Kako složiti genomsku slagalicu od milion delova?

Grafovski algoritmi

Bioinformatics Algorithms:
an Active Learning Approach
~Poglavlje 3~

Pregled

- Šta je sekvencioniranje genoma?
- Eksplozija u štampariji
- Problem rekonstrukcije niske
- Rekonstrukcija niske kao problem Hamiltonove putanje
- Rekonstrukcija niske kao problem Ojlerove putanje
- De Brujinovi grafovi
- Ojlerova teorema
- Spajanje parova očitavanja
- U realnosti

Genom

- Genom jednog organizma predstavlja njegov genetski materijal
- Kod većine organizama, genetski materijal je sadržan u DNK
- Kod čoveka, genom sadrži oko tri milijarde nukleotida
- Genomi nekih organizama su i 100 puta veći od humanog genoma

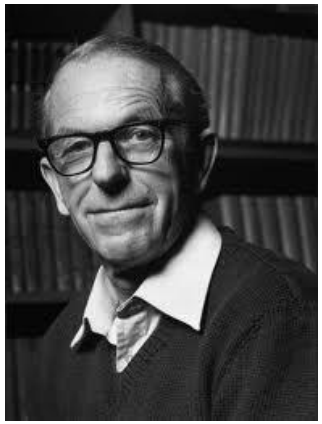
Amoeba Dubia
~ 670 milijardi

Paris Japonica
~ 150 milijardi

Kratka istorija sekvencioniranja genoma

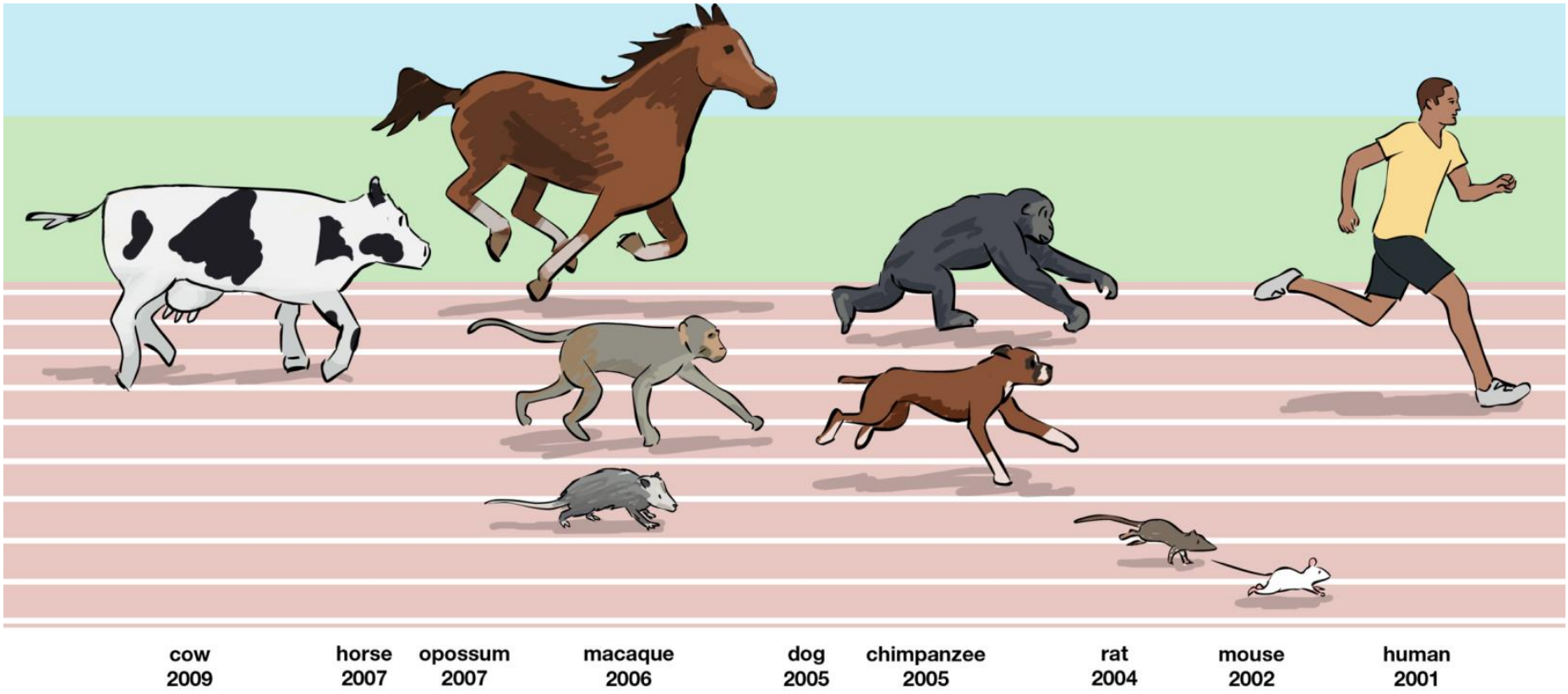
- **1977**: Walter Gilbert i Frederick Sanger razvijaju nezavisne metode sa sekvencioniranje DNK
- **1980**: Podelili su Nobelovu nagradu.
- Njihove metode za sekvencioniranje su bile veoma skupe ($3 milijarde za sekvencioniranje humanog genoma).

Walter Gilbert



Frederick Sanger

Početak sekvencioniranje genoma



- Krajem 2000-tih Sanger metodom je sekvencioniran veliki broj genoma
- Visoka cena je bila ograničavajući faktor i za dalji napredak je bila neophodna nova tehnologija sekvencioniranja

Sekvencioniranje nove generacije

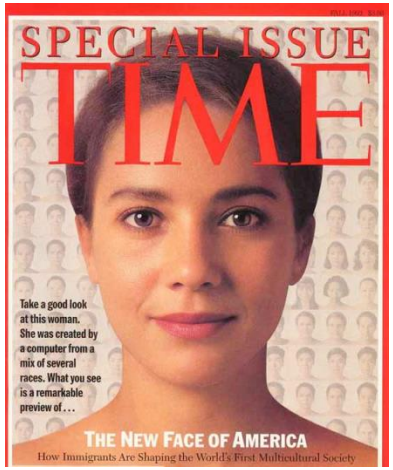
- *Next Generation Sequencing* (NGS)
- **Krajem 2000-tih**: Na tržištu se pojavljuju nove mašine za sekvencioniranje
  - *Illumina* smanjuje trošak sekvencioniranja humanog gemona sa 3 milijarde na 10 hiljada dolara
  - Kompanija *Complete Genomics* otvara genomsku fabriku u Silikonskoj dolini koja sekvencionira stotine genoma mesečno
  - Pekinški genomski institut (BGI - *Beijing Genome Institute*) preuzima *Complete Genomics* 2013. godine i postaje najveći svetski centar za sekvencioniranje genoma

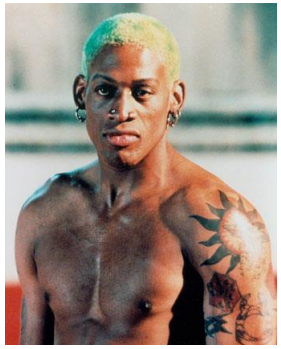
illumina

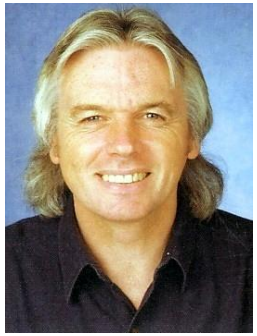
Complete genomics

华大基因 BGI

Sekvencioniranje ličnih genoma







Sekvencioniranje ličnih genoma

- Genomi se kod različitih ljudi razlikuju na malom broju pozicija (u proseku sadrže jednu mutaciju na hiljadu nukleotida)
- Ova razlika je odgovorna za različite visine kod ljudi, da li će imati sklonost ka visokom holesterolu ili ne, za veliki broj genetskih bolesti, ...

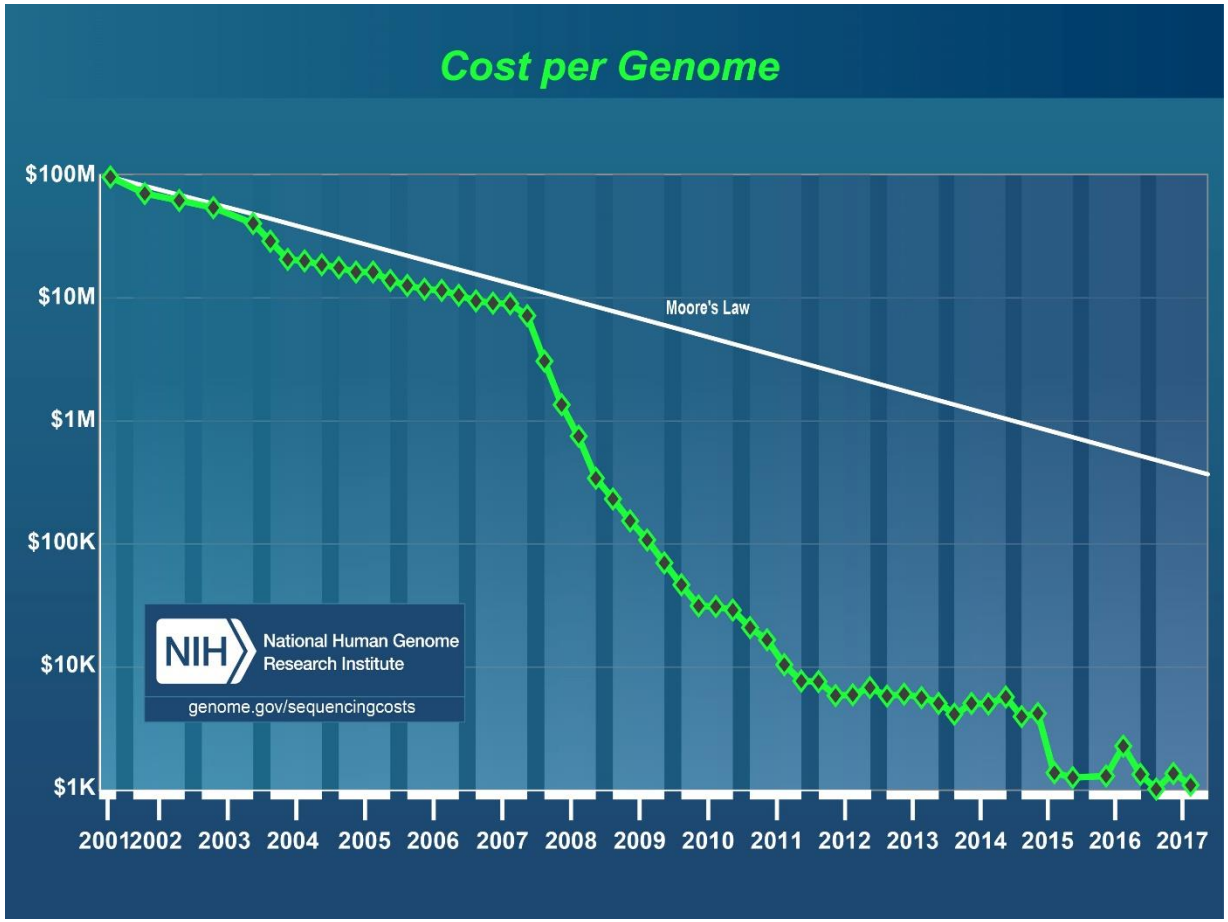
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGA
TCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTAT
CGATCGATCGATCGATTATCTACGATCGATCGATCGATCA
CTATACGAGCTACTACGTACGTACGATCGCGGGACTATTA
TCGACTACAGATAAAACATGCTAGTACAACAGTATACATA
GCTGCGGGATACGATTAGCTAATAGCTGACGATATCCGAT

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGA
TCAGCTACAACATCGTAGCTACGATGCATTAGCAAGCTAT
CGATCGATCGATCGATTATCTACGATCGATCGATCGATCA
CTATACGAGCTACTACGTACGTACGATCGCGTGACTATTA
TCGACTACAGATGAAACATGCTAGTACAACAGTATACATA
GCTGCGGGATACGATTAGCTAATAGCTGACGATATCCGAT

Sekvencioniranje ličnih genoma

- **2010:** *Nicholas Volker* je postao prvo ljudsko biće čiji je život spašen zahvaljujući genomskom sekvencioniranju
  - Lekari nisu mogli da postave tačnu dijagnozu i morali su da ga podvrgnu velikom broju operacija pokušavajući da je utvrde
  - Sekvencioniranje je otkrilo retku mutaciju na jednom genu (*XIAP)* koja je bila povezana sa oštećenjem njegovog imunog sistema
  - Ovo otkriće je navelo lekare na adekvatnu terapiju koja je rešila problem

Sekvencioniranje ličnih genoma



Pregled

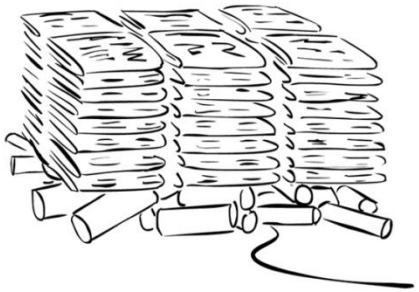
- Šta je sekvencioniranje genoma?
- **Eksplozija u štampariji**
- Problem rekonstrukcije niske
- Rekonstrukcija niske kao problem Hamiltonove putanje
- Rekonstrukcija niske kao problem Ojlerove putanje
- De Brujinovi grafovi
- Ojlerova teorema
- Spajanje parova očitavanja
- U realnosti

Problem novina

stack of NY Times, June 27, 2000

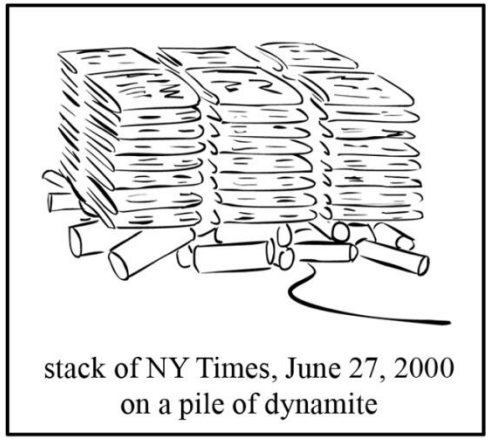
Problem novina

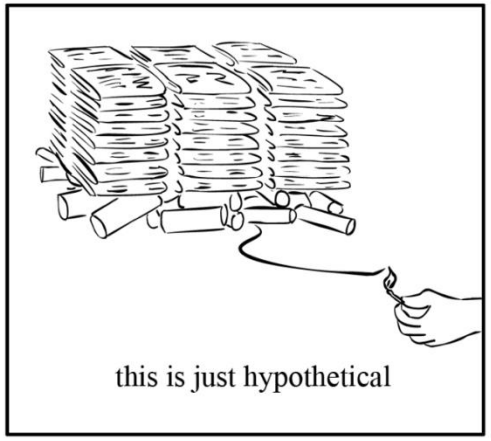
stack of NY Times, June 27, 2000



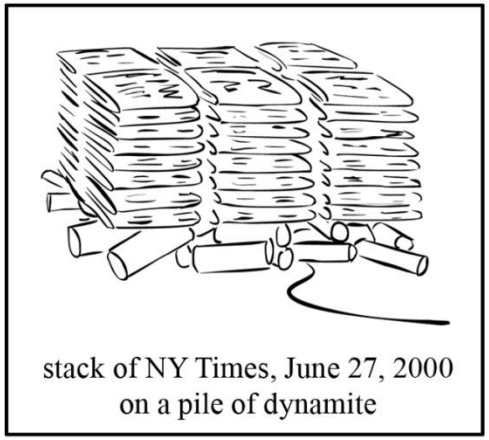
stack of NY Times, June 27, 2000
on a pile of dynamite

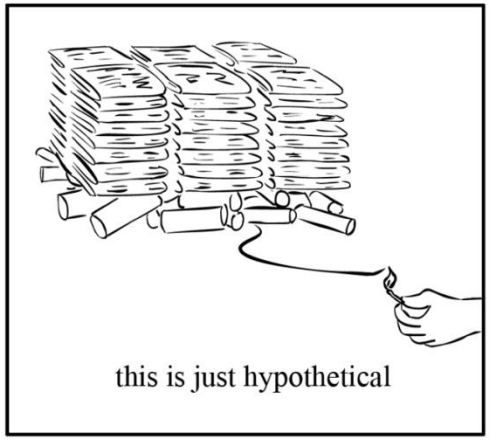
Problem novina

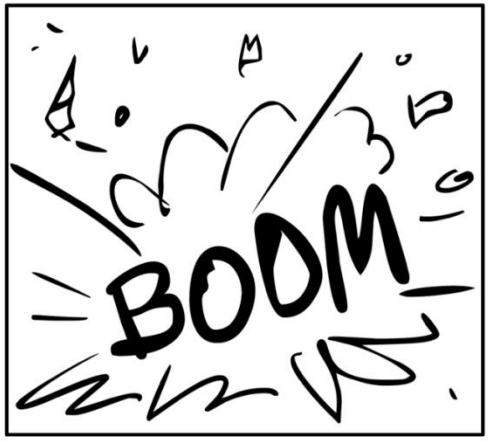




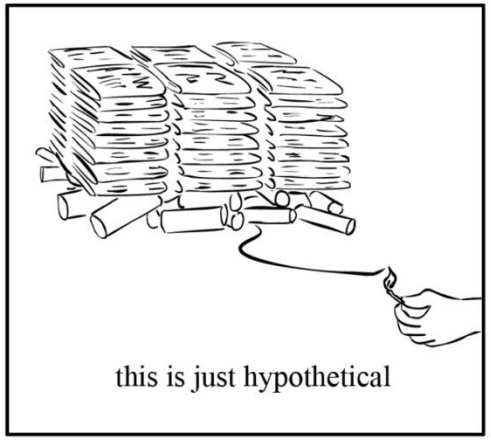
Problem novina

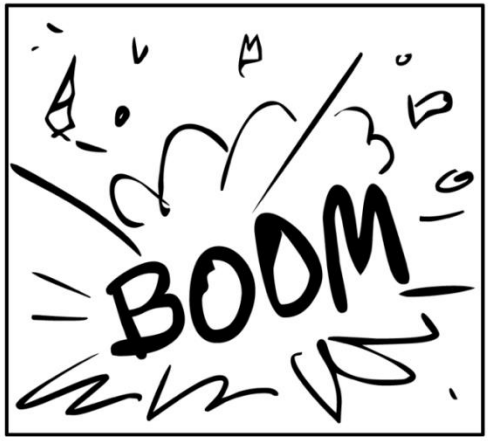


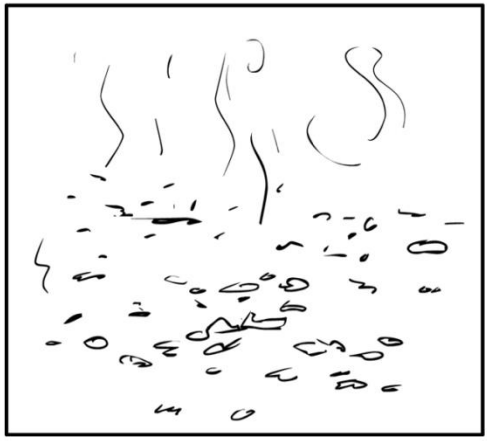




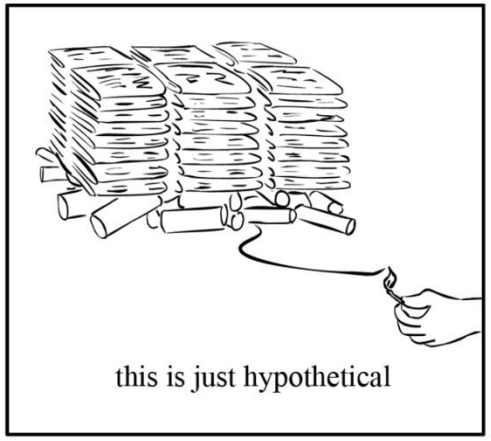
Problem novina

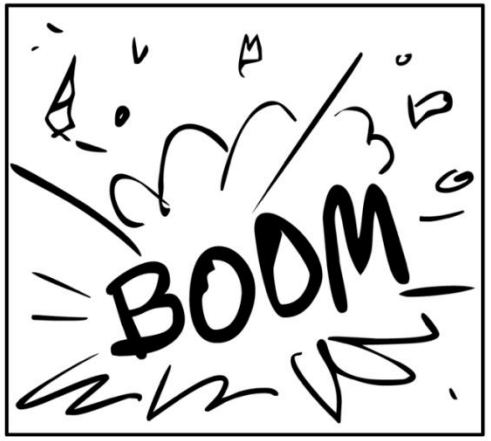


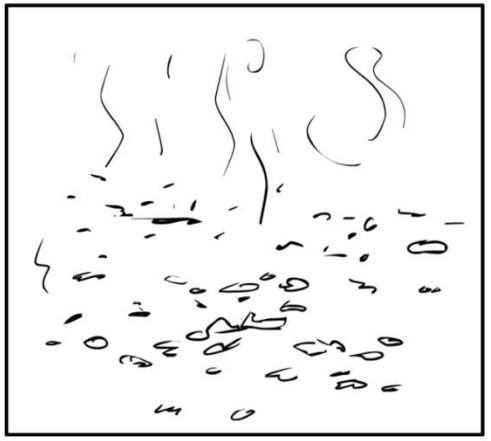


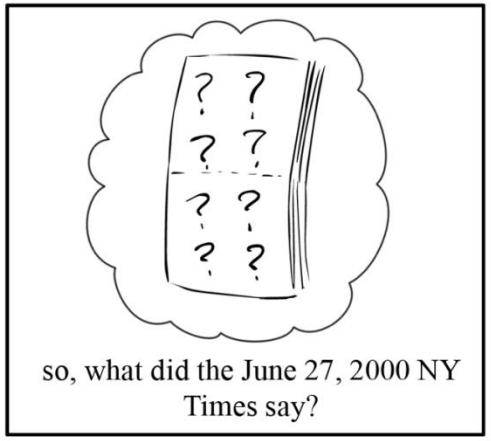


Problem novina

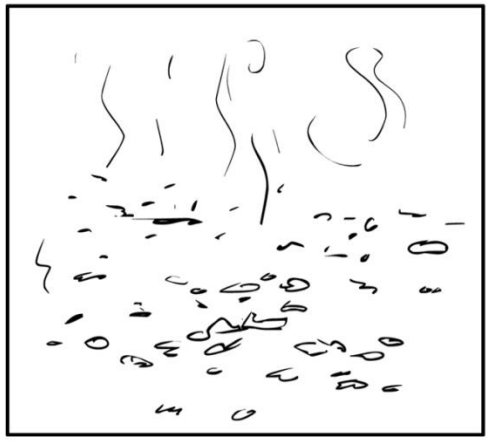


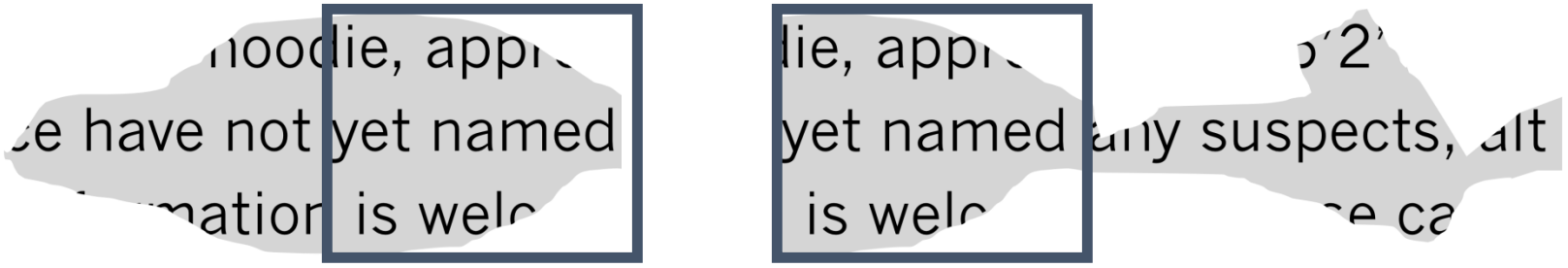




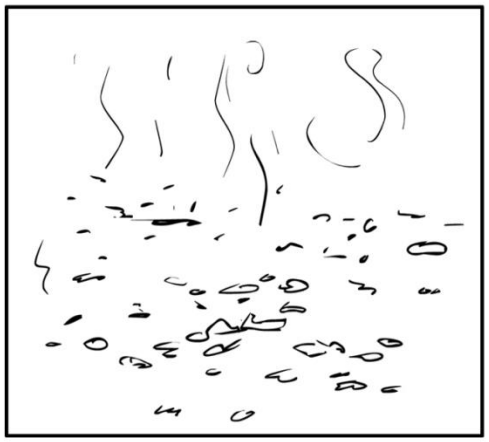


Problem novina kao delovi slagalice koji se preklapaju





Problem novina kao delovi slagalice koji se preklapaju



noodie, appr
ce have not yet named
rmation is welc

die, appr
yet named any suspects, alt
is welc
e ca

Milion kopija genoma

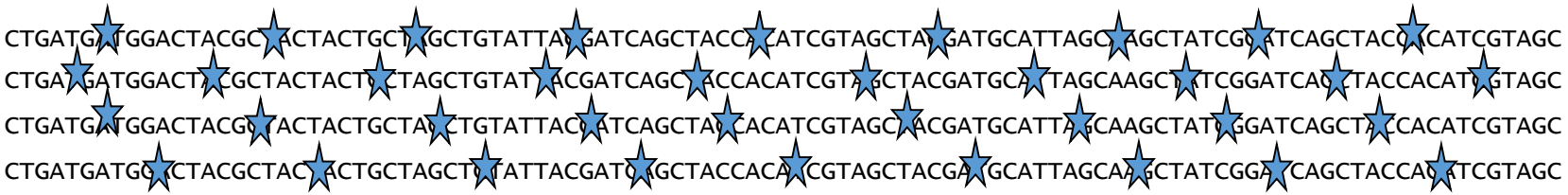
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

