Shortest common superstring

Given set of strings S, find $SCS(S)$: shortest string containing the strings in S as substrings

S: BAA AAB BBA ABA ABB BBB AAA BAB

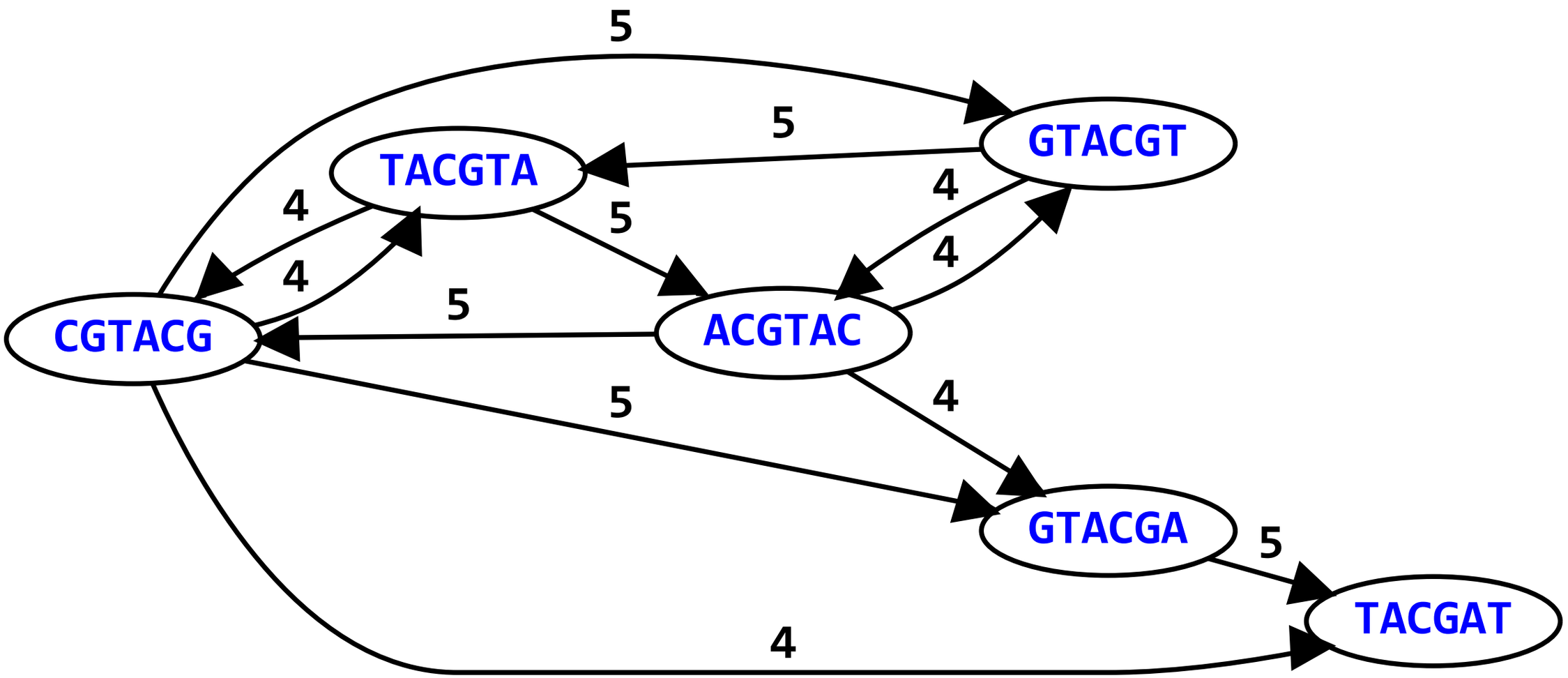
Concat(S): BAAAABBBAABAABBBBBBAAABAB

24

SCS(S): AAABBBABAA

10

Reads: all 6-mers from GTACGTACGAT




```
>>> scs(['GTACGT', 'TACGTA', 'ACGTAC',
         'CGTACG', 'GTACGA', 'TACGAT'])
'GTACGTACGAT'
```

Shortest common superstring

NP-complete: no efficient algorithms for large inputs

Idea: pick order for strings in S *and* construct superstring

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAA

Idea: pick order for strings in *S and* construct superstring

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAAB

Idea: pick order for strings in *S and* construct superstring

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAABA

Idea: pick order for strings in *S and* construct superstring

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAABABB

Idea: pick order for strings in *S and* construct superstring

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAABABBAABABBABBB ⟵ superstring 1

Idea: pick order for strings in *S and* construct superstring

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAABABBAABABBABBB ⟵ superstring 1

order 2: AAA AAB ABA BAB ABB BBB BAA BBA

AAABABBBAABBA ⟵ superstring 2

Try all possible orderings and pick shortest superstring

If *S* contains *n* strings, *n* ! (*n* factorial) orderings possible