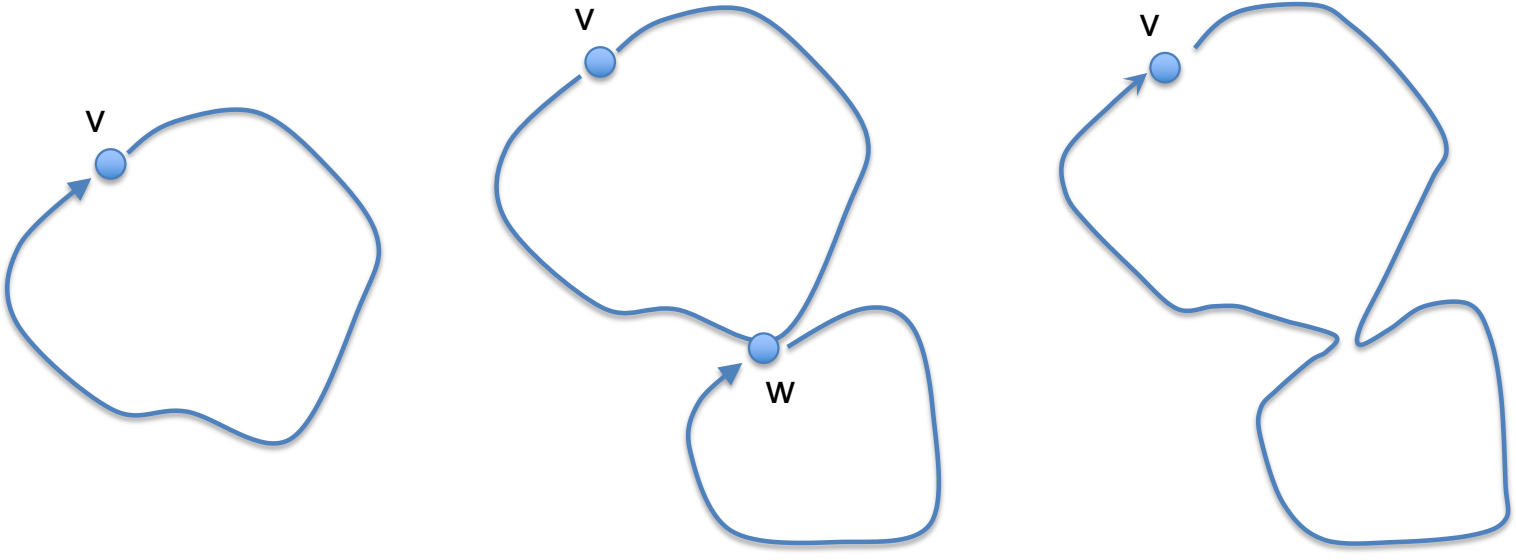
Euler answered the question: “Is there a walk through the city that traverses each bridge exactly once?”

Eulerian cycle algorithm

- Start at any vertex *v*, traverse unused edges until returning to *v*
- While the cycle is not Eulerian
  - Pick a vertex *w* along the cycle for which there are untraversed outgoing edges
  - Traverse unused edges until ending up back at *w*
  - Join two cycles into one cycle