Genom je razbijen na slučajno odabranim pozicijama



Generisana su očitavanja (*reads*)

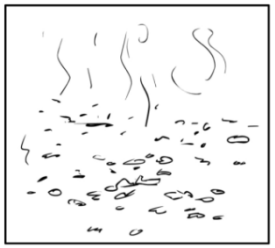
CTGATGA TGGACTACGCTAC TACTGCTAG CTGTATTACG ATCAGCTACCACA TCGTAGCTACG ATGCATTAGCAA GCTATCGGA TCAGCTACCA CATCGTAGC

CTGATGATG GACTACGCT ACTACTGCTA GCTGTATTACG ATCAGCTACC ACATCGTAGCT ACGATGCATTA GCAAGCTATC GGATCAGCTAC CACATCGTAGC

CTGATGATGG ACTACGCTAC TACTGCTAGCT GTATTACGATC AGCTACCAC ATCGTAGCTACG ATGCATTAGCA AGCTATCGG A TCAGCTACCA CATCGTAGC

CTGATGATGGACT ACGCTACTACT GCTAGCTGTAT TACGATCAGC TACCACATCGT AGCTACGATGCA TTAGCAAGCT ATCGGATCA GCTACCACATC GTAGC

Neka očitavanja su nestala u eksploziji



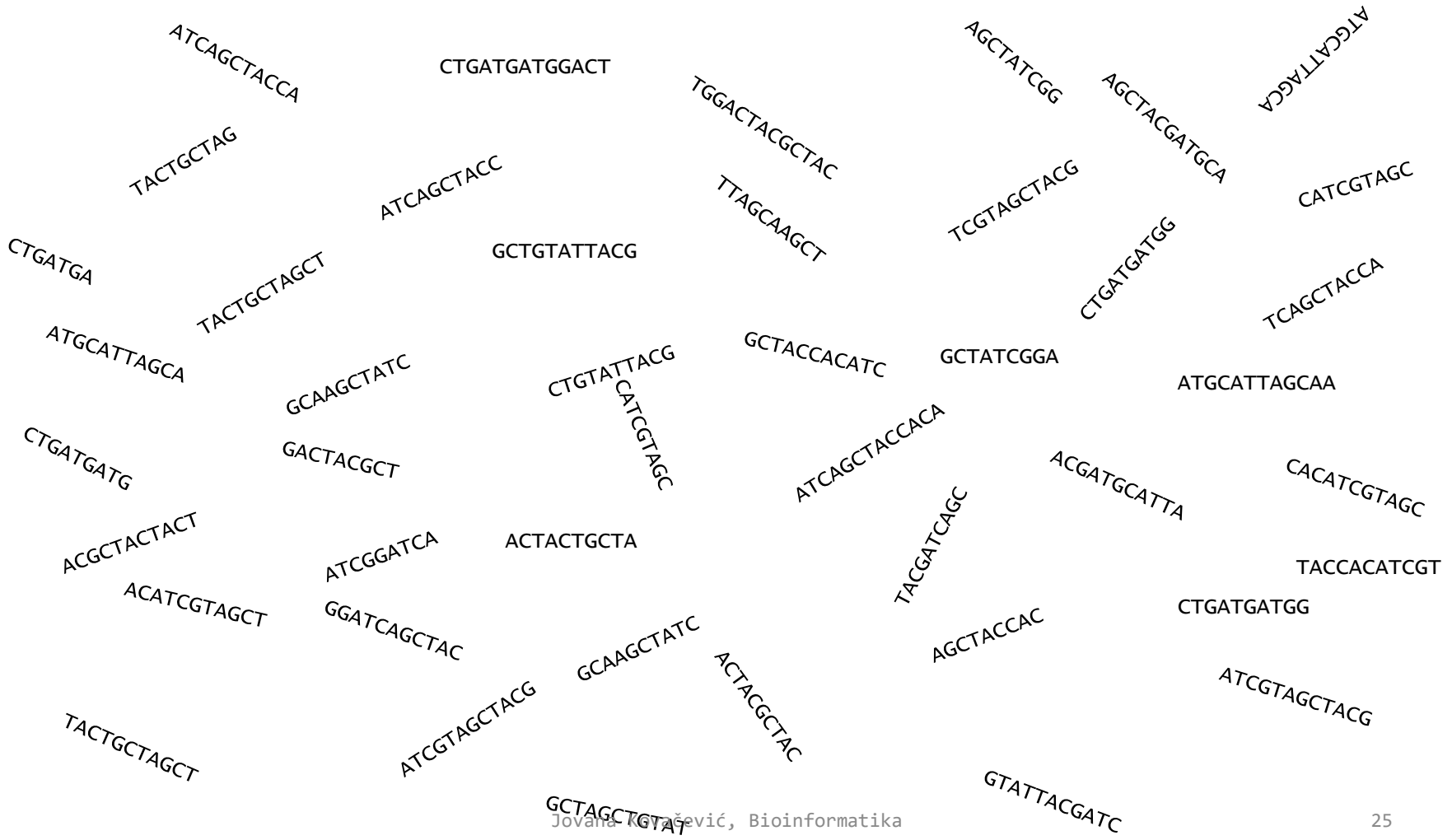
CTGATGA TGGACTACGCTAC TACTGCTAG CTGTATTACG ATCAGCTACCACA TCGTAGCTACG ATGCATTAGCAA GCTATCGGA TCAGCTACCA CATCGTAGC

CTGATGATG GACTACGCT ACTACTGCTA GCTGTATTACG ATCAGCTACC ACATCGTAGCT ACGATGCATTA GCAAGCTATC GGATCAGCTAC CACATCGTAGC

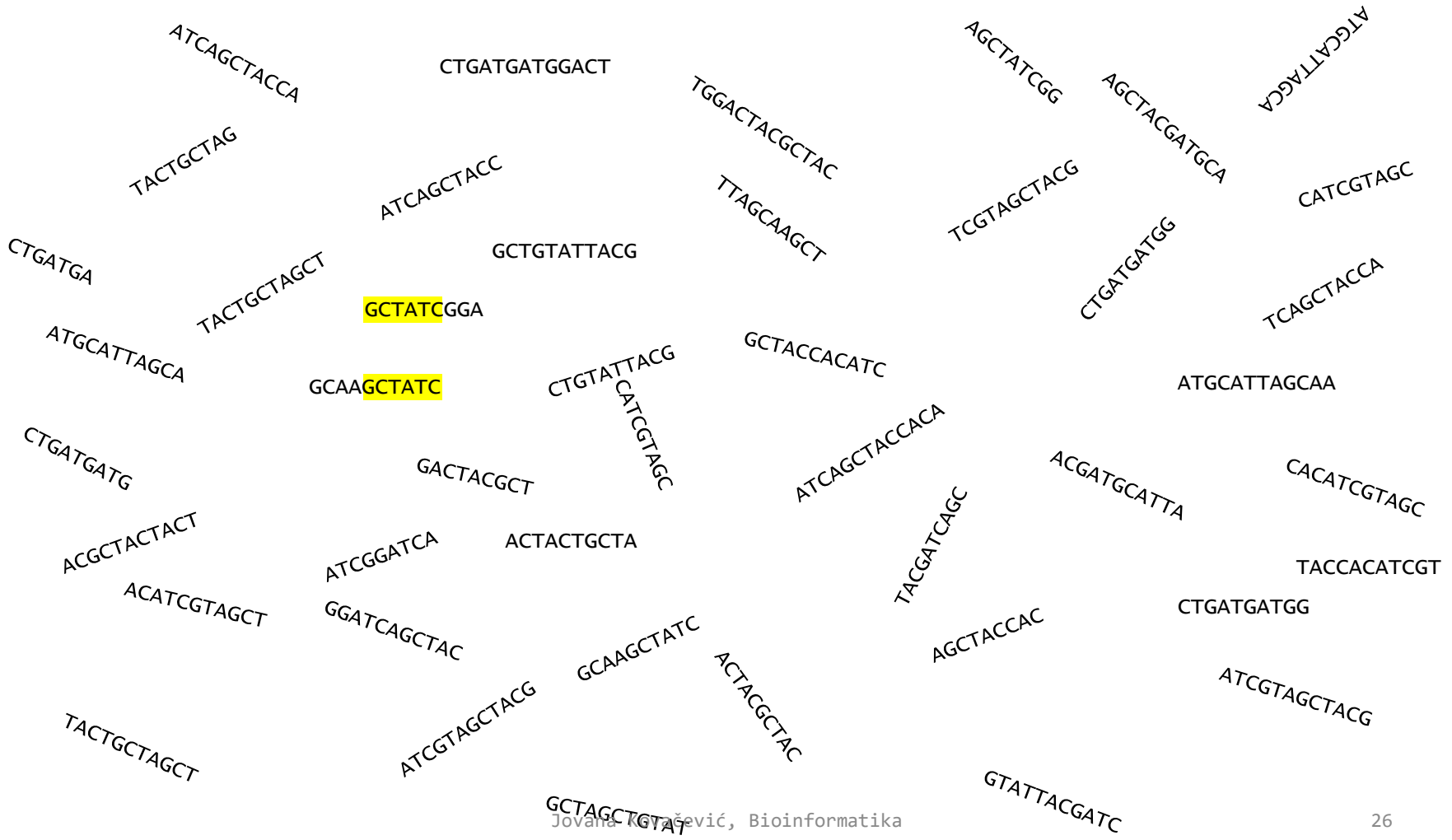
CTGATGATGG ACTACGCTAC TACTGCTAGCT GTATTACGATC AGCTACCAC ATCGTAGCTACG ATGCATTAGCA AGCTATCGG A TCAGCTACCA CATCGTAGC

CTGATGATGGACT ACGCTACTACT GCTAGCTGTAT TACGATCAGC TACCACATCGT AGCTACGATGCA TTAGCAAGCT ATCGGATCA GCTACCACATC GTAGC

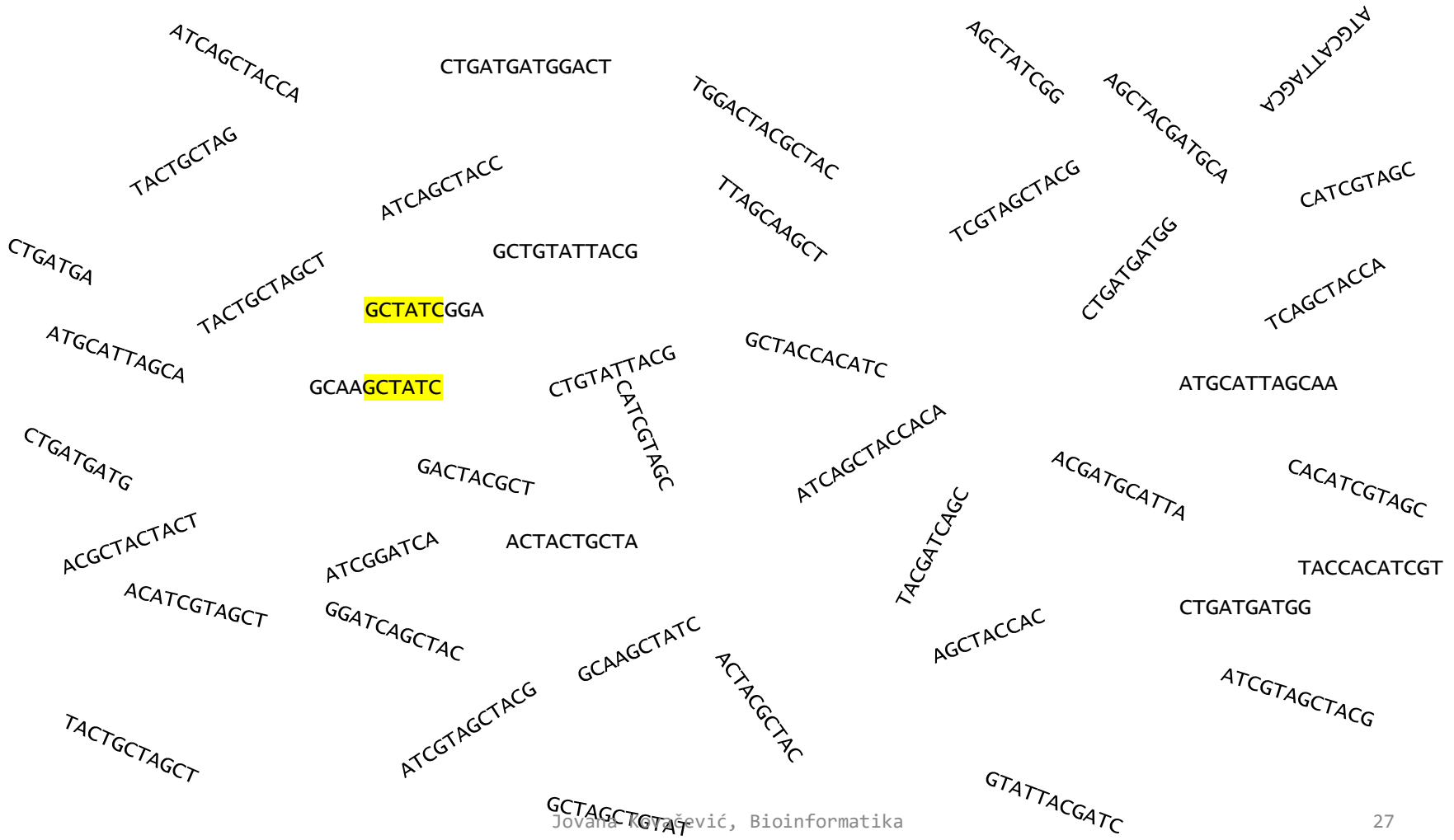
Ne znamo sa kojih pozicija su očitavanja došla



Ne znamo sa kojih pozicija su očitavanja došla



Ne znamo sa kojih pozicija su očitavanja došla



Od eksperimentalnih do računarskih problema

Više kopija genoma (nesekvencioniranog)

Generisanje očitavanja

Očitavanja

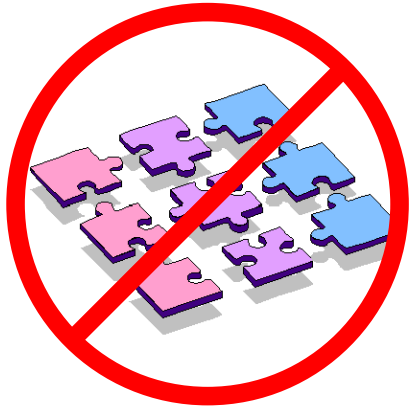
Sastavljanje genoma

Sastavaljen genom

...GGCATGCGTCAGAAACTATCATAGCTAGATCGTACGTAGCC...

Zašto je sekvencioniranje genoma teško?

- Moderne mašine za sekvencioniranje (sekvenceri) ne mogu da pročitaju ceo genom nukleotid po nukleotid od početka do kraja (kao što bismo pročitali knjigu)
- Mogu samo da iseckaju genom i generišu njegova kratka očitavanja
- Sastavljanje genoma nije isto kao i slaganje slagalice: moramo da koristimo preklapajuća očitavanja da bismo rekonstruisali genom



Pregled

- Šta je sekvencioniranje genoma?
- Eksplozija u štampariji
- **Problem rekonstrukcije niske**
- Rekonstrukcija niske kao problem Hamiltonove putanje
- Rekonstrukcija niske kao problem Ojlerove putanje
- De Brujinovi grafovi
- Ojlerova teorema
- Spajanje parova očitavanja
- U realnosti

Problem sekvencioniranja genoma

Problem sekvencioniranja genoma. Rekonstruisati genom na osnovu očitavanja.

- Ulaz. Kolekcija niski *Reads*.
- Izlaz. Niska *Genome* rekonstruisana na osnovu *Reads*.

k-gramski sastav niske

Composition$_3$(**TAATGCCATGGGATGTT**)**=**

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
GAT
ATG
TGT
GTT

k-gramski sastav niske

$Composition_3$(TAATGCCATGGGATGTT)=

TAA AAT ATG TGC GCC CCA CAT ATG TGG GGG GGA GAT ATG TGT GTT

=

AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

leksikografski poredak

Rekonstrukcija niske na osnovu njenog *k*-gramskog sastava

Problem rekonstrukcije niske.
Rekonstruisati nisku na osnovu njenog *k*-gramskog sastava.

- Ulaz. Kolekcija *k*-grama.
- Izlaz. Niska *Genome* takva da je $Composition_k(Genome)$ ekvivalentno kolekciji *k*-grama.

Naivni pristup

AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

Naivni pristup

AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA

Naivni pristup

AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA

Naivni pristup

ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA
AAT

Naivni pristup

ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA
AAT

Naivni pristup

ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA
AAT
ATG

Naivni pristup

ATG ATG CAT CCA GAT GCC GGA GGG GTT

TGC TGG TGT

TAA
AAT
ATG

Naivni pristup

ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA
AAT
ATG

Naivni pristup

ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG

TAA
AAT
ATG
TGT

Naivni pristup

ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG

TAA
AAT
ATG
TGT

Šta je sledeće?

ATG ATG CAT CCA GAT GCC GGA GGG TGC TGG

TAA
AAT
ATG
TGT
GTT

Pregled

- Šta je sekvencioniranje genoma?
- Eksplozija u štampariji
- Problem rekonstrukcije niske
- **Rekonstrukcija niske kao problem Hamiltonove putanje**
- Rekonstrukcija niske kao problem Ojlerove putanje
- De Brujinovi grafovi
- Ojlerova teorema
- Spajanje parova očitavanja
- U realnosti

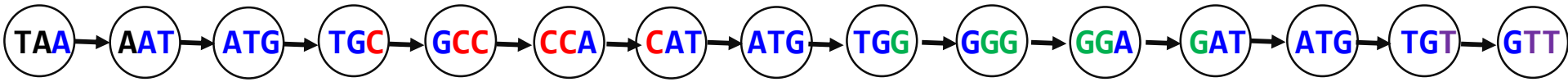
Genom kao putanja

$\text{Composition}_3(\text{TAATGCCATGGGATGTT}) =$

TAA AAT ATG TGC GCC CCA CAT ATG TGG GGG GGA GAT ATG TGT GTT

Genom kao putanja

$\text{Composition}_3(\text{TAATGCCATGGGATGTT}) =$



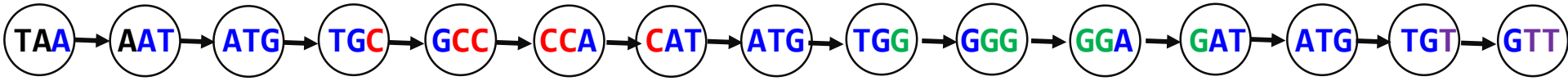
Genom kao putanja

Composition_3(TAATGCCATGGGATGTT)=

Da li možemo konstruisati ovu genomsku putanju ako ne znamo sam genom TAATGCCATGGGATGTT ali znamo njegov *k*-gramski sastav?

Genom kao putanja

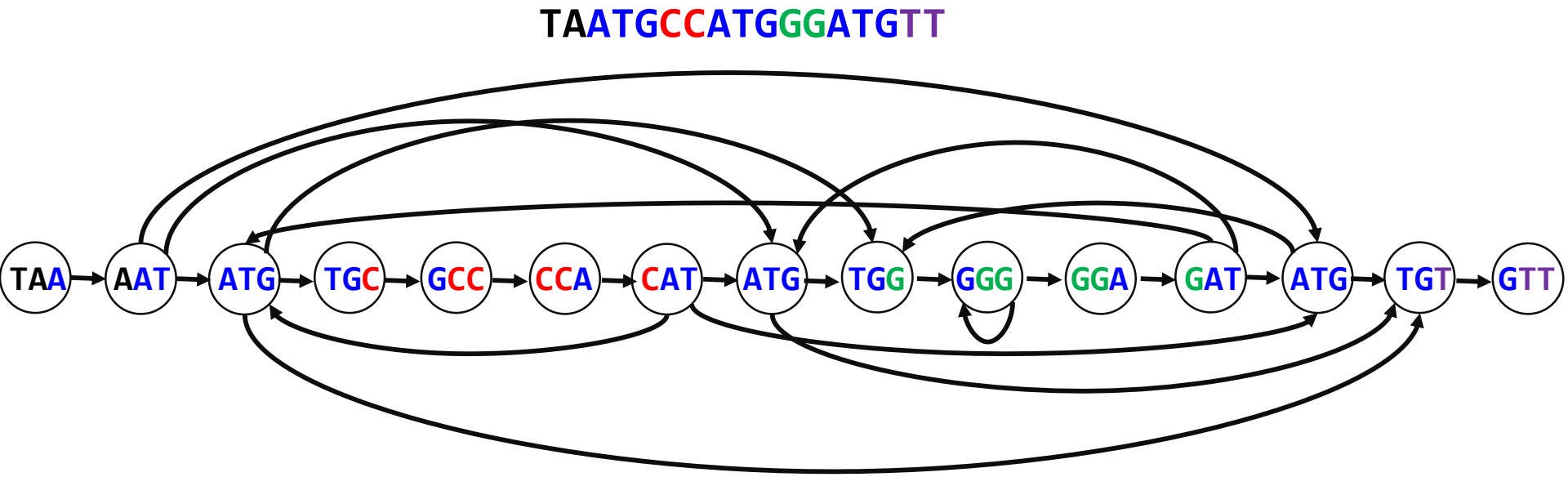
Composition_3(TAATGCCATGGGATGTT)=



Da li možemo konstruisati ovu genomsku putanju ako ne znamo sam genom TAATGCCATGGGATGTT ali znamo njegov *k*-gramski sastav?

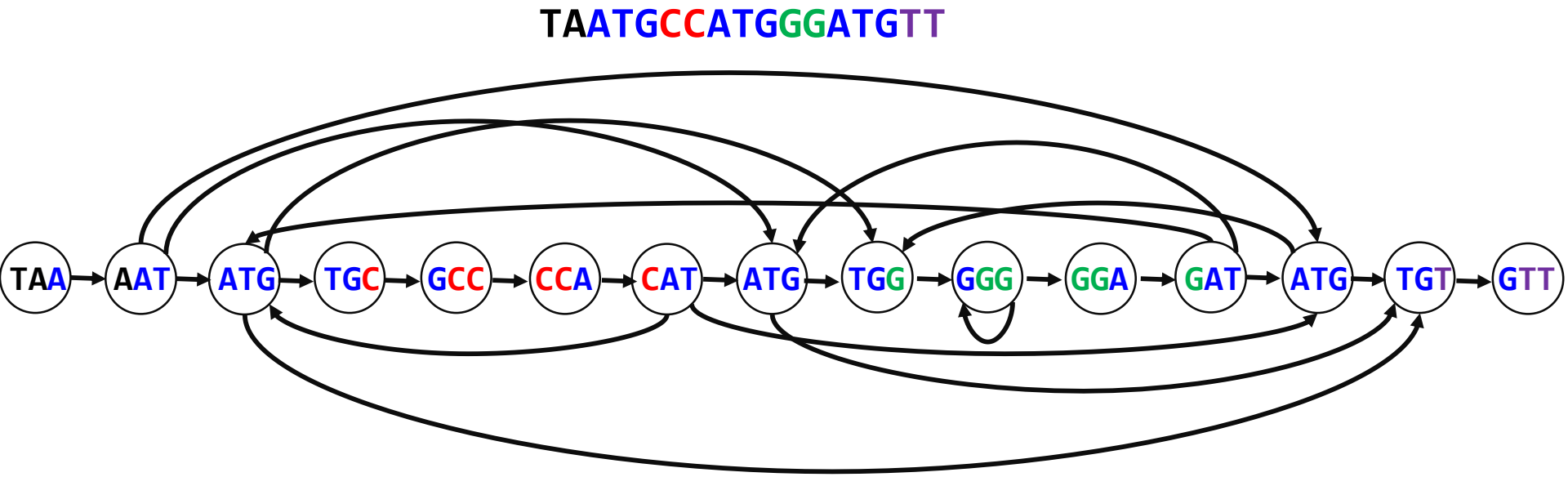
Možemo. Treba da povežemo $k\text{-mer}_1$ sa $k\text{-mer}_2$ ako
$suffix(k\text{-mer}_1)=prefix(k\text{-mer}_2)$.
E.g. TAA → AAT

Graf na osnovu *k*-gramskog sastava



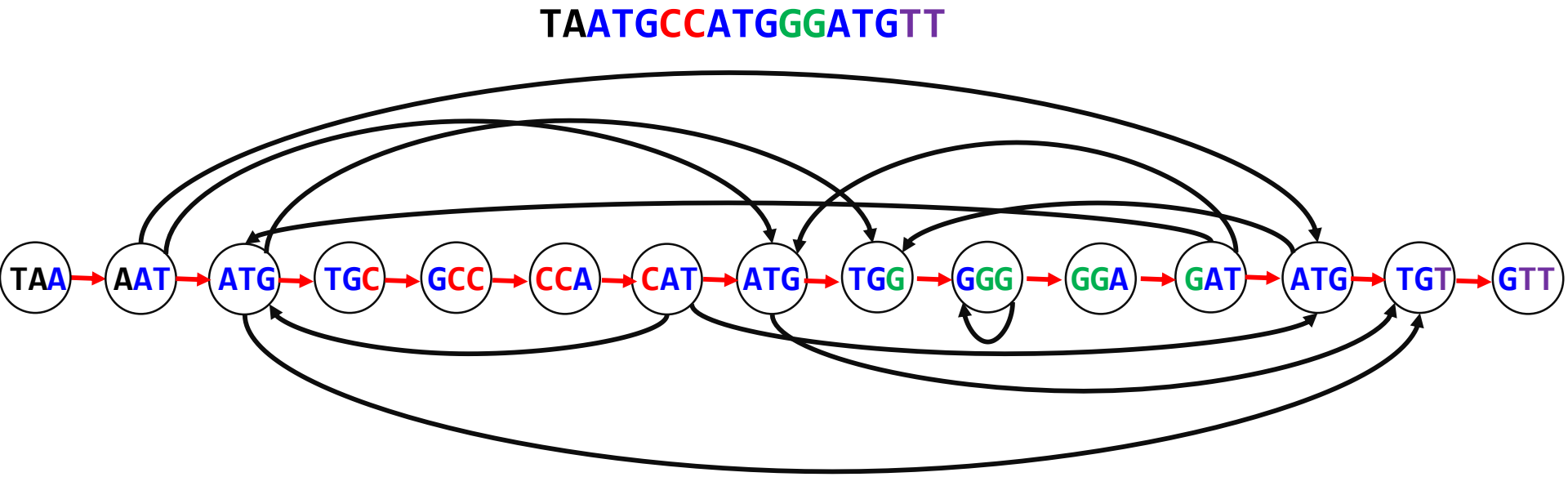
Možemo. Treba da povežemo $k\text{-mer}_1$ sa $k\text{-mer}_2$ ako
$suffix(k\text{-mer}_1)=prefix(k\text{-mer}_2)$.
E.g. TAA → AAT

Graf na osnovu *k*-gramskog sastava



Od svih putanja, da li možemo da pronađemo genomsku putanju u ovom grafu?

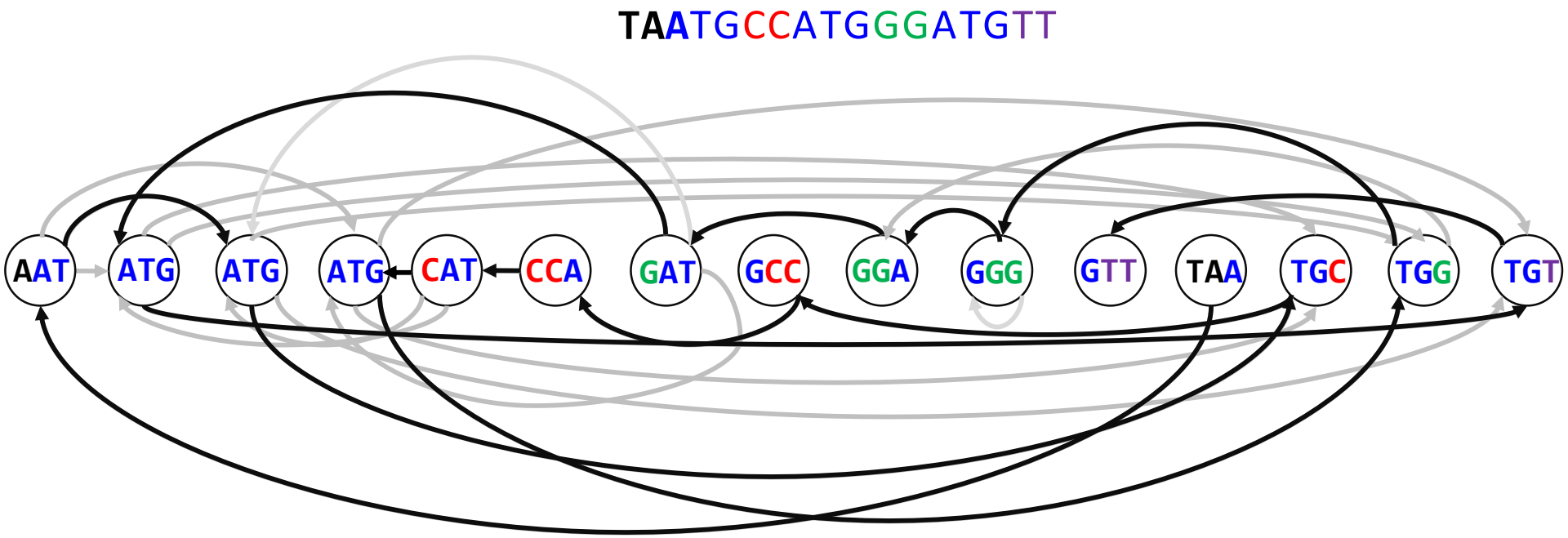
Graf na osnovu *k*-gramskog sastava



Od svih putanja, da li možemo da pronađemo genomsku putanju u ovom grafu?

Gde je genomska putanja?

Hamiltonova putanja: putanja koja posećuje svaki čvor u grafu tačno jednom.



Šta pokušavamo da pronađemo na ovom grafu?

Problem Hamiltonove putanje

Problem Hamiltonove putanje. Naći Hamiltonovu putanju u grafu.

- **Ulaz**. Graf.
- **Izlaz**. Putanja koja posećuje svaki čvor u grafu tačno jednom

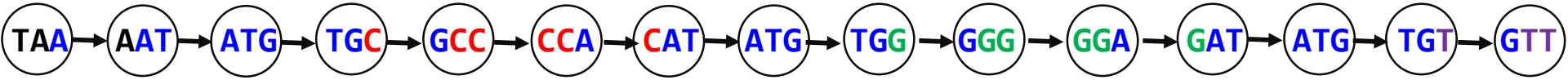
Nalaženje Hamiltonove putanje je NP kompletan problem!

Pregled

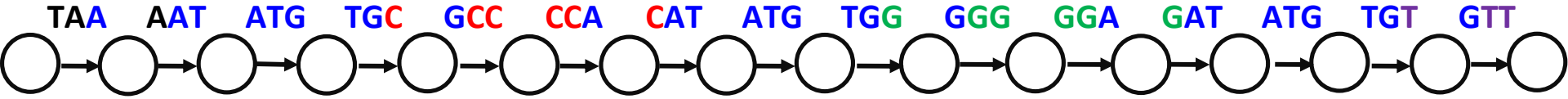
- Šta je sekvencioniranje genoma?
- Eksplozija u štampariji
- Problem rekonstrukcije niske
- Rekonstrukcija niske kao problem Hamiltonove putanje
- **Rekonstrukcija niske kao problem Ojlerove putanje**
- De Brujinovi grafovi
- Ojlerova teorema
- Spajanje parova očitavanja
- U realnosti

Malo drugačija putanja

TAATGCCATGGGATGTT

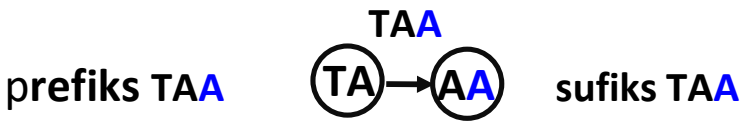


3-grami kao čvorovi



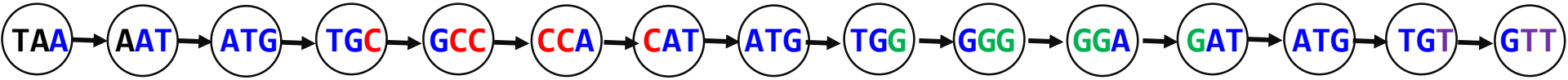
3-grami kao grane

Kako obeležavamo početni i krajnji čvor grane?

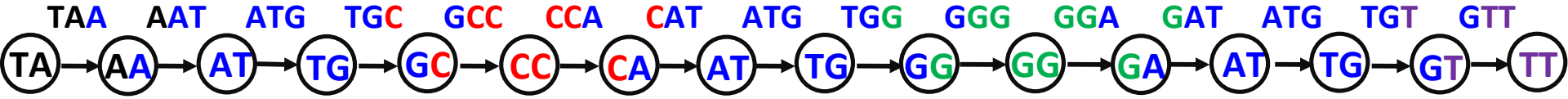


Obeležavanje čvorova u novoj putanji

TAATGCCATGGGATGTT

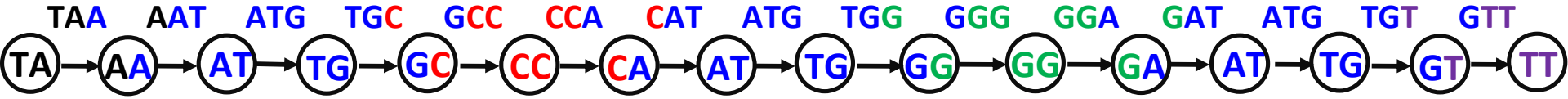


3-grami su čvorovi



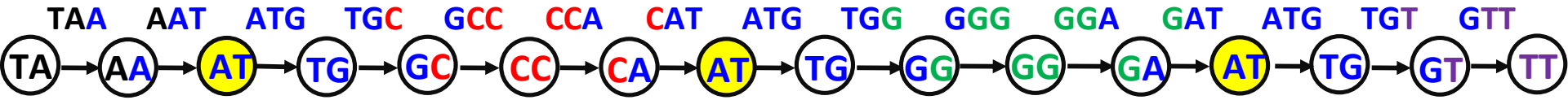
3-grami su grane a 2-grami su čvorovi

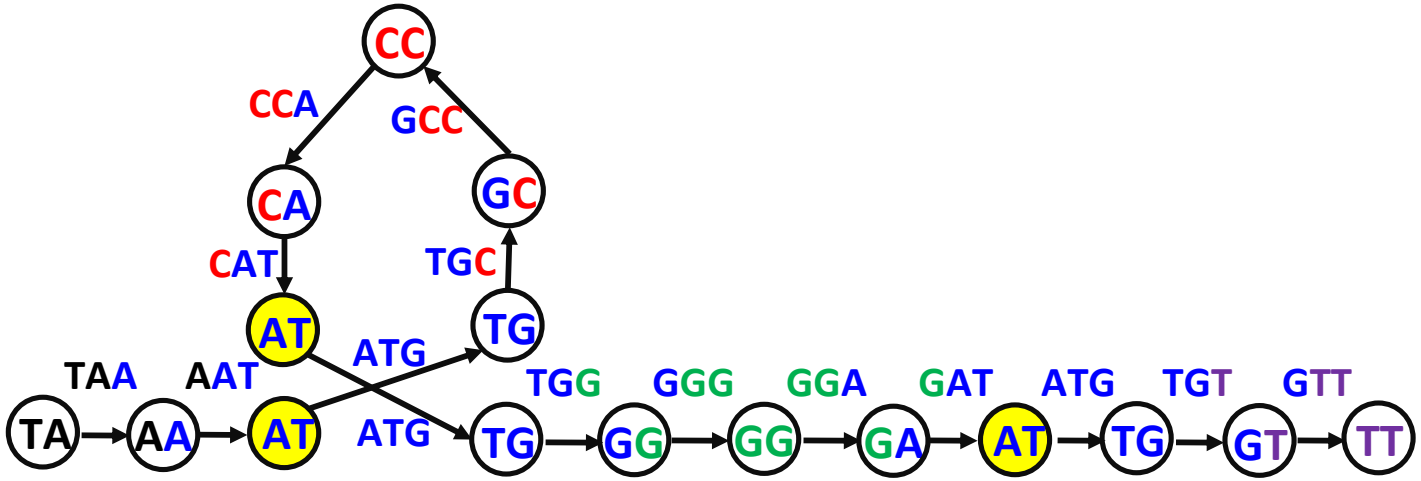
Obeležavanje čvorova u novoj putanji



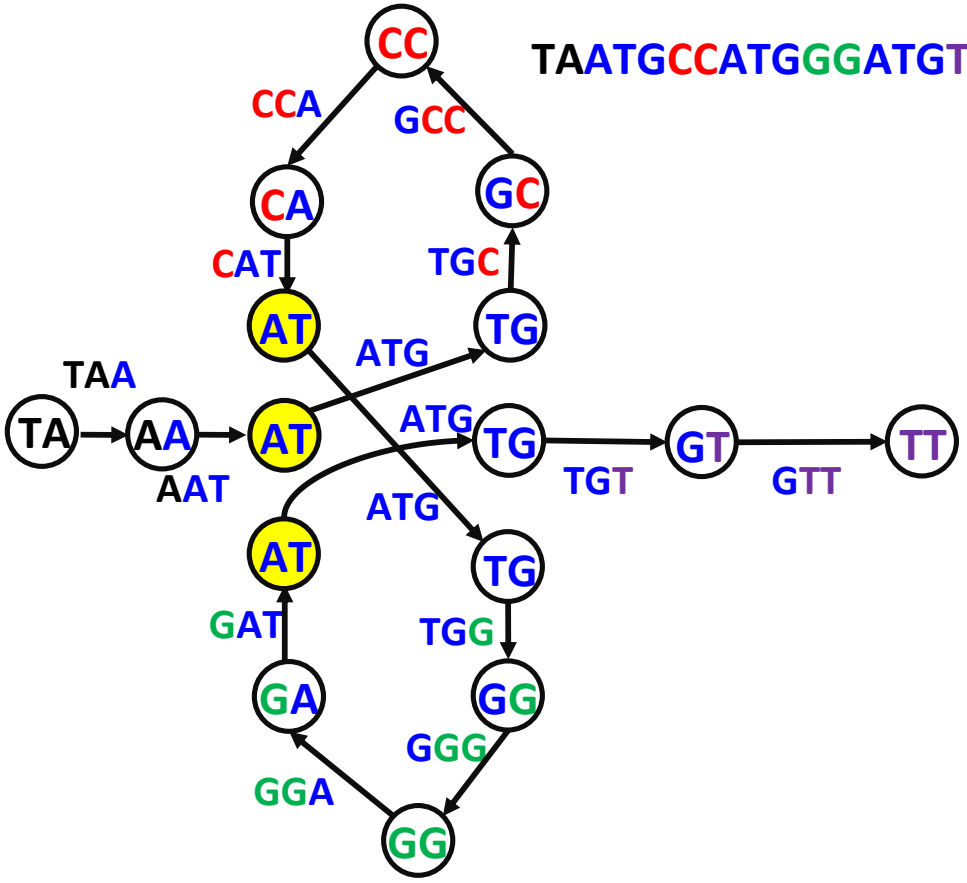
3-grami su grane a 2-grami su čvorovi

Lepljenje identično obeleženih čvorova

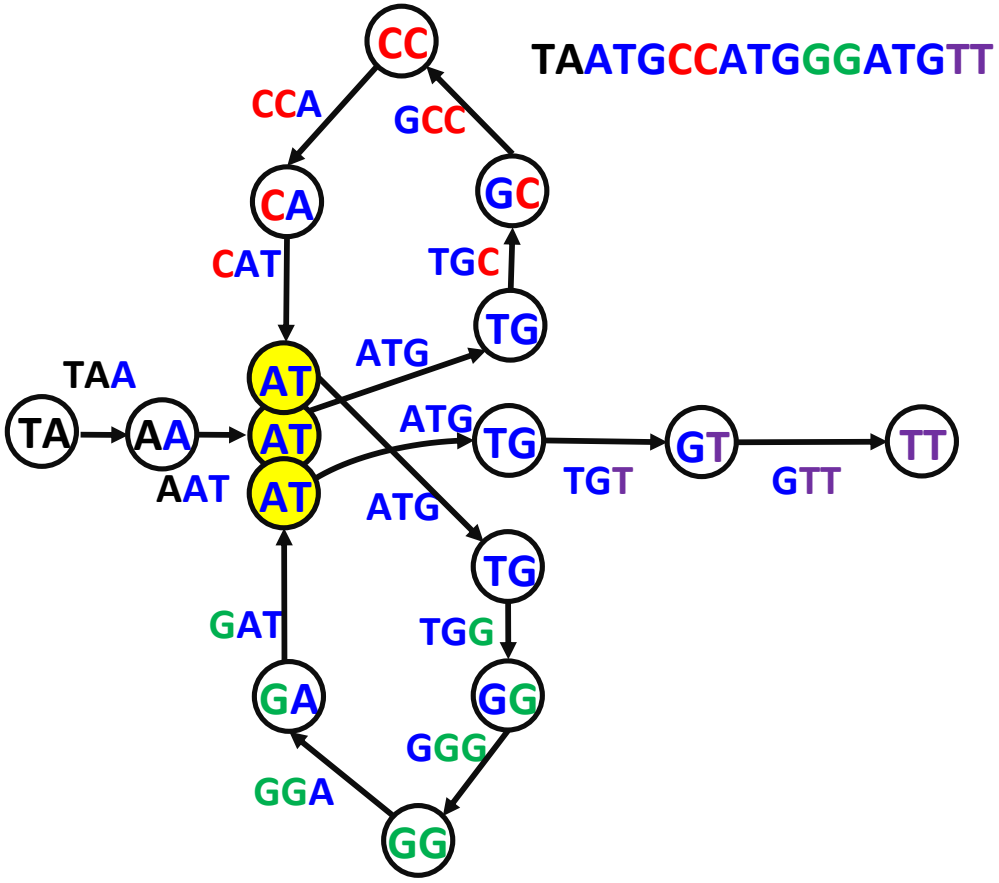




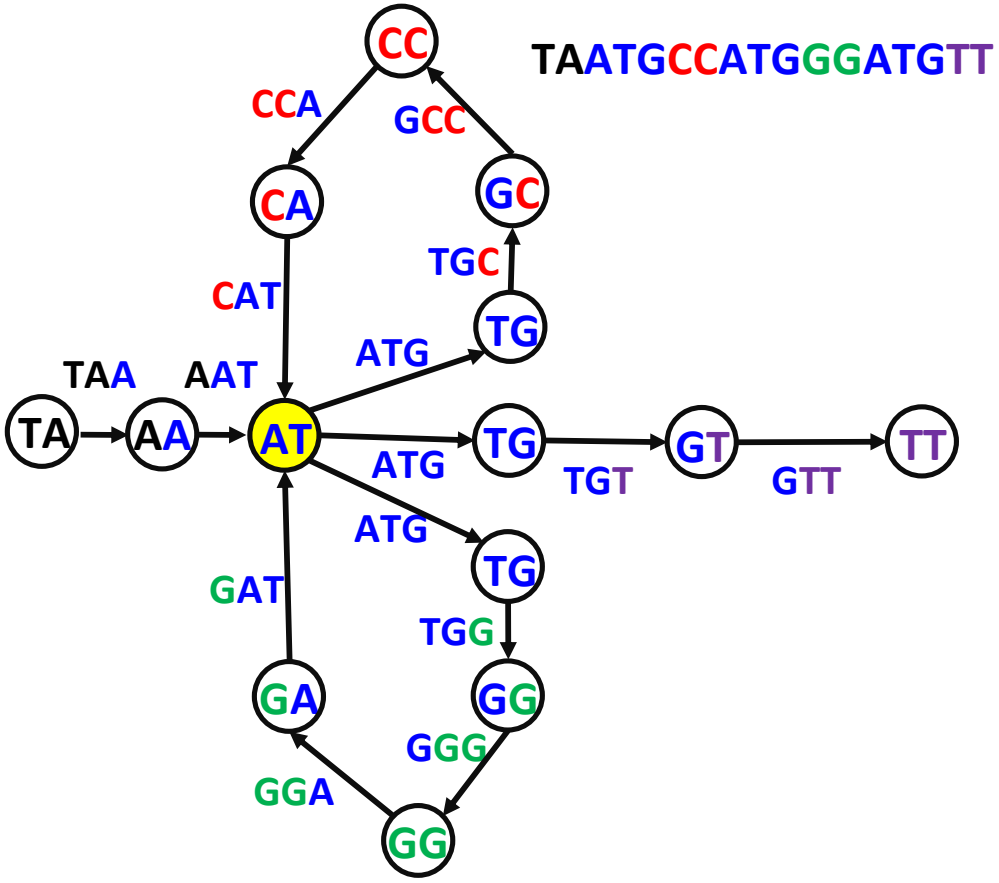
Lepljenje identično obeleženih čvorova



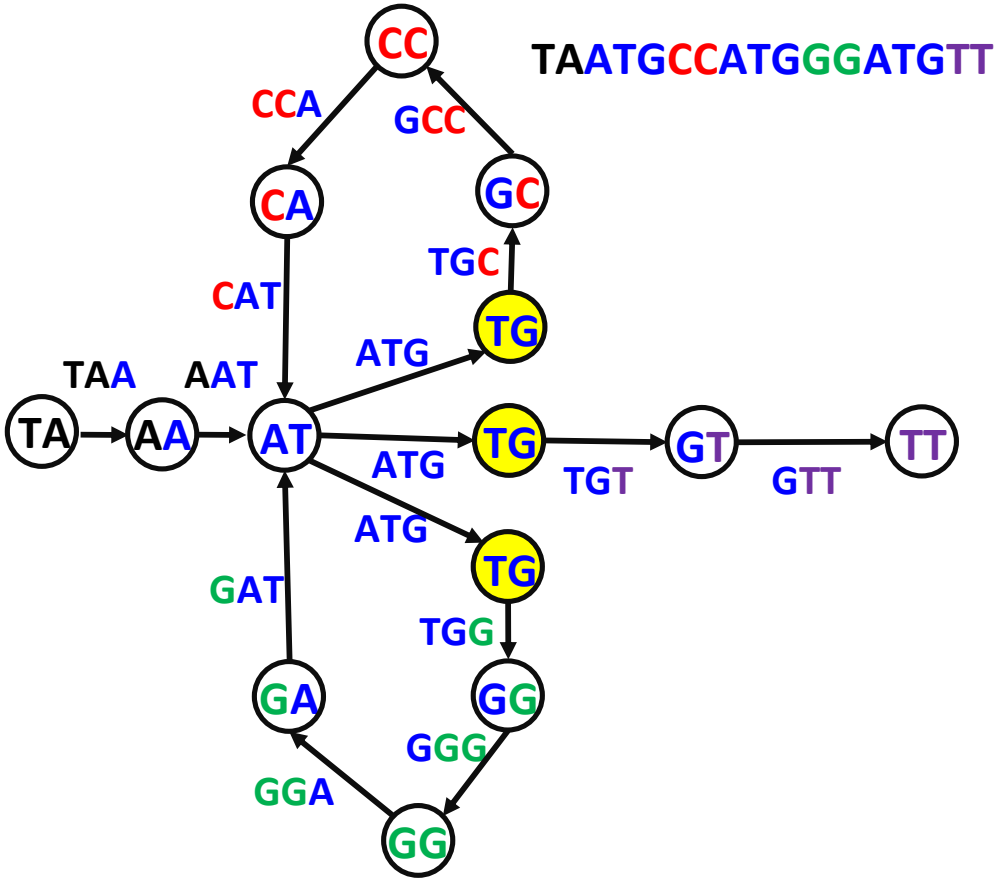
Lepljenje identično obeleženih čvorova



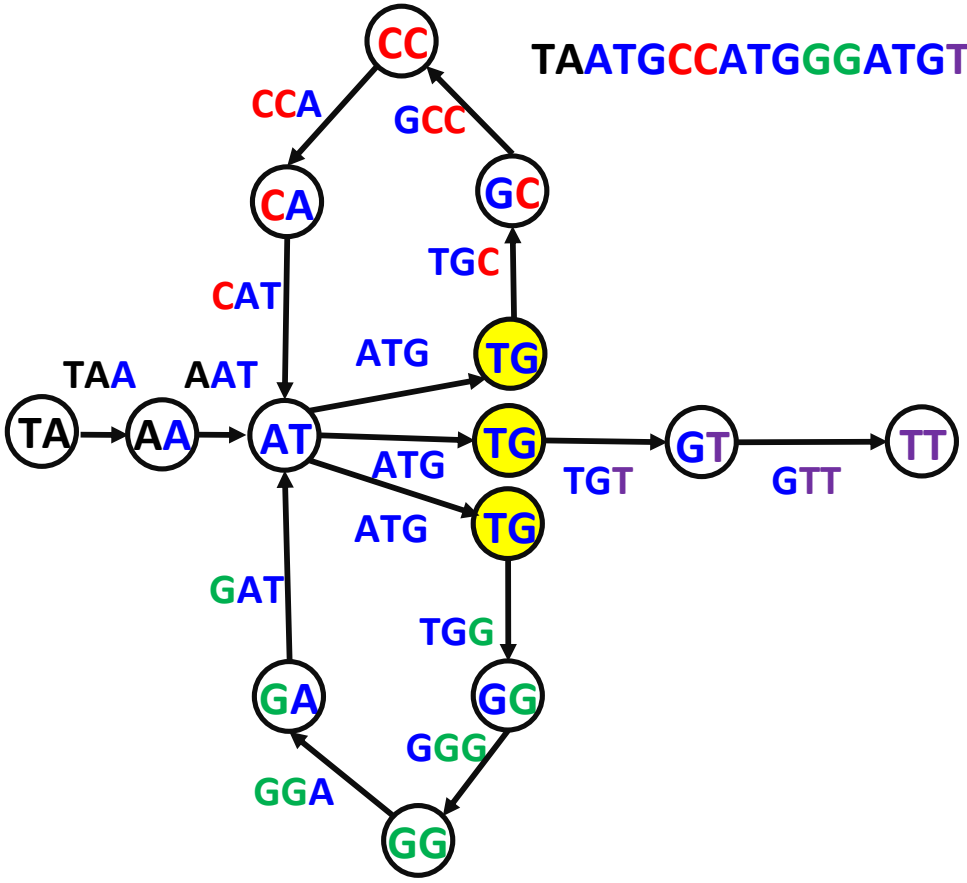
Lepljenje identično obeleženih čvorova



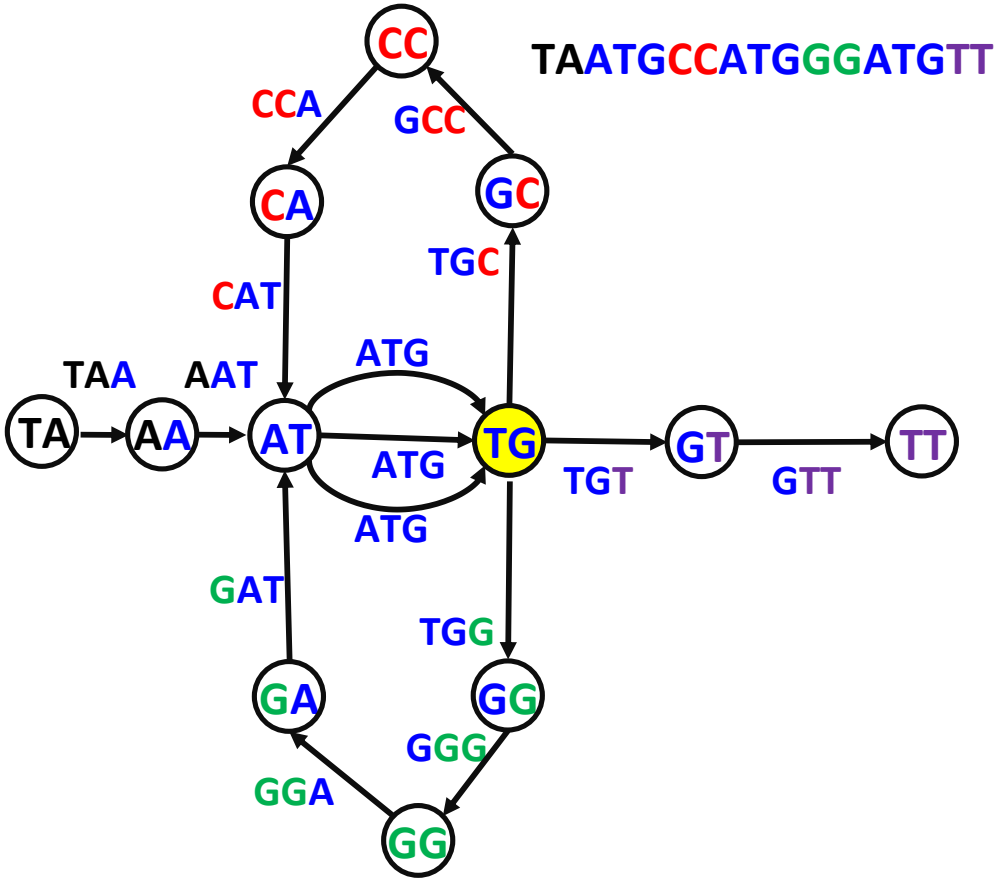
Lepljenje identično obeleženih čvorova



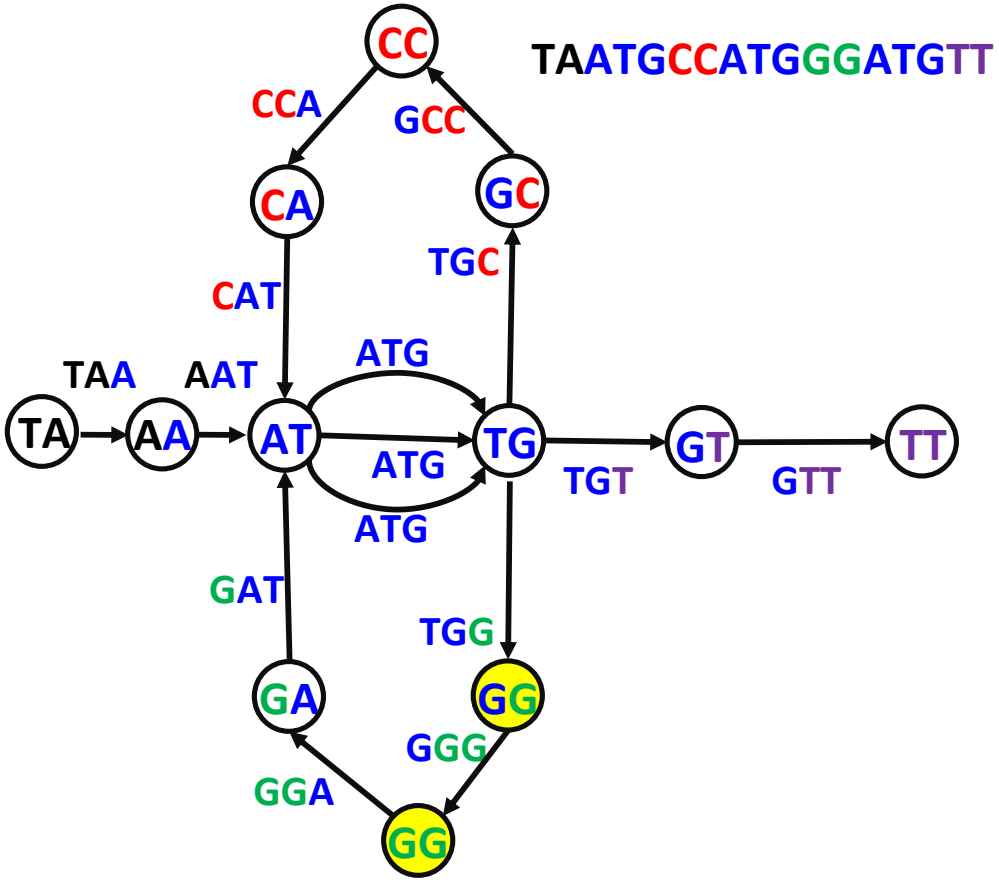
Lepljenje identično obeleženih čvorova



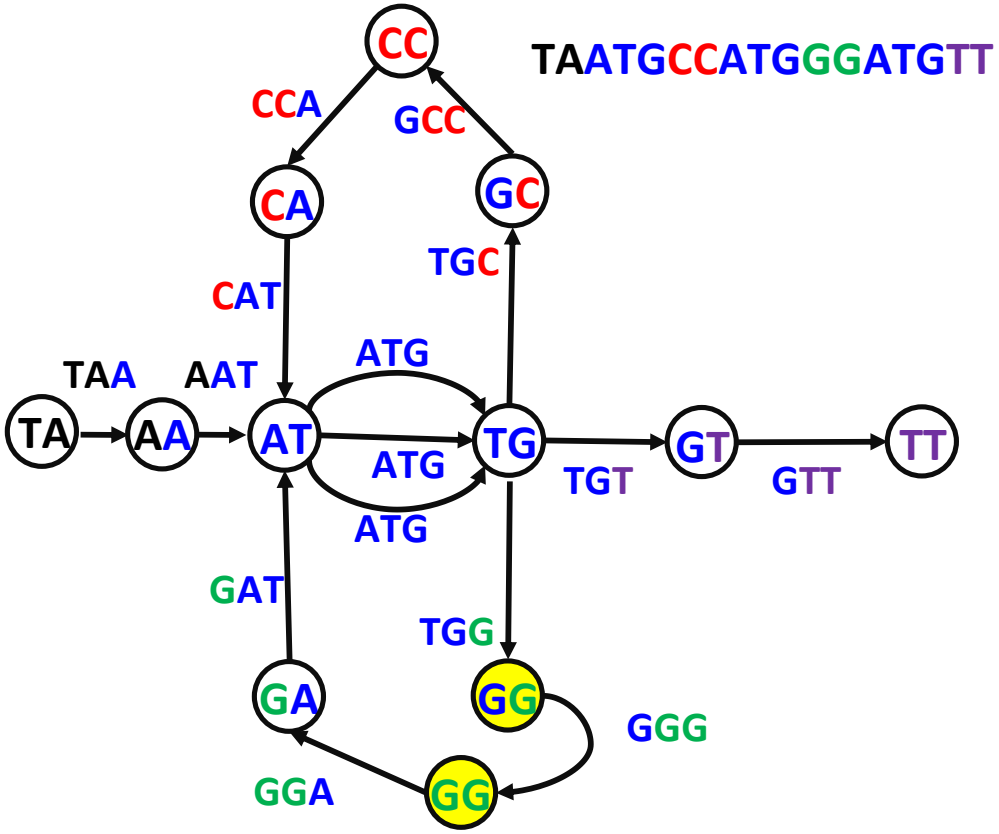
Lepljenje identično obeleženih čvorova



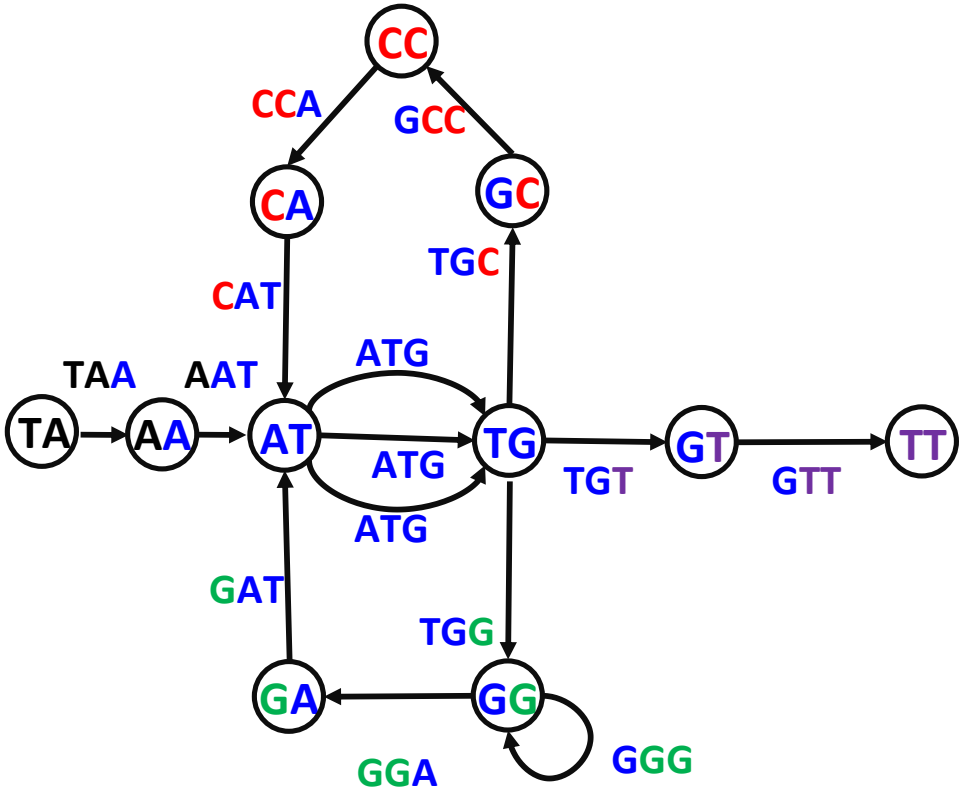
Lepljenje identično obeleženih čvorova



Lepljenje identično obeleženih čvorova

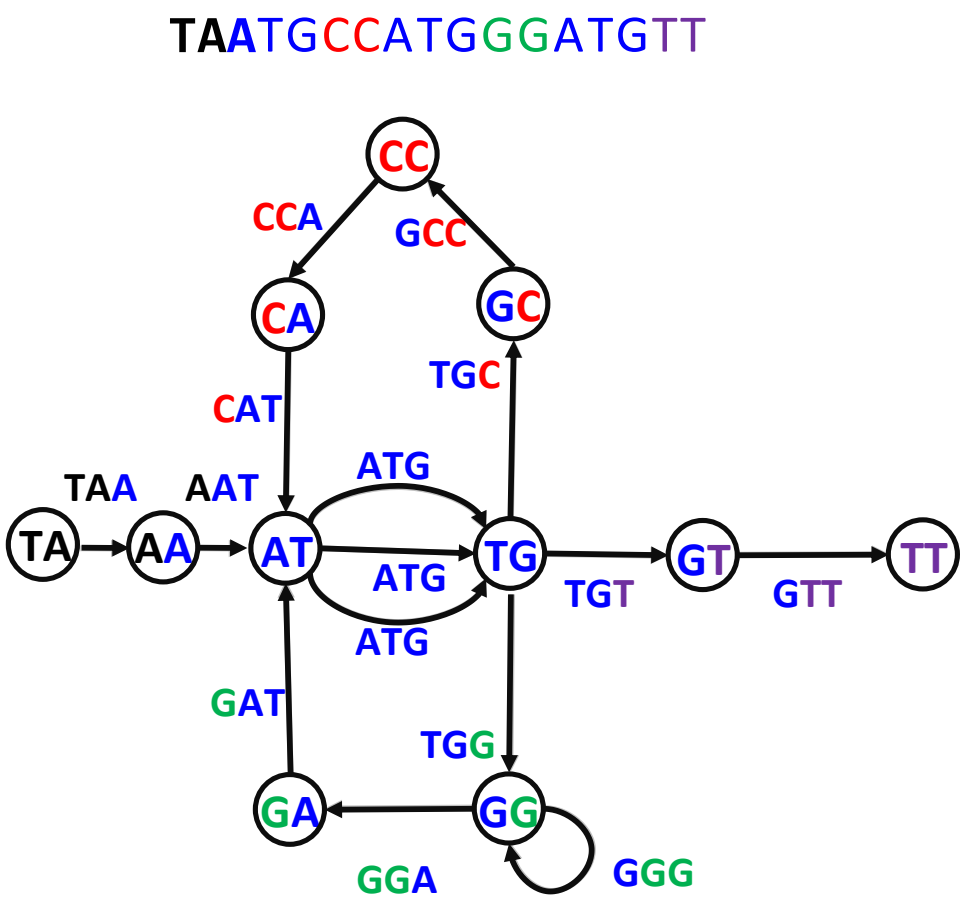


De Brujinov graf za nisku TAATGCCATGGGATGTT



Gde se *Genome* krije u ovom grafu?

Gde je *Genome* u De Brujinovom grafu?

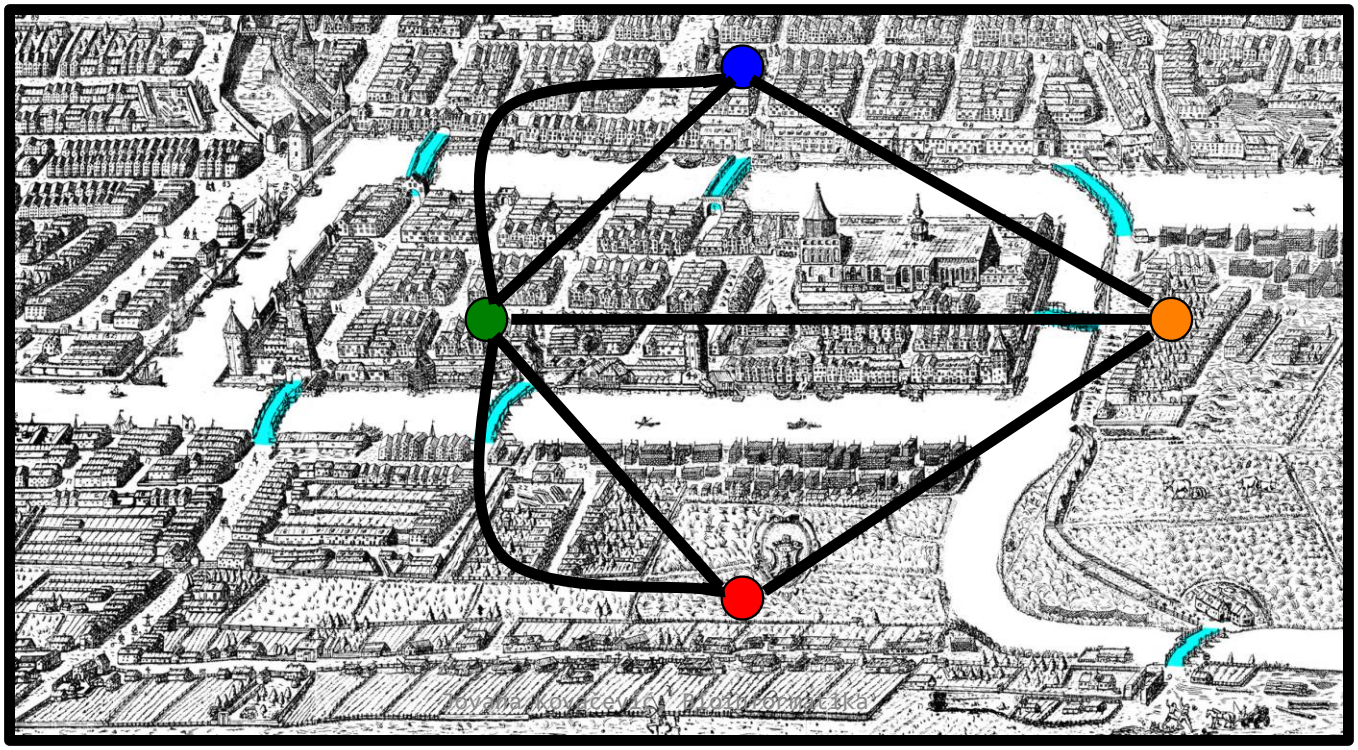


Ojlerova **putanja** u grafu je putanja koja posećuje svaku granu tačno jednom.

Problem Ojlerove putanje

Problem Ojlerove putanje. Pronaći Ojlerovu putanju u grafu.

- **Ulaz.** Graf.
- **Izlaz.** Putanja koja posećuje svaku granu u grafu tačno jednom.



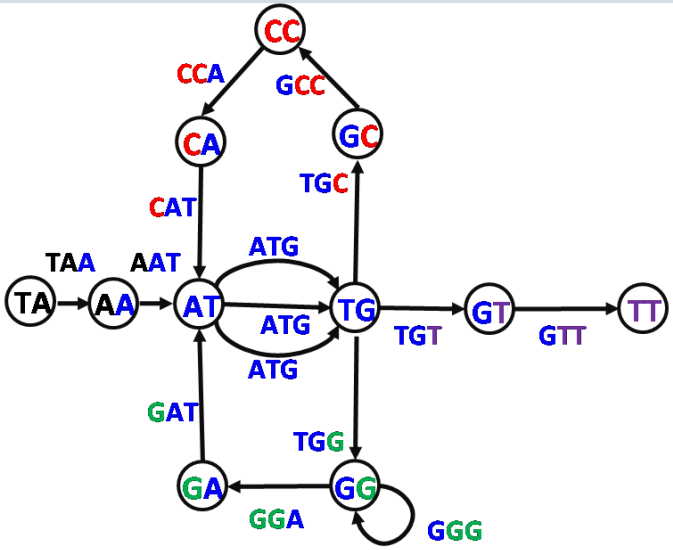
Pregled

- Šta je sekvencioniranje genoma?
- Eksplozija u štampariji
- Problem rekonstrukcije niske
- Rekonstrukcija niske kao problem Hamiltonove putanje
- Rekonstrukcija niske kao problem Ojlerove putanje
- **De Brujinovi grafovi**
- Ojlerova teorema
- Spajanje parova očitavanja
- U realnosti

Problem Ojlerove putanje

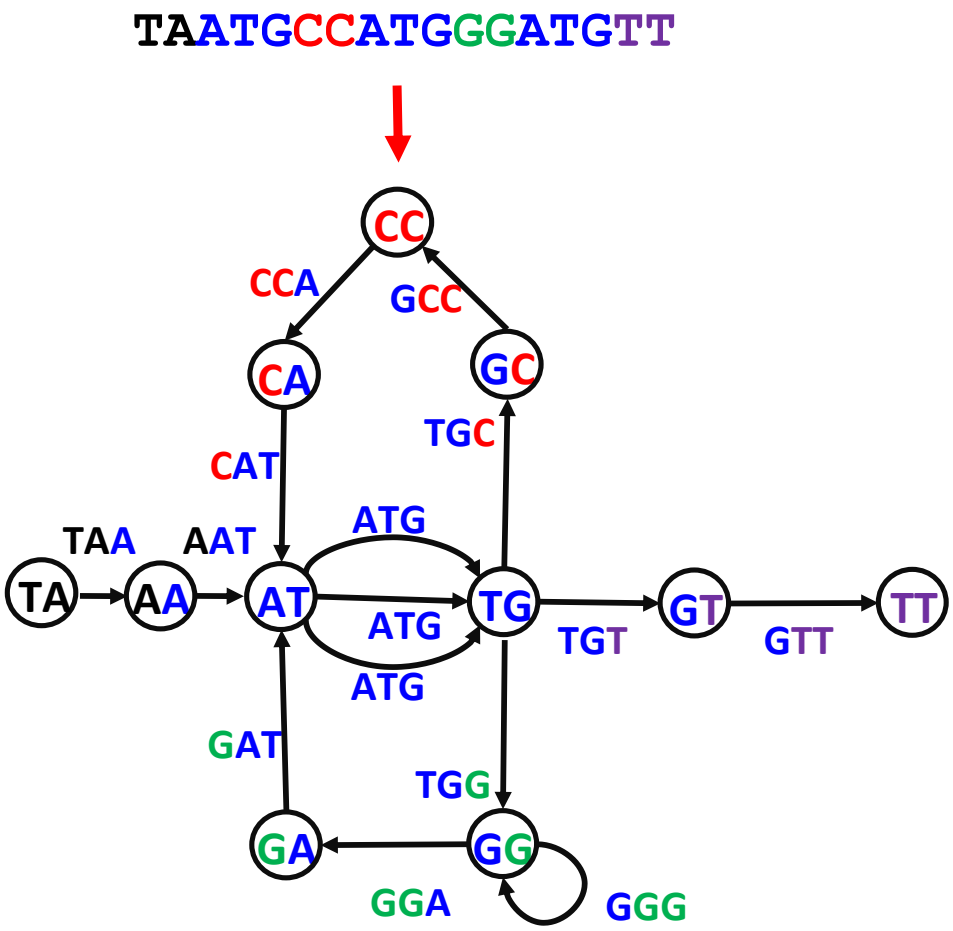
Problem Ojlerove putanje. Pronaći Ojlerovu putanju u grafu.

- **Ulaz.** Graf.
- **Izlaz.** Putanja koja posećuje svaku granu u grafu tačno jednom.



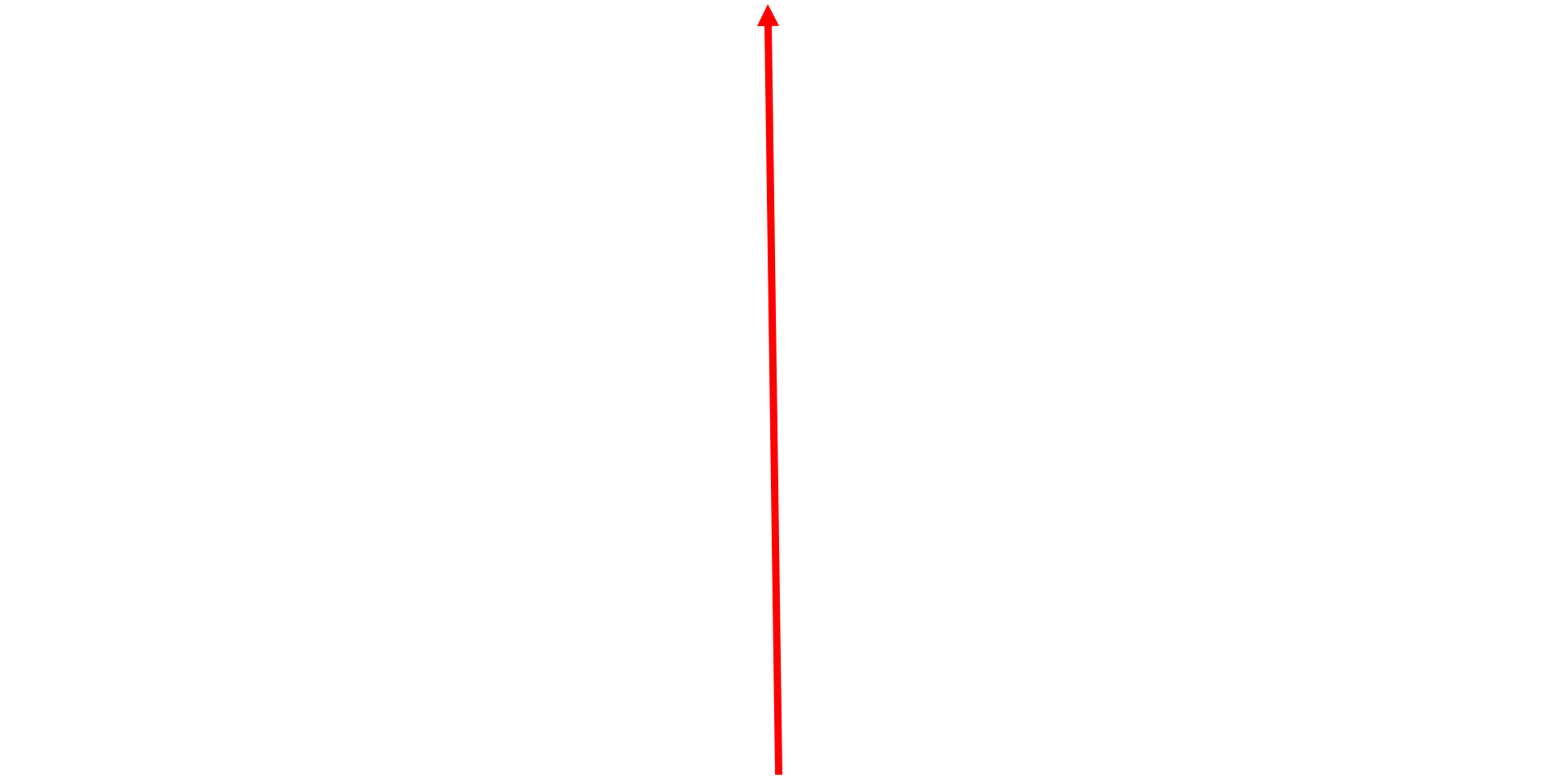
Konstruisali smo De Brujinov graf na osnovu genoma, ali u realnim primenama, genom je nepoznat!

Urađeno: Od genoma do De Brujinovog grafa



Želimo da uradimo: Od očitavanja (kolekcije *k*-grama) do genoma

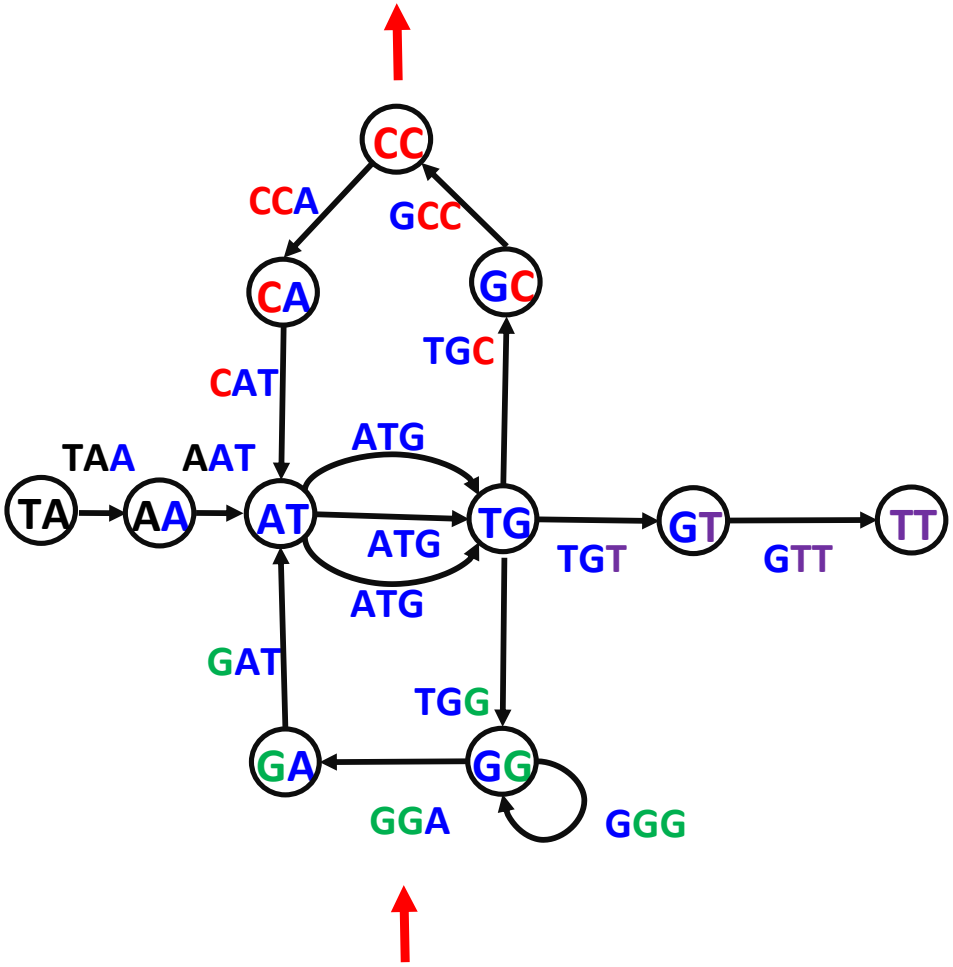
TAATGCCATGGGATGTT



AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

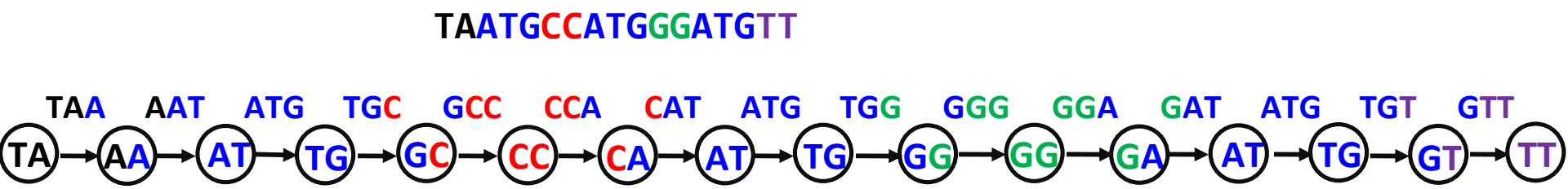
Pokazaćemo: Od očitavanja do De Brujinovog grafa do genoma

TAATGCCATGGGATGTT



AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

Konstrukcija De Brujinovog grafa kada je genom poznat



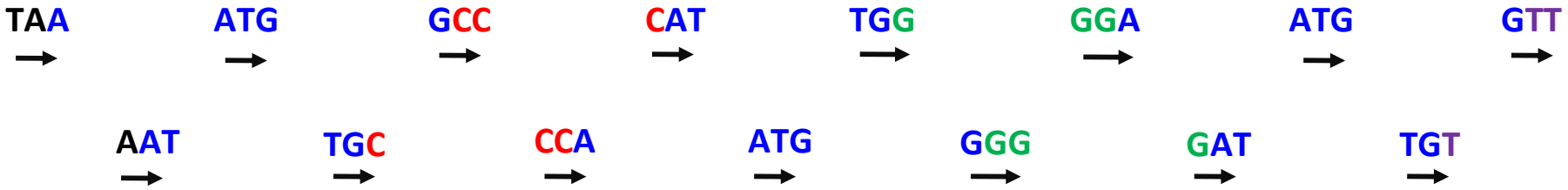
Konstrukcija De Brujinovog grafa kada je genom nepoznat

TAA ATG GCC CAT TGG GGA ATG GTT

AAT TGC CCA ATG GGG GAT TGT

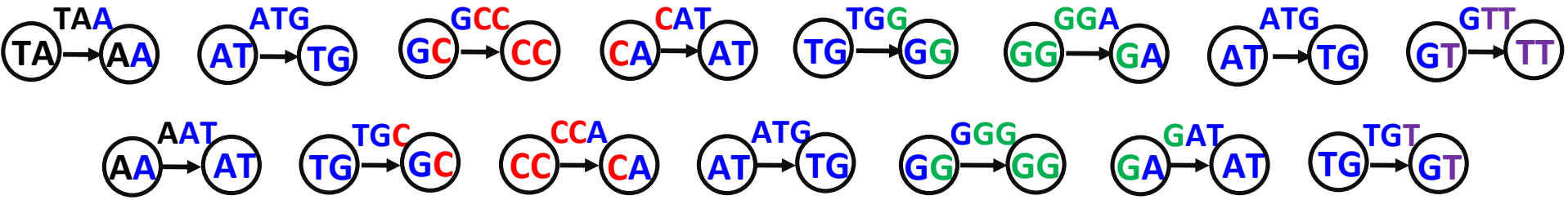
Composition$_3$(TAATGCCATGGGATGTT)

Predstavimo *k*-gramski sastav kao graf koji se sastoji od nepovezanih grana



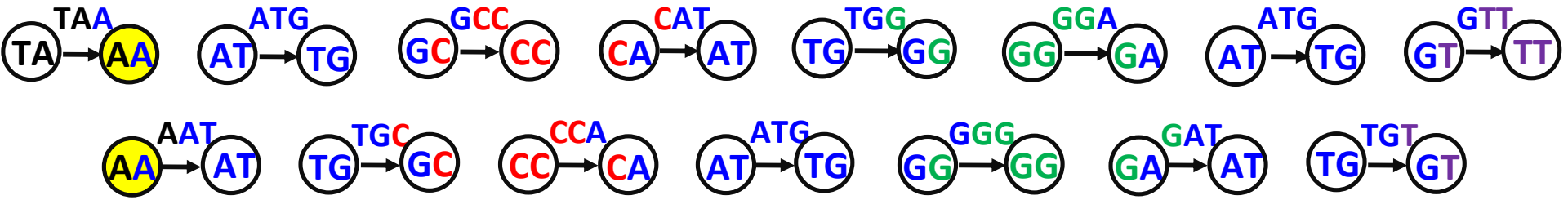
Composition_3(**TAATGCCATGGGATGTT**)

Konstruišemo De Brujinov graf na osnovu *k*-gramskog sastava

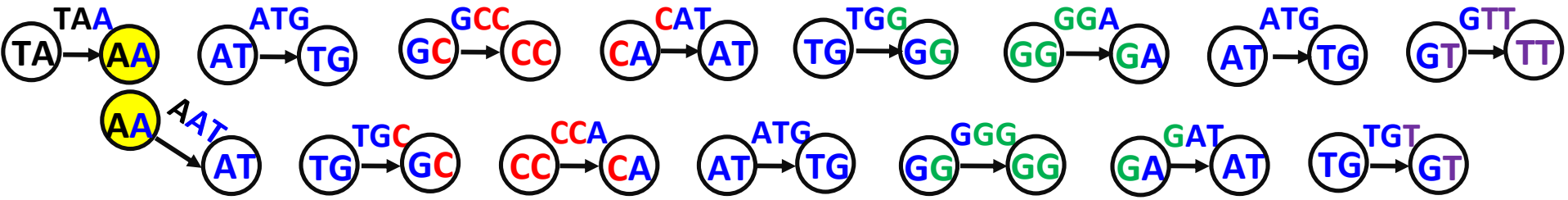


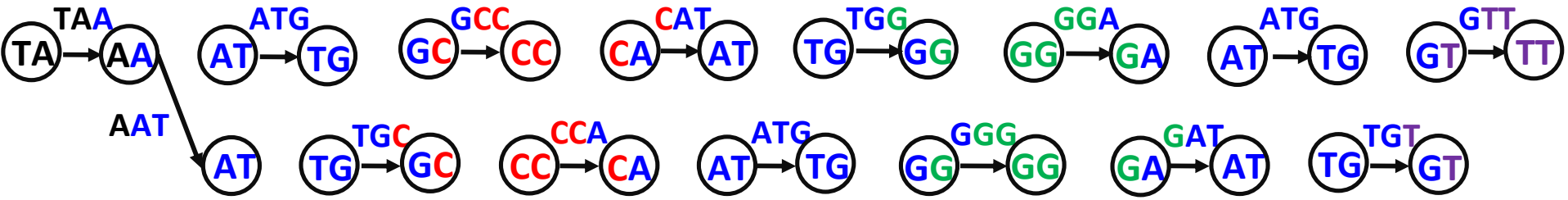
$\text{Composition}_3(\mathbf{TAATGCCATGGGATGTT})$

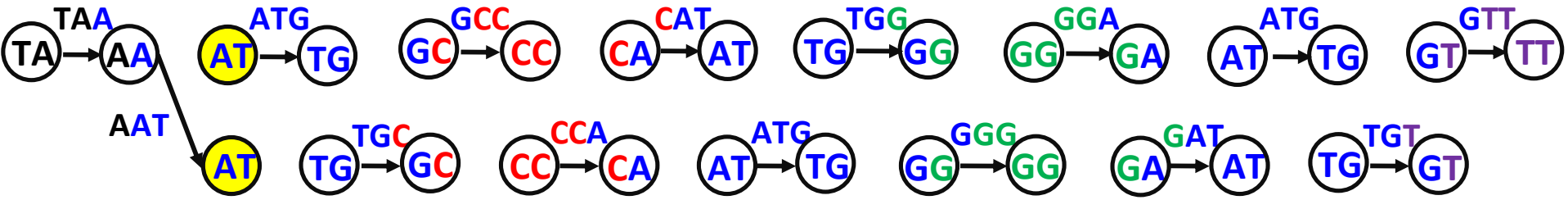
Zalepimo identično obeležene čvorove

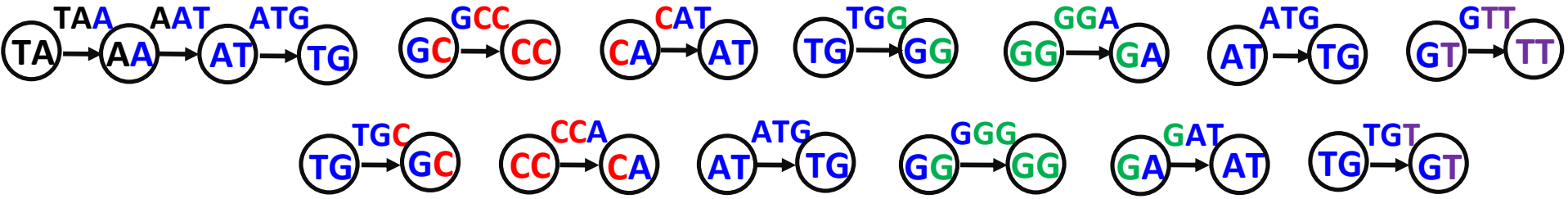


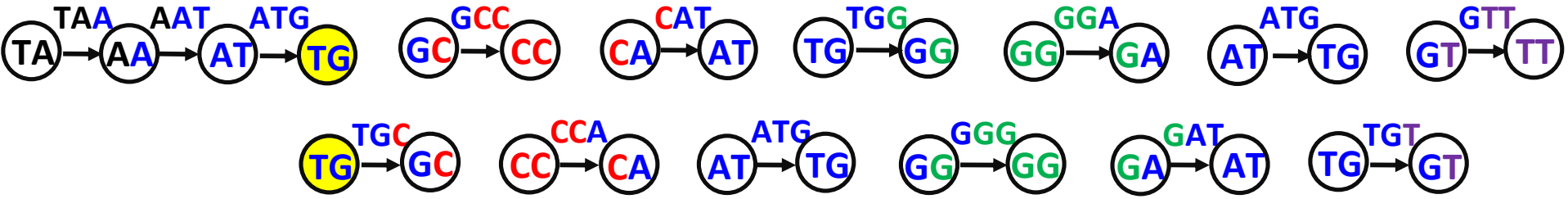
Zalepimo identično obeležene čvorove

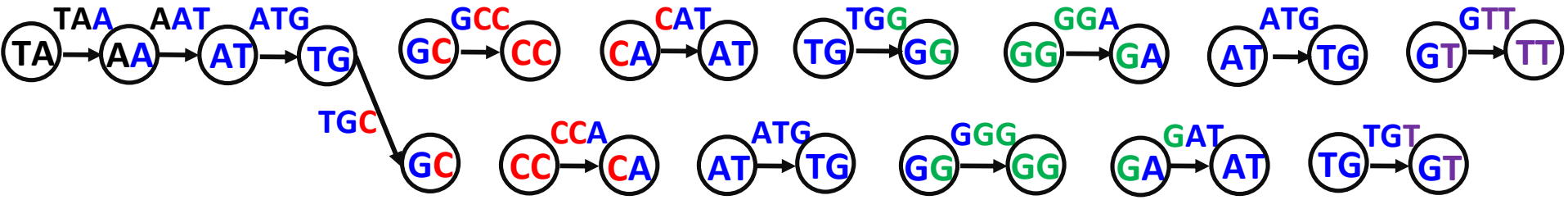


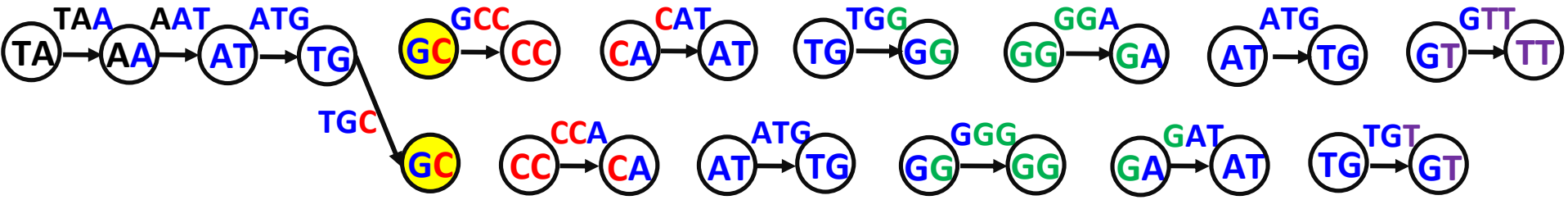

TAA
TA
AA
AAT
AT
ATG
AT
TG
GCC
GC
CC
CAT
CA
AT
TGG
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
TGC
TG
GC
CCA
CC
CA
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

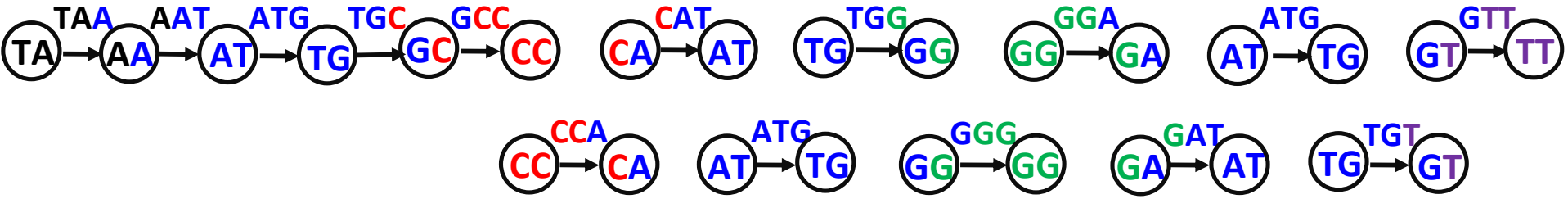

TAA
TA
AA
AAT
AT
ATG
AT
TG
GCC
GC
CC
CAT
CA
AT
TGG
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
TGC
TG
GC
CCA
CC
CA
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

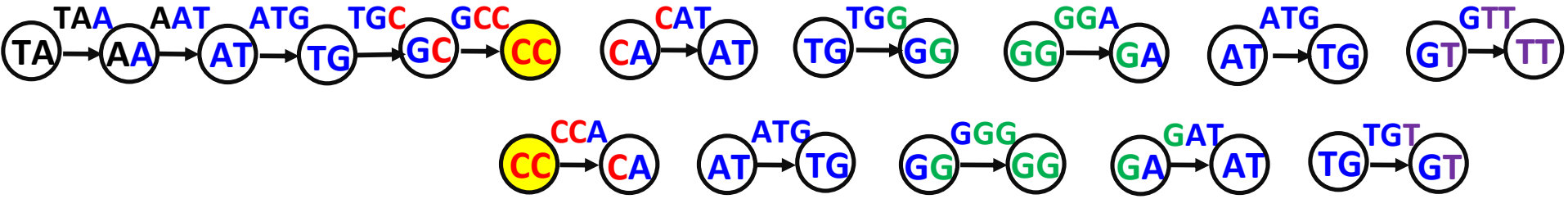

TAA
AAT
ATG
TA
AA
AT
TG
GCC
GC
CC
CAT
CA
AT
TGG
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
TGC
TG
GC
CCA
CC
CA
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

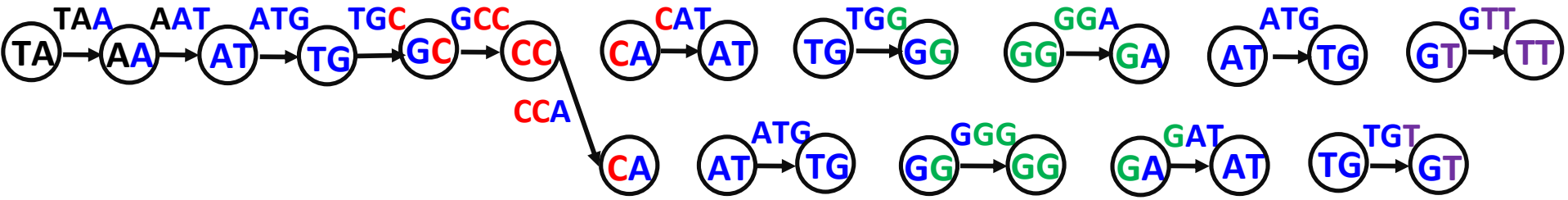

TAA
AAT
ATG
TA
AA
AT
TG
GCC
GC
CC
CAT
CA
AT
TGG
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
TGC
TG
GC
CCA
CC
CA
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

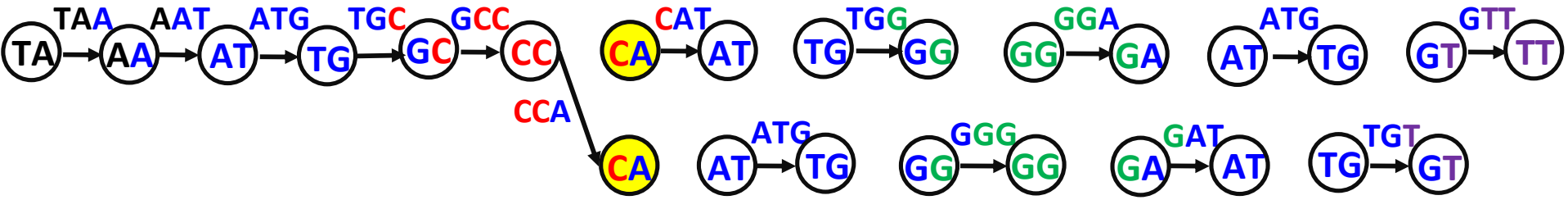

TAA
AAT
ATG
GCC
CAT
TGG
GGA
ATG
GTT
TA
AA
AT
TG
GC
CA
TG
GG
AT
GT
TGC
CCA
ATG
GGG
GAT
TGT
GC
CC
AT
GG
GA
TG
CA
TG
GG
AT
GT
TT

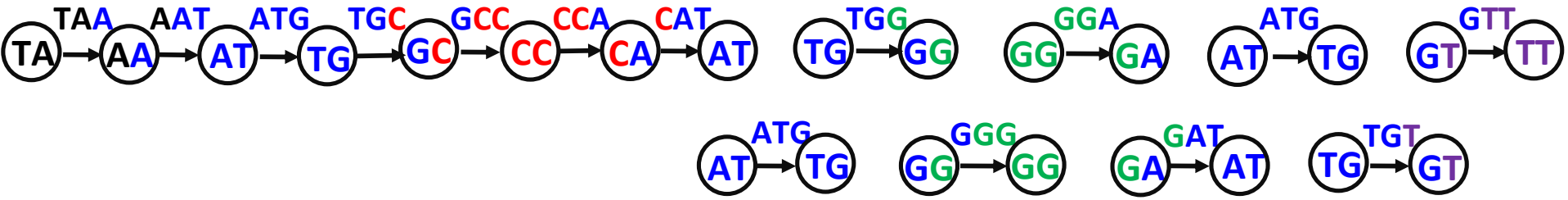

TAA
AAT
ATG
GCC
CAT
TGG
GGA
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
TGC
GC
CCA
CC
CA
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

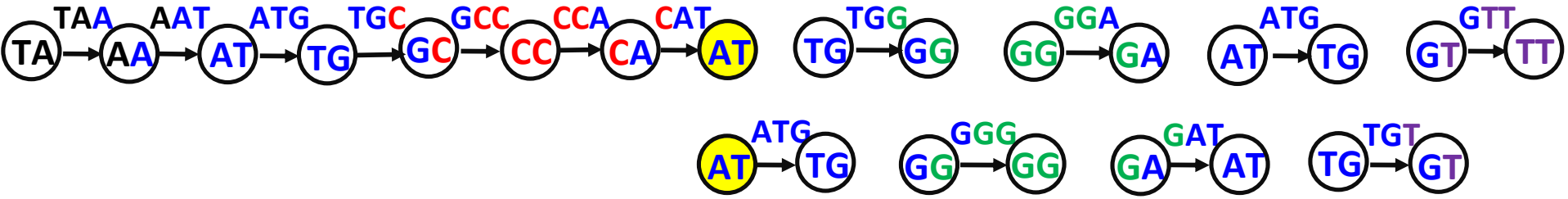

TAA
AAT
ATG
TGC
GCC
TA
AA
AT
TG
GC
CC
CAT
CA
AT
TGG
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
CCA
CC
CA
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

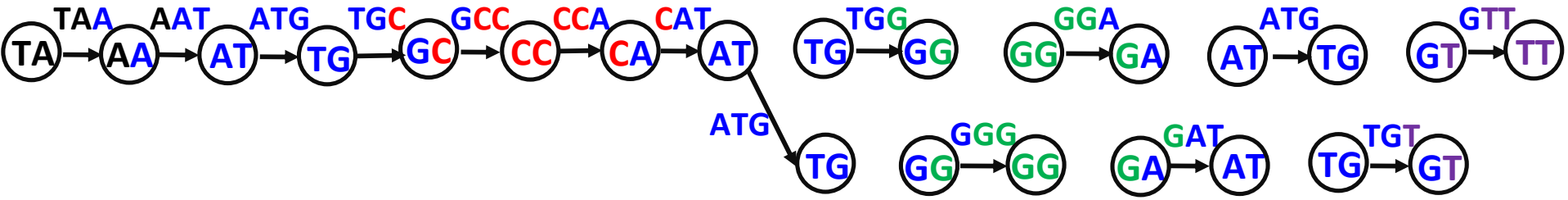

TAA
AAT
ATG
TGC
GCC
TA
AA
AT
TG
GC
CC
CAT
CA
AT
TGG
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
CCA
CC
CA
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

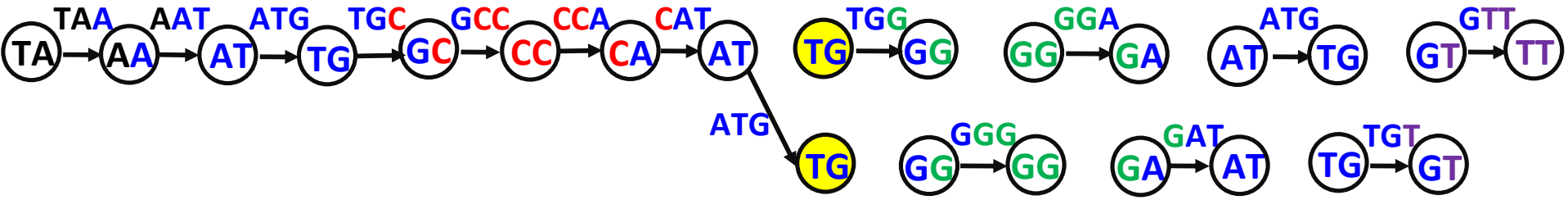

TAA
AAT
ATG
TGC
GCC
CAT
TGG
GGA
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
CCA
CA
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

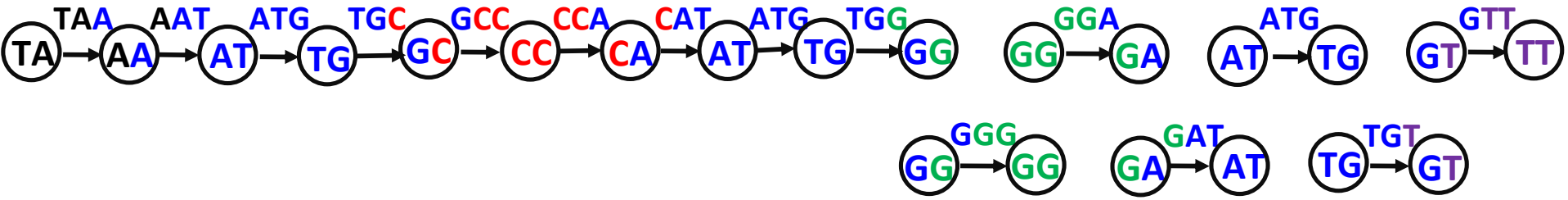

TAA
AAT
ATG
TGC
GCC
CAT
TGG
GGA
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
CCA
CA
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

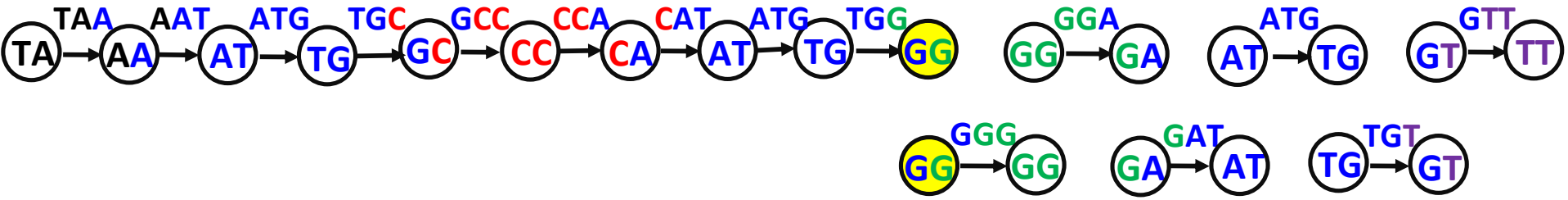

TAA
AAT
ATG
TGC
GCC
CCA
CAT
TA
AA
AT
TG
GC
CC
CA
AT
TGG
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

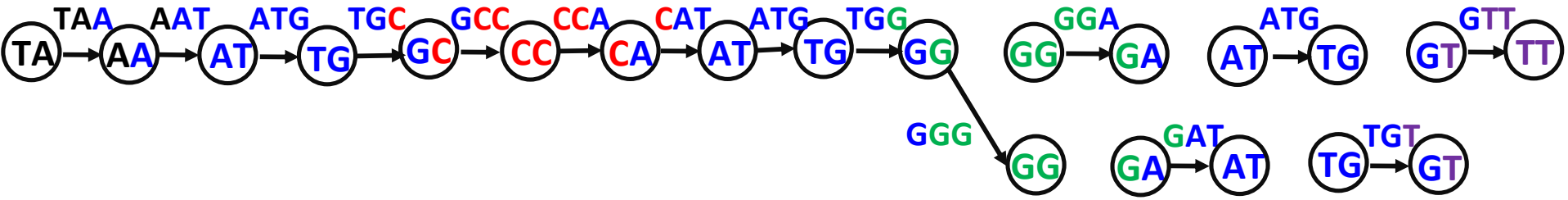

TAA
AAT
ATG
TGC
GCC
CCA
CAT
TA
AA
AT
TG
GC
CC
CA
AT
TGG
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
ATG
AT
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

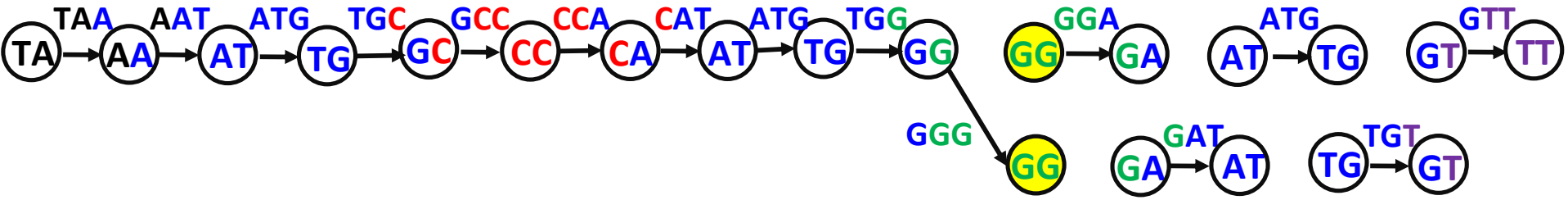

TAA
AAT
ATG
TGC
GCC
CCA
CAT
TA
AA
AT
TG
GC
CC
CA
AT
TGG
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
ATG
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

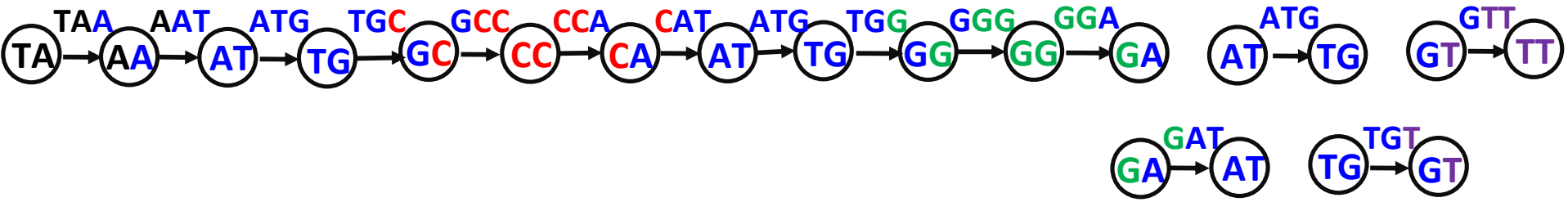

TAA
AAT
ATG
TGC
GCC
CCA
CAT
TGG
GGA
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
ATG
TG
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

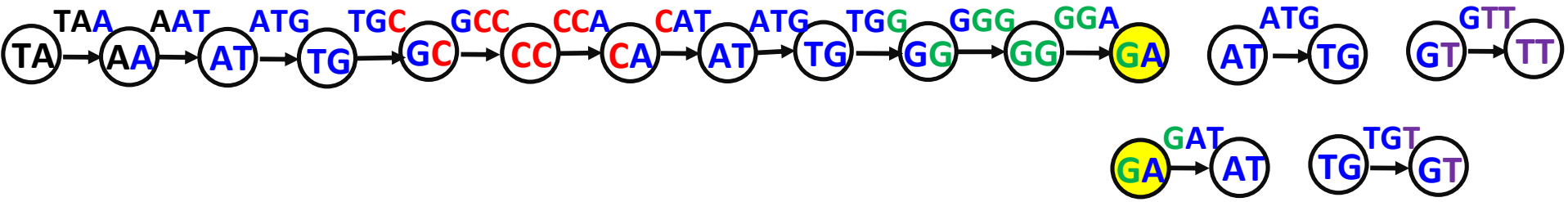

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

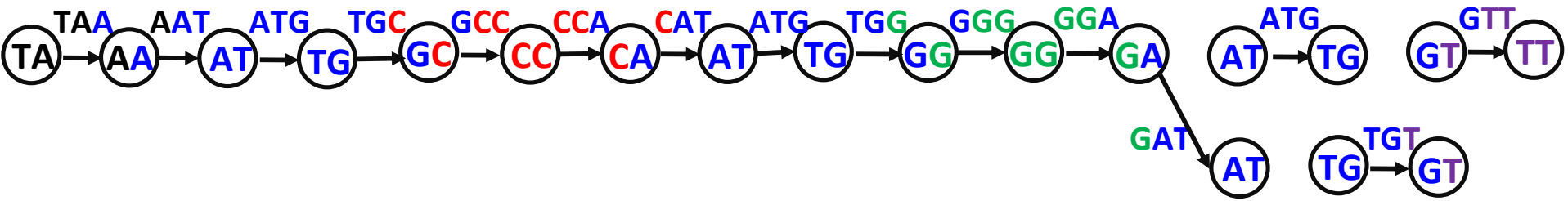

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GGA
GG
GA
ATG
AT
TG
GTT
GT
TT
GGG
GG
GG
GAT
GA
AT
TGT
TG
GT

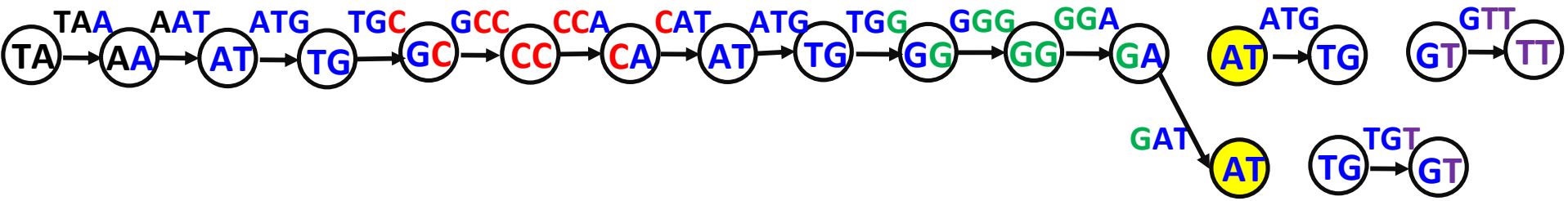

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGA
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
GGG
GAT
TGT
GG
GA
AT
TG
GT

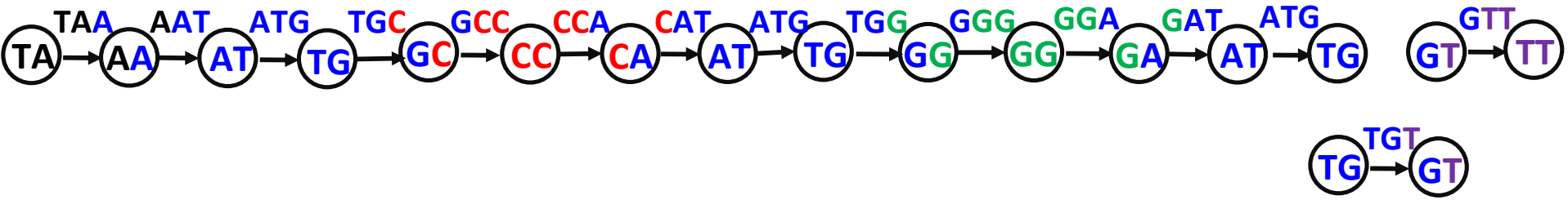

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGA
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
GGG
GG
GAT
GA
AT
TGT
TG
GT

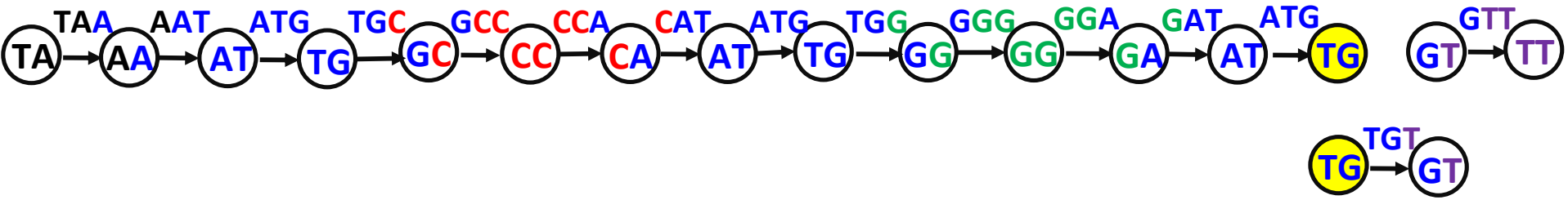

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
ATG
AT
TG
GTT
GT
TT
GAT
GA
AT
TGT
TG
GT

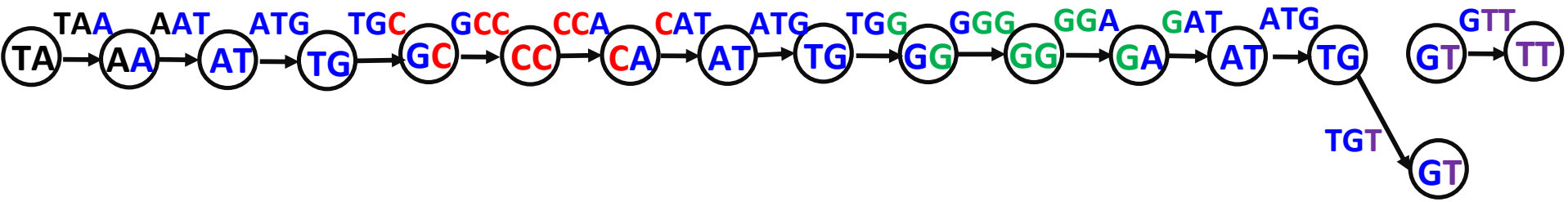

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
ATG
AT
TG
GTT
GT
TT
GAT
GA
AT
TGT
TG
GT

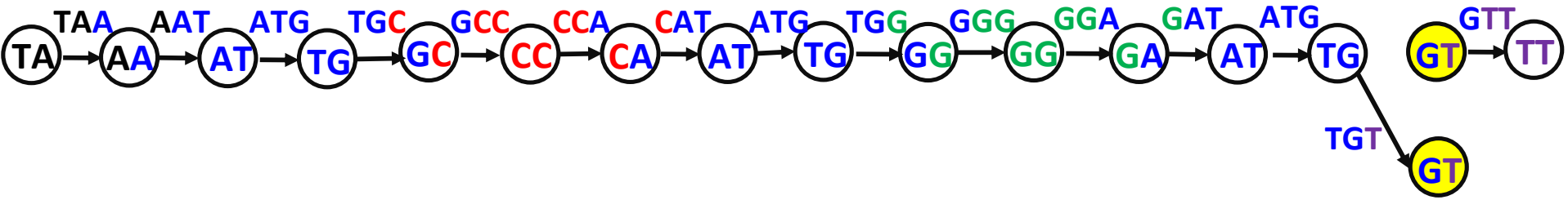

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
GAT
AT
TGT
TG
GT


TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
ATG
GTT
GAT
AT
TG
GT
TGT

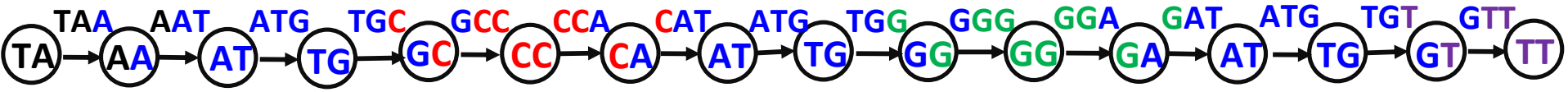

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
GAT
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
TGT
TG
GT


TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
GAT
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
TGT
TG
GT

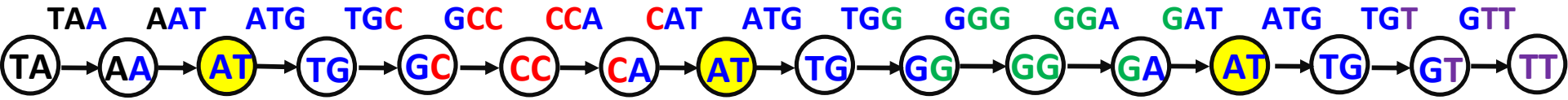

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
GAT
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
TGT
GT

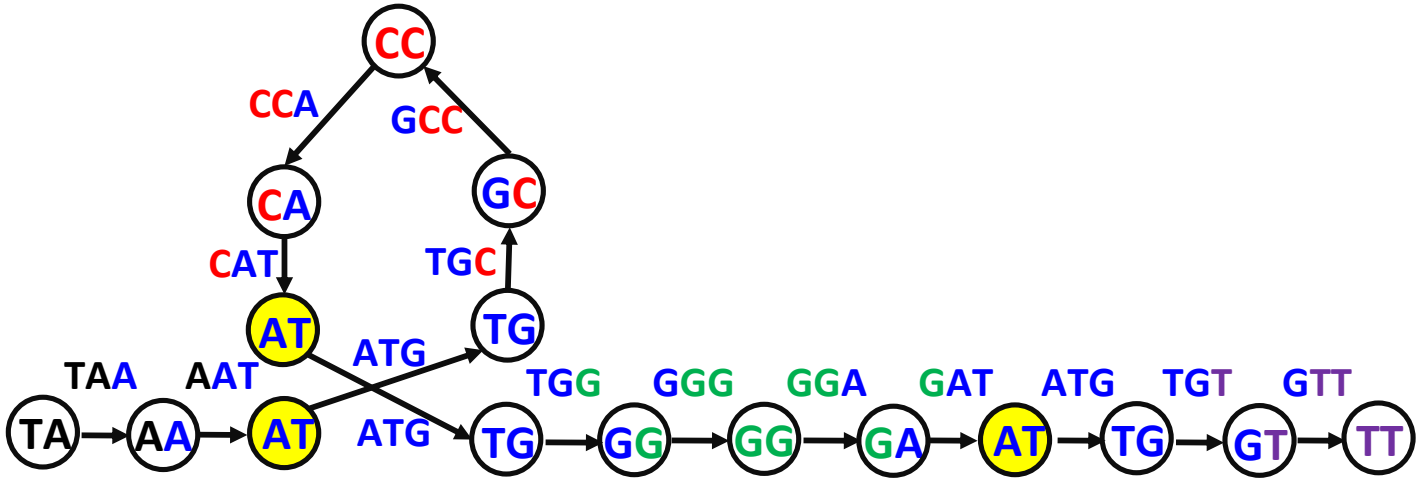

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
GAT
ATG
GTT
TA
AA
AT
TG
GC
CC
CA
AT
TG
GG
GG
GA
AT
TG
GT
TT
TGT
GT

Lepljenje nije završeno

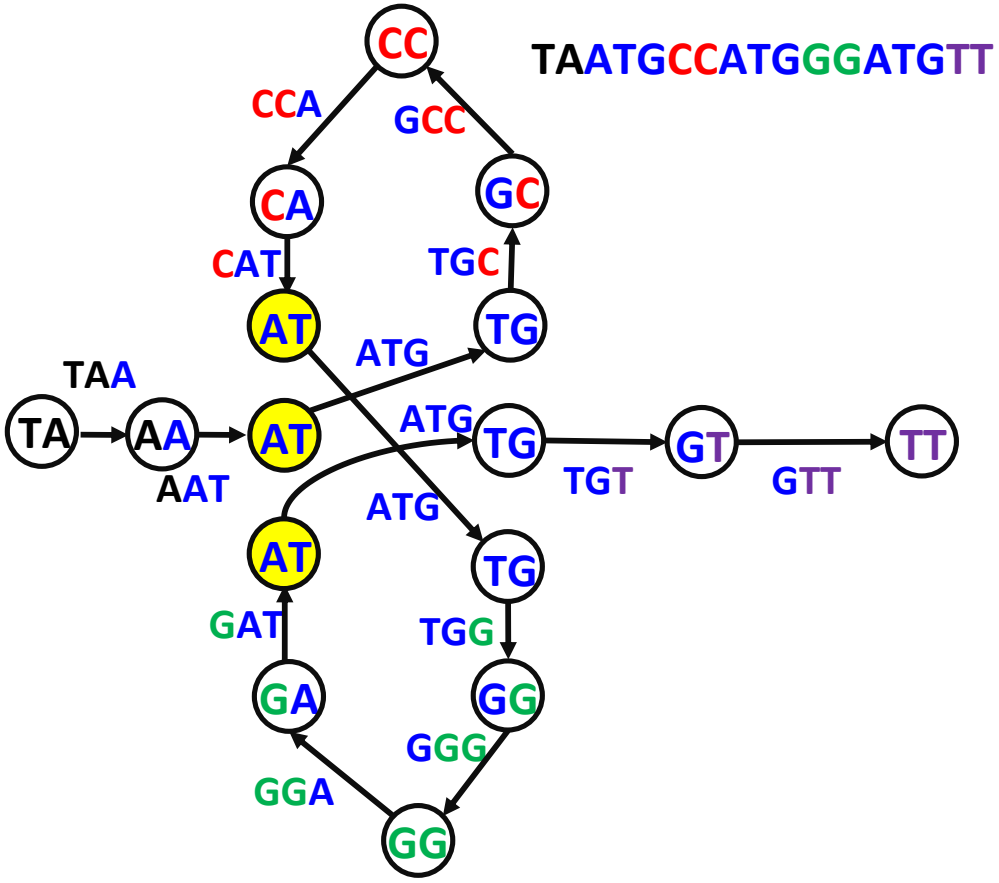


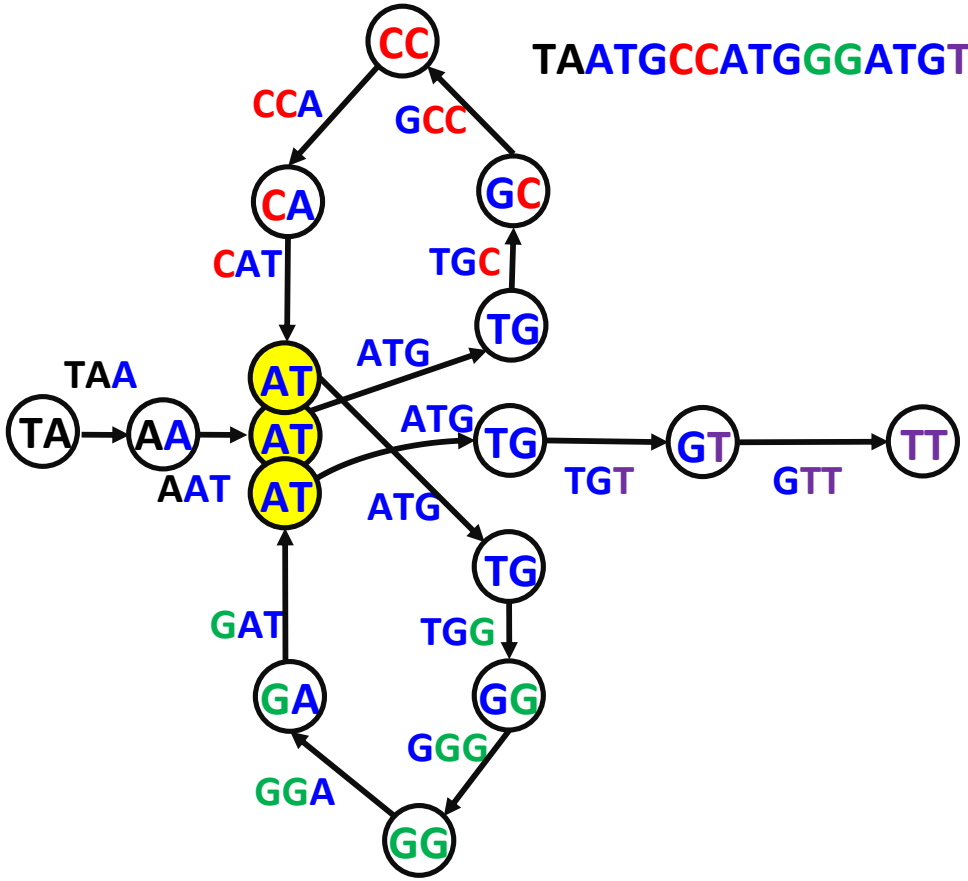
Zalepimo identično obeležene čvorove



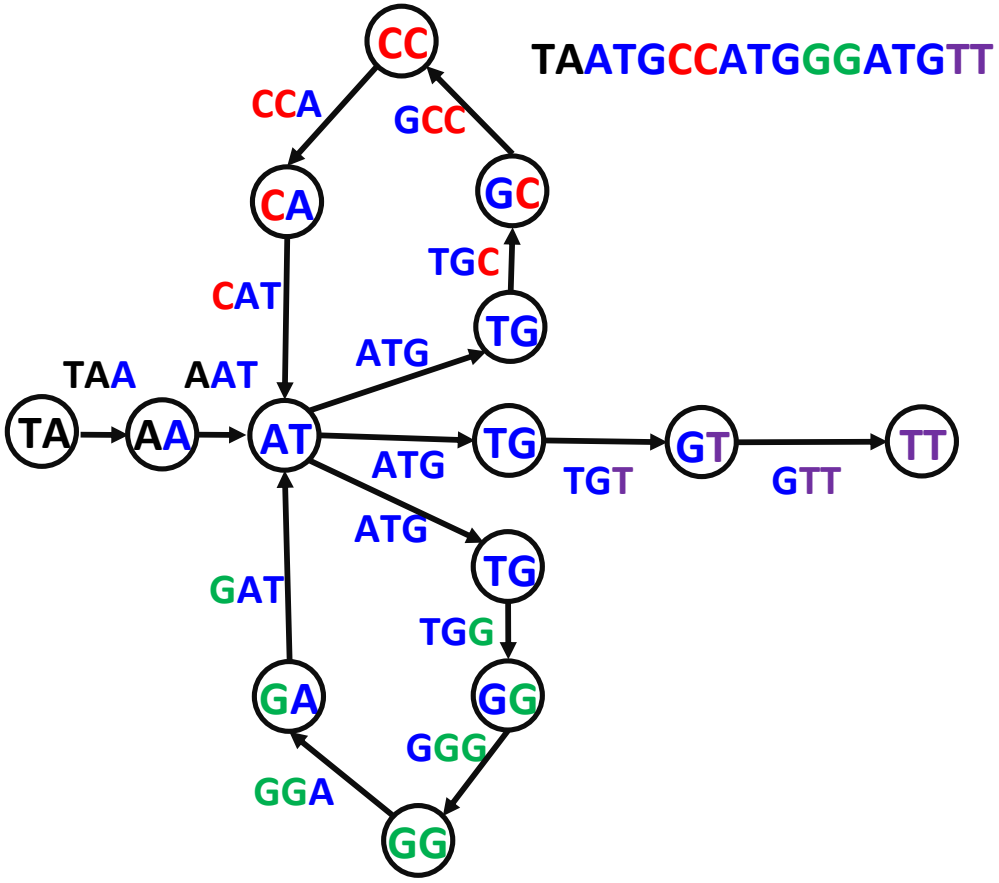


Zalepimo identično obeležene čvorove

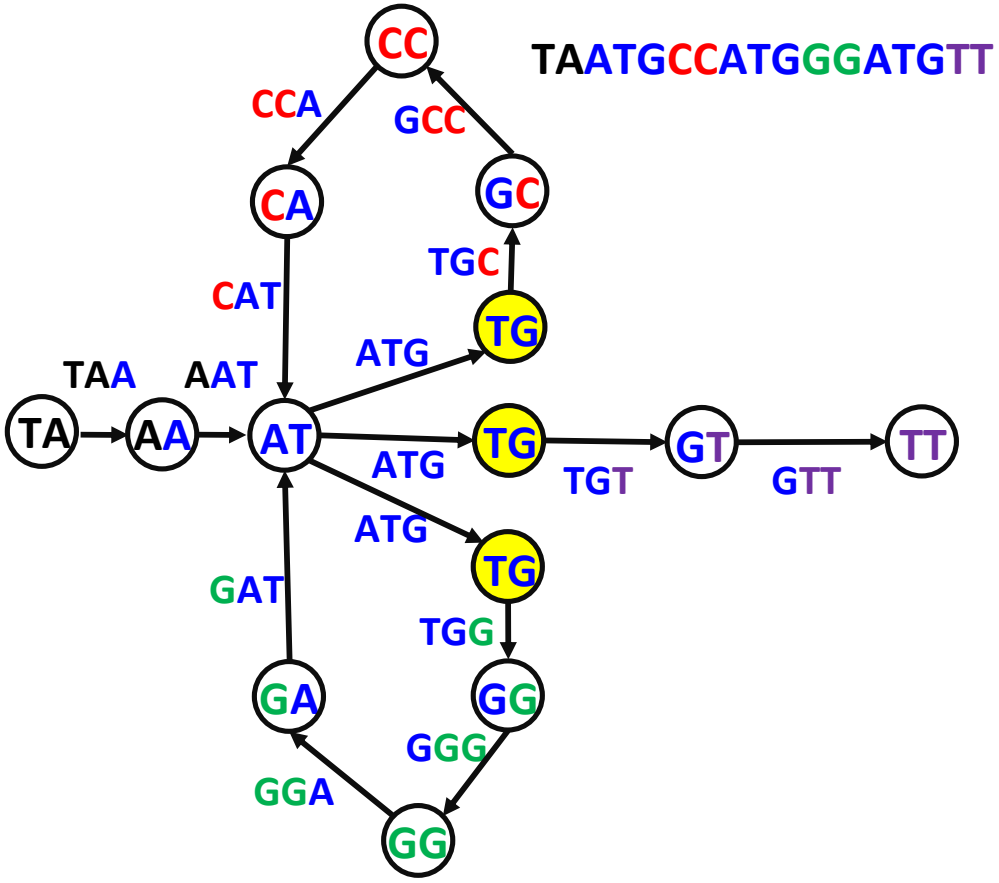



TAATGCCATGGGATGTT
CC
CCA
GCC
CA
GC
CAT
TGC
TG
ATG
TAA
AT
TA
AA
AT
ATG
TG
GT
TT
AAT
AT
TGT
GTT
ATG
TG
GAT
TGG
GA
GG
GGG
GGA
GG

