On NP-Hardness of the Paired de Bruijn Sound Cycle Problem

Evgeny Kapun Fedor Tsarev

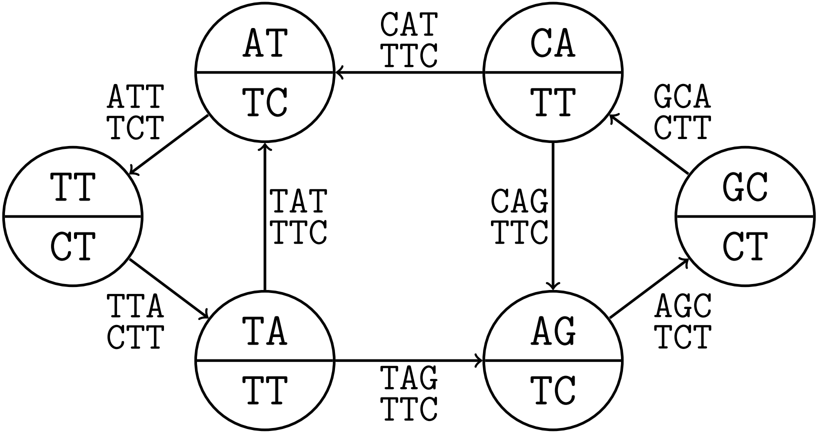
Genome Assembly Algorithms Laboratory
University ITMO, St. Petersburg, Russia

September 2, 2013

Genome assembly models

- Shortest common superstring – NP-hard.
- Shortest common superwalk in a de Bruijn graph – NP-hard.
- Superwalk in a de Bruijn graph with known edge multiplicities – NP-hard.
- Path in a paired de Bruijn graph?

Paired de Bruijn graph



Sound path

Sound path: strings match with shift $d = 6$.

TAGCTCACCCGTTGGT

ACCCGTTGGTAATTGC

Sound cycle: cyclic strings match with shift $d = 6$.

TGATAAGTAGGCTAAG

GTAGGCTAAGTGATAA

Paired de Bruijn Sound Cycle Problem

Given a paired de Bruijn graph G and an integer d (represented in unary coding), find if G has a sound cycle with respect to shift d.

Paired de Bruijn Covering Sound Cycle Problem

Given a paired de Bruijn graph G and an integer d (represented in unary coding), find if G has a *covering* sound cycle with respect to shift d.

Parameters

- $|\Sigma|$: size of the alphabet.
- k: length of vertex labels.
- $|V|$, $|E|$: size of the graph (bounded in terms of $|\Sigma|$ and k).
- d: shift distance.

Simple cases

- $|\Sigma| = 1$: at most one vertex, at most one edge.
- $k = 0$: at most one vertex, at most $|\Sigma|^2$ edges, reduces to the problem of computing strongly connected components.
- d is fixed: find a cycle in a graph of $|V||\Sigma|^d$ states.

Interesting case: $k = 1$

With fixed $k = 1$, the problem is NP-hard.
Proof outline:

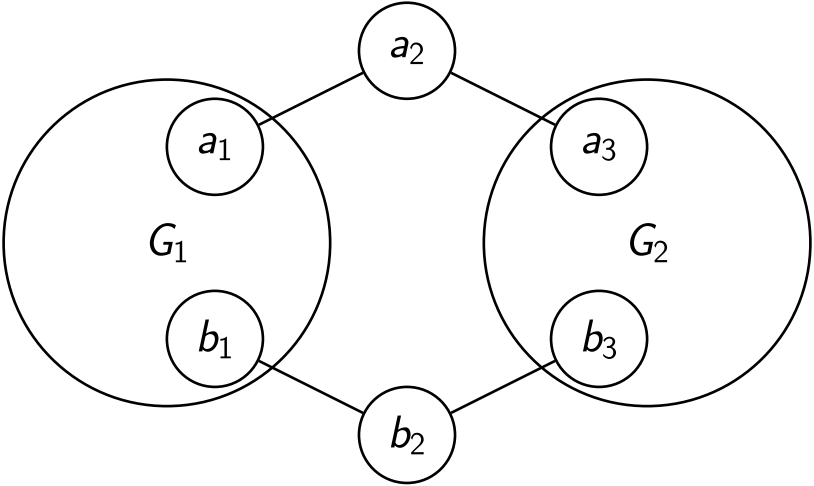
1. Reduce Hamiltonian Cycle Problem, which is NP-hard, to an intermediate problem.
2. Reduce that problem to Paired de Bruijn (Covering) Sound Cycle Problem.

The intermediate problem

Given an undirected graph,

- if it contains a hamiltonian cycle, output 1.
- if it doesn't contain hamiltonian paths, output 0.
- otherwise, invoke undefined behavior.