Joining cycles

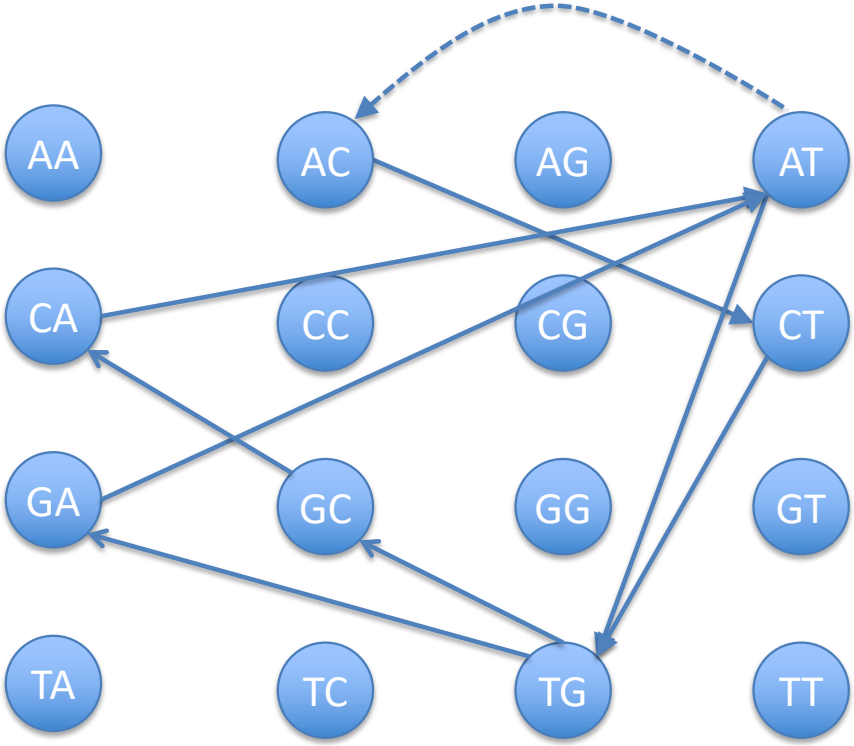


Eulerian Path -> Eulerian Cycle

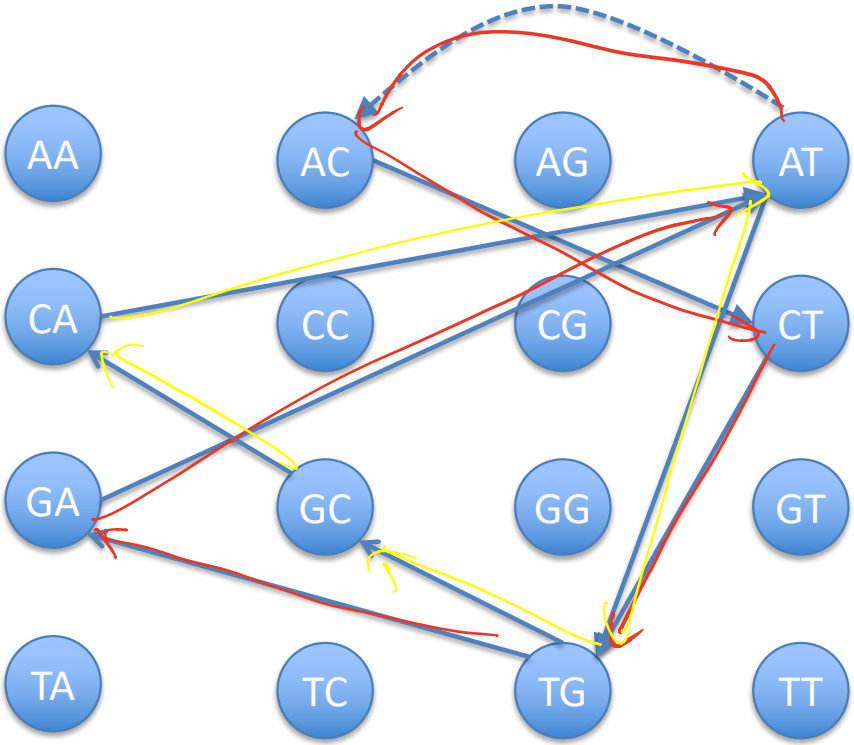
- If a graph has an Eulerian Path starting at *s* and ending at *t* then
  - All vertices must be balanced, except for *s* and *t* which may have $|indegree(v) - outdegree(v)| = 1$
  - If and *s* and *t* are not balanced, add an edge between them to balance
    - Graph now has an Eulerian cycle which can be converted to an Eulerian path by removal of the added edge

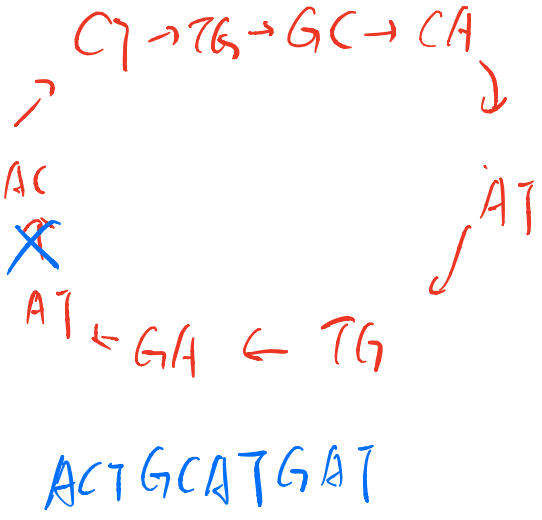
SBH graph example

{ACT, ATG, CAT, CTG, GAT, GCA, TGA, TGC}



Eulerian cycle algorithm example





SBH difficulties

- In practice, sequencing by hybridization is hard
  - Arrays are often inaccurate -> incorrect spectra
    - False positives/negatives
  - Need long probes to deal with repetitive sequence
    - But the number of probes needed is exponential in the length of the probes!
    - There is a limit to the number of probes per array (currently between 1-10 million probes / array)

K-mer spectrum approach with read data (de Bruijn approach)

- Generate spectrum from set of all *k*-mers contained within reads
- Choose *k* to be small enough such that the majority of the genome's *k*-mers will be found within the reads
- Particularly useful for short-read data, such as that produced by Illumina
- Made popular by methods such as Euler and Velvet

Summary

- The SBH task can be cast as finding an Eulerian cycle in a certain graph
- There exists an efficient algorithm for finding Eulerian cycles in graphs
- Unfortunately, it is not feasible to obtain ideal SBH data
- Fortunately, algorithmic insights from the spectral assembly approach can be applied to shotgun sequencing assembly data