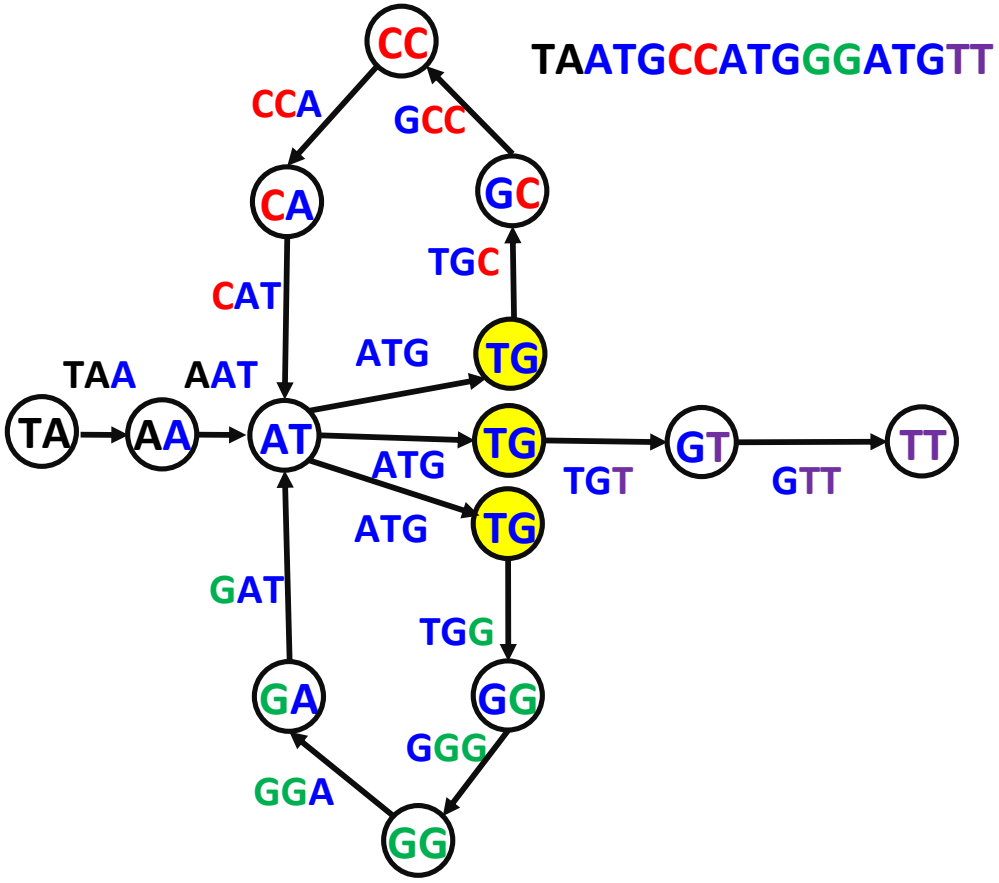
Zalepimo identično obeležene čvorove



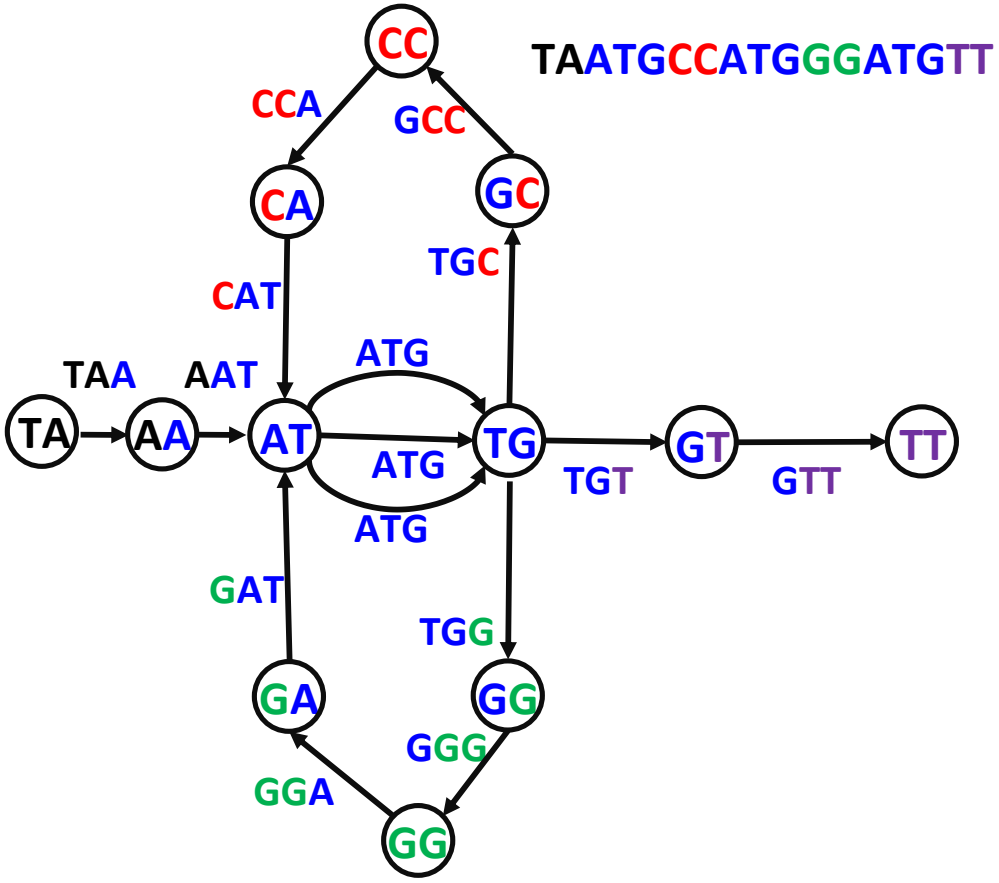
Zalepimo identično obeležene čvorove



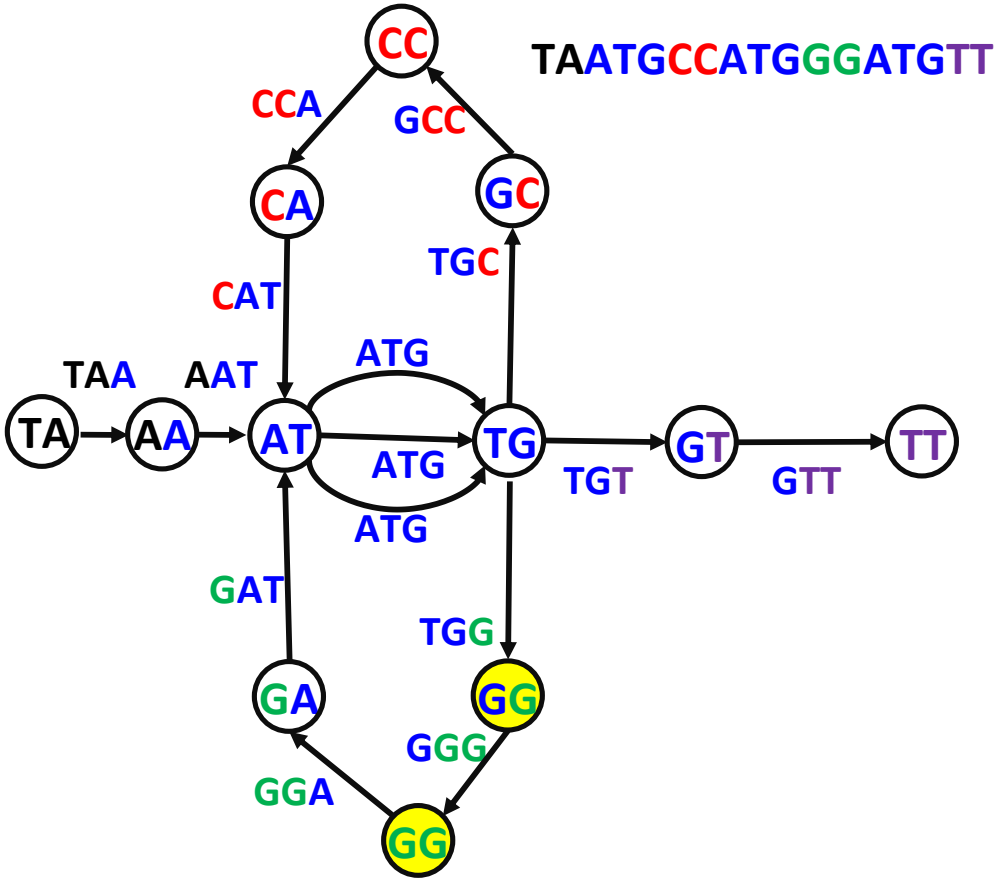
Zalepimo identično obeležene čvorove



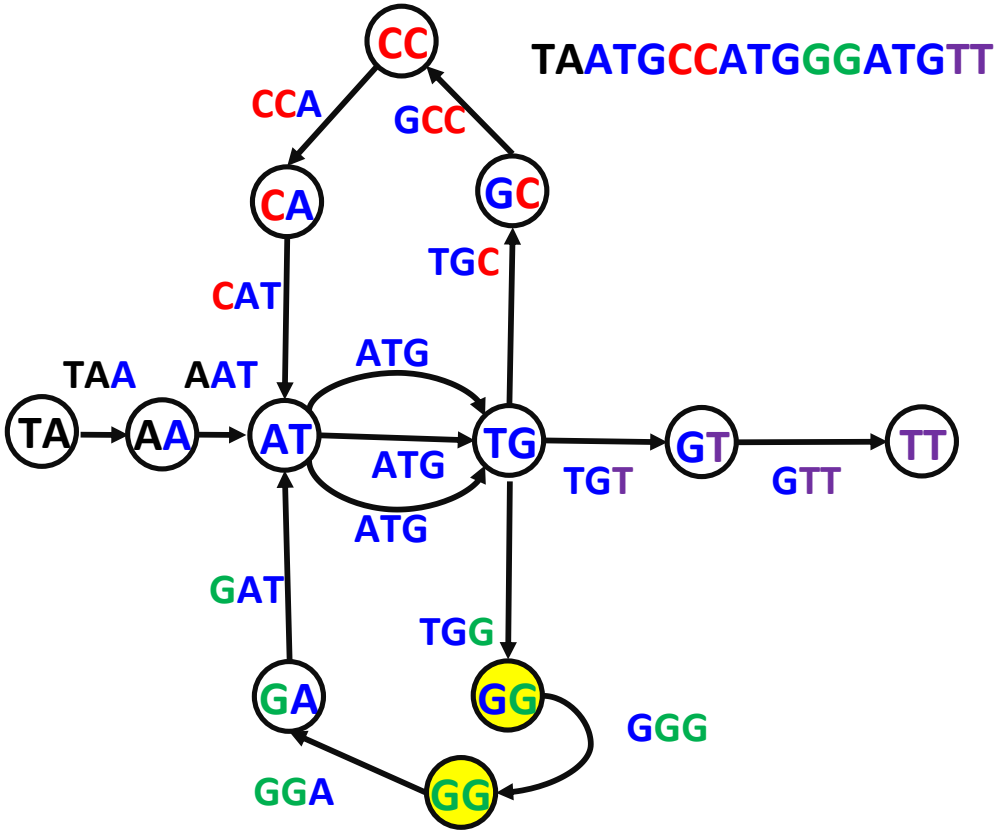
Zalepimo identično obeležene čvorove



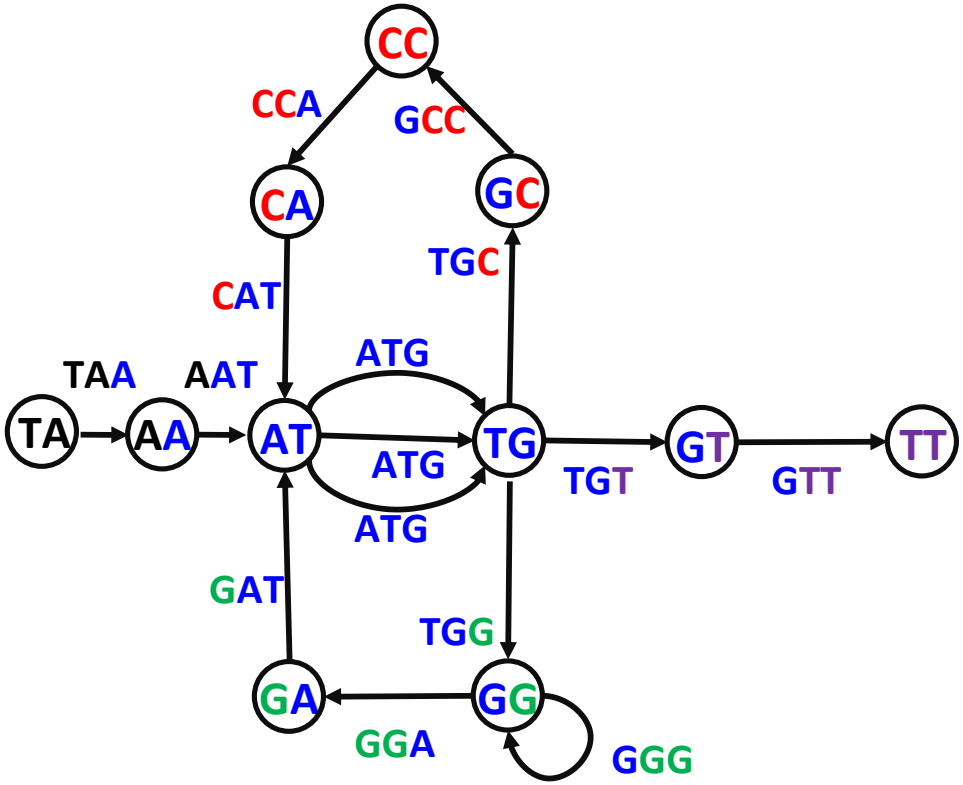
Zalepimo identično obeležene čvorove



Zalepimo identično obeležene čvorove



Isti De Brujinov graf: *DeBruin(Genome)= DeBruin(Genome Composition)*



Konstrukcija De Brujinovog grafa

De Brujinov graf na osnovu kolekcije *k*-grama:

- Svaka grana je označena jednim *k*-gramom
- Svaki čvor je označen prefiksom/sufiksom izlazne/ulazne grane
- Zalepljeni su svi čvorovi sa identičnim oznakama.

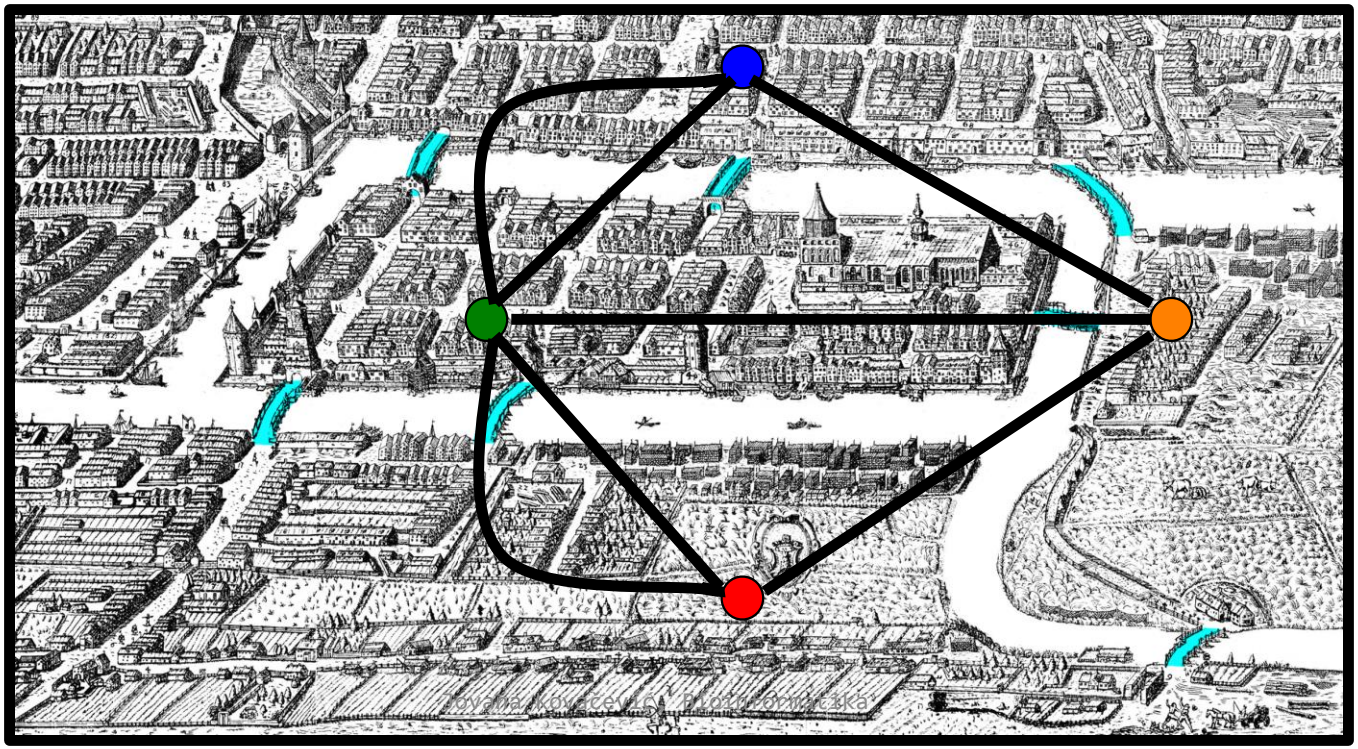
Pregled

- Šta je sekvencioniranje genoma?
- Eksplozija u štampariji
- Problem rekonstrukcije niske
- Rekonstrukcija niske kao problem Hamiltonove putanje
- Rekonstrukcija niske kao problem Ojlerove putanje
- De Brujinovi grafovi
- **Ojlerova teorema**
- Spajanje parova očitavanja
- U realnosti

Problem Ojlerovog *ciklusa*

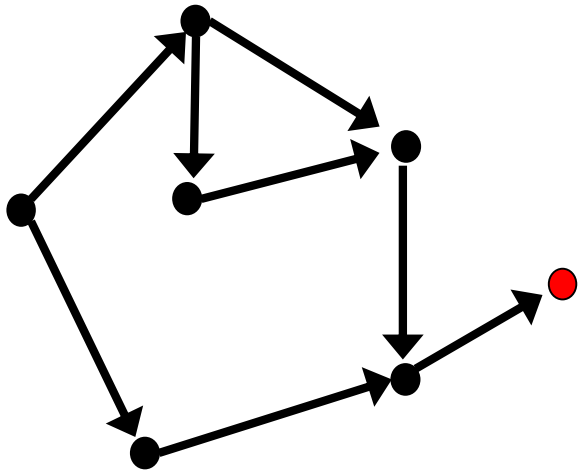
Problem Ojlerovog ciklusa. Pronaći Ojlerov ciklus u grafu.

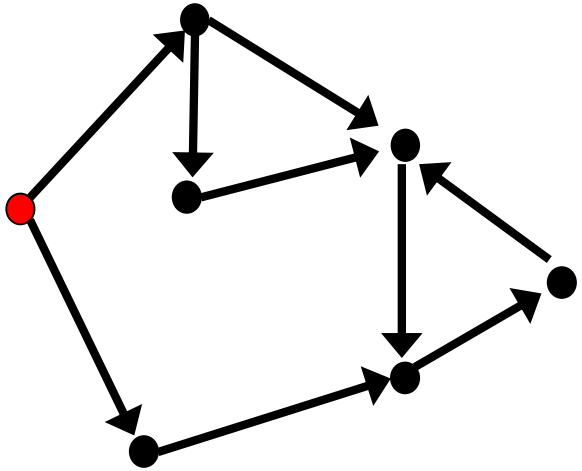
- **Ulaz.** Graf.
- **Izlaz.** Ciklus koja posećuje svaku granu u grafu tačno jednom.

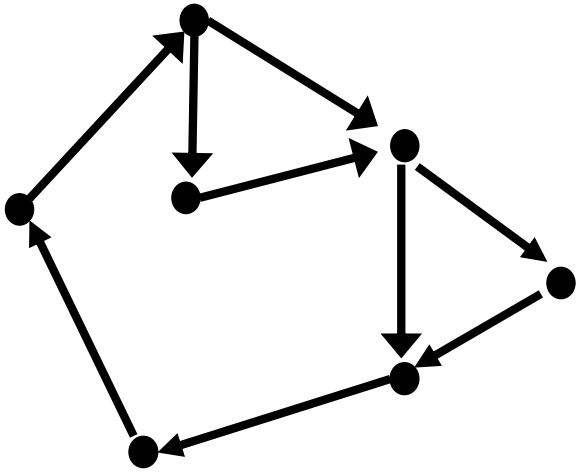


Kažemo da je graf **Ojlerov** ako sadrži Ojlerov ciklus.

Da li je ovaj graf Ojlerov?

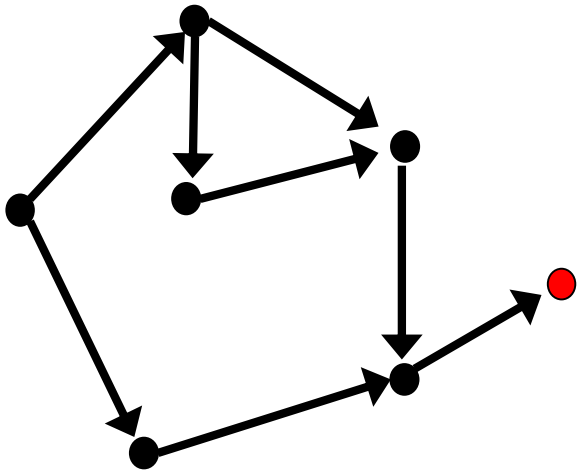


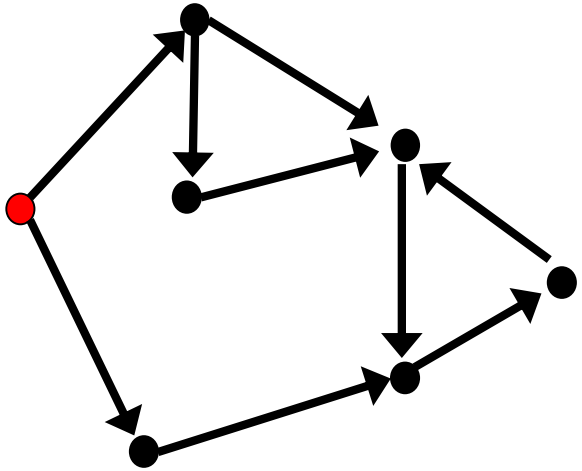


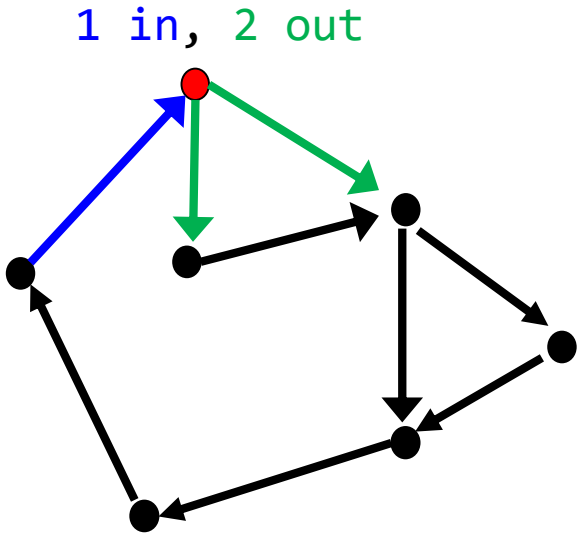


Kažemo da je graf **Ojlerov** ako sadrži Ojlerov ciklus.

Da li je ovaj graf Ojlerov?







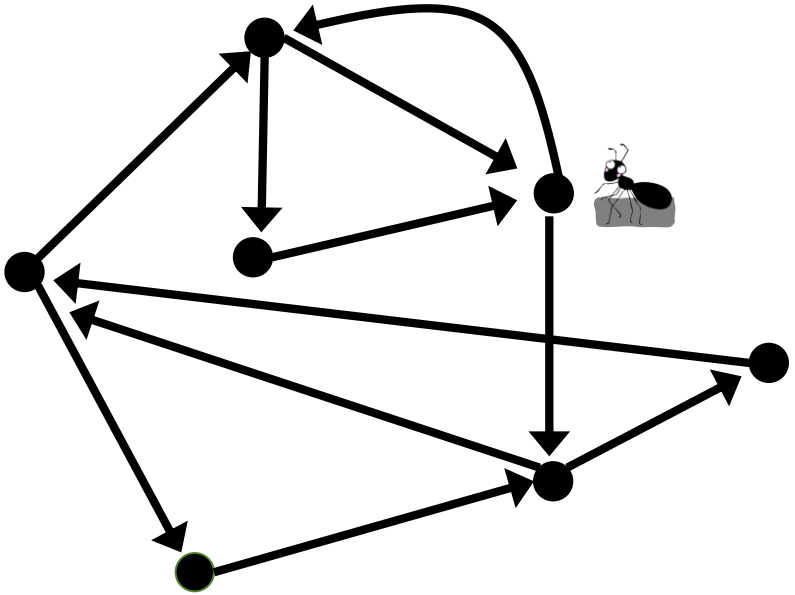
Kažemo da je graf balansiran ako za svaki čvor važi
indegree = outdegree

Ojlerova teorema

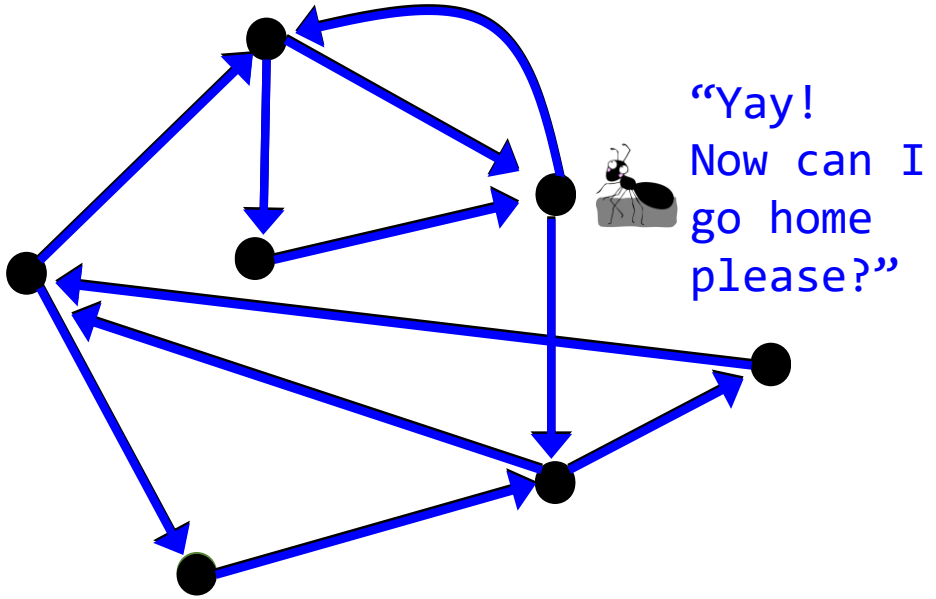
- Svaki Ojlerov graf je balansiran
- Svaki povezan i balansiran graf je Ojlerov
  - Kažemo da je graf povezan ako za ma koja dva čvora postoji putanja koja ih povezuje.

Kako bi mrav dokazao Ojlerovu teoremu?

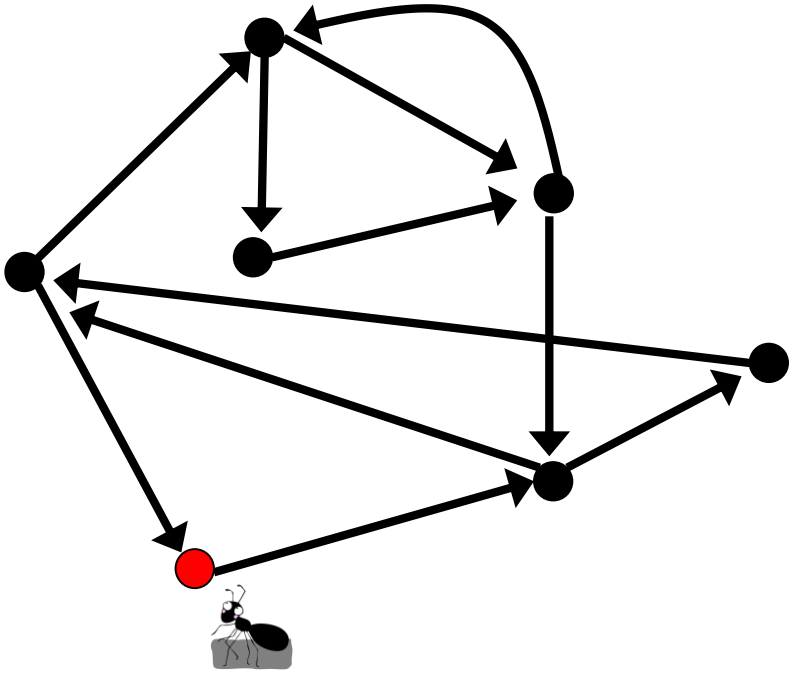
Mrav na slučajan način bira grane kojima će se kretati u grafu. **Ne može da obiđe istu granu dvaput!**



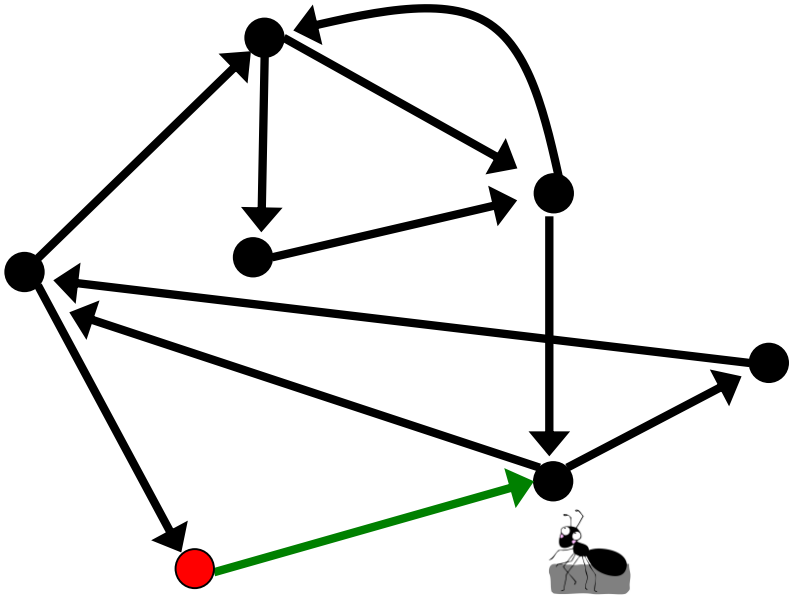
Veoma pametan mrav



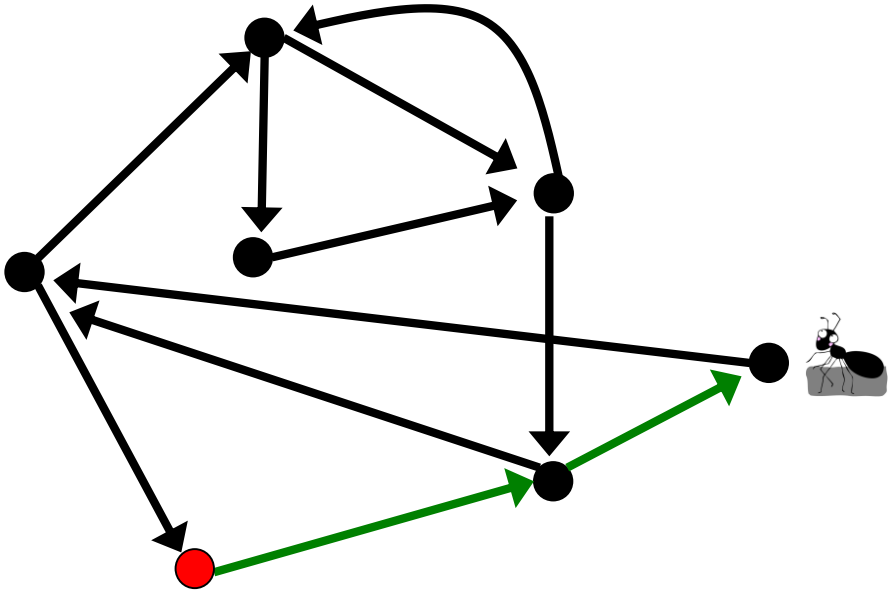
Manje pametan mrav



... obilazi ...

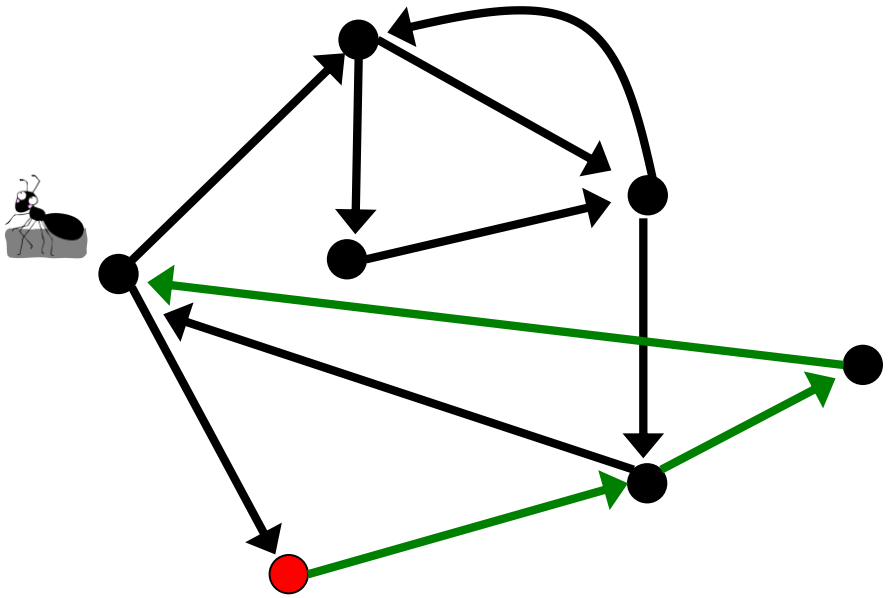


... obilazi ...