Solution, step 1

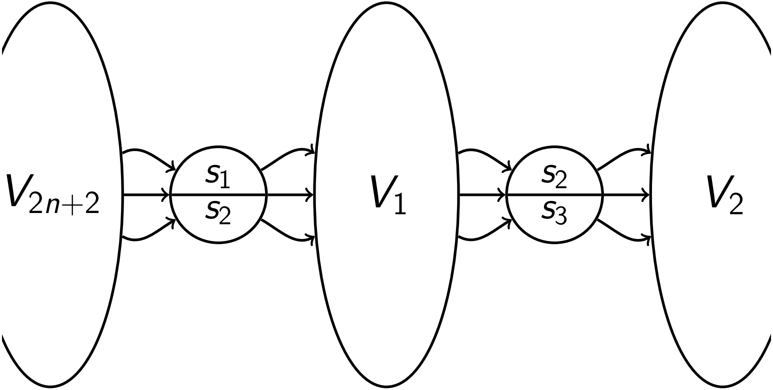


Properties of a graph with a hamiltonian cycle

Such graph contains hamiltonian paths

- ending at any vertex.
- passing through any edge.
- for any edge $\{i, j\}$ and vertex $k \neq i, j$, passing through $\{i, j\}$ such that j is between i and k on the path.

Solution, step 2

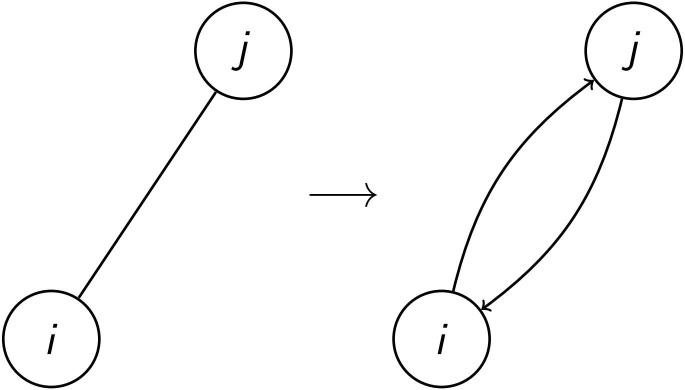


$\ldots\ldots s_1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots s_2 \ldots\ldots$

$\ldots\ldots s_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots s_3 \ldots\ldots$

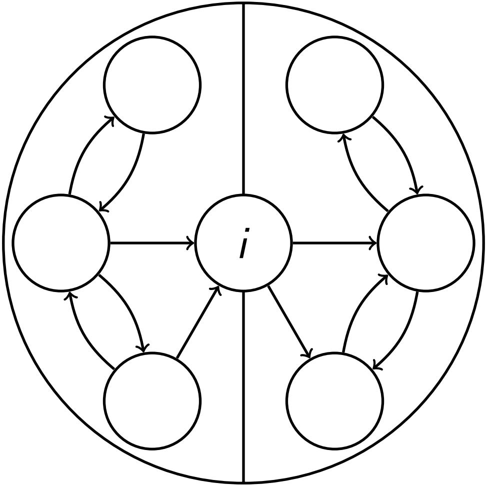
Solution, step 2

In V_1, V_{2n+2}:



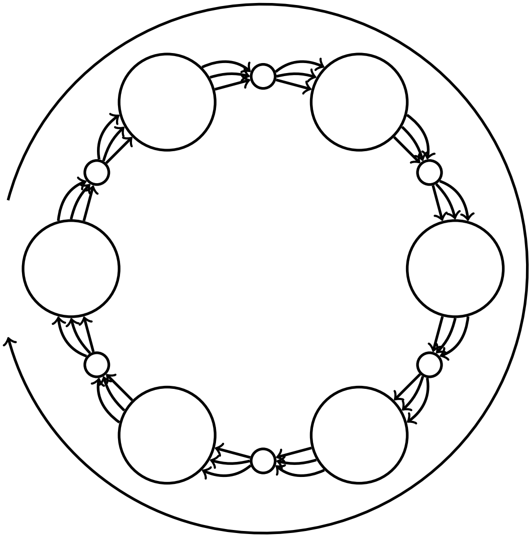
Solution, step 2

In $V_3 \ldots V_{2n+1}$:



How to make a covering cycle

Pass along the loop multiple times, covering more and more edges with each iteration.



Interesting case: $|\Sigma| = 2$

With fixed $|\Sigma| = 2$, the problem is NP-hard. The proof is done by reduction from the case $k = 1$. The characters are replaced with binary sequences, and a transformation is done to avoid undesired overlaps.

Fix both k and $|\Sigma|$

If both k and $|\Sigma|$ are fixed, the number of different de Bruijn graphs is finite. The only argument which can take infinitely many values is d, and it is an integer represented in unary coding. As a result, the number of valid problem instances having any fixed length is bounded. So, the language defined by the problem is sparse. Therefore, the problem is *not* NP-hard unless P=NP.

Results

Paired de Bruijn (Covering) Sound Cycle Problem is

- NP-hard for any fixed $k \geq 1$ (can be reduced from $k = 1$).
- NP-hard for any fixed $|\Sigma| \geq 2$ (trivially reduced from $|\Sigma| = 2$).
- NP-hard in the general case.

Results

Paired de Bruijn (Covering) Sound Cycle Problem is

- Not NP-hard if both k and $|\Sigma|$ are fixed, unless P=NP.
- Solvable in polynomial time if $k = 0$, $|\Sigma| = 1$, or d is fixed.

Thank you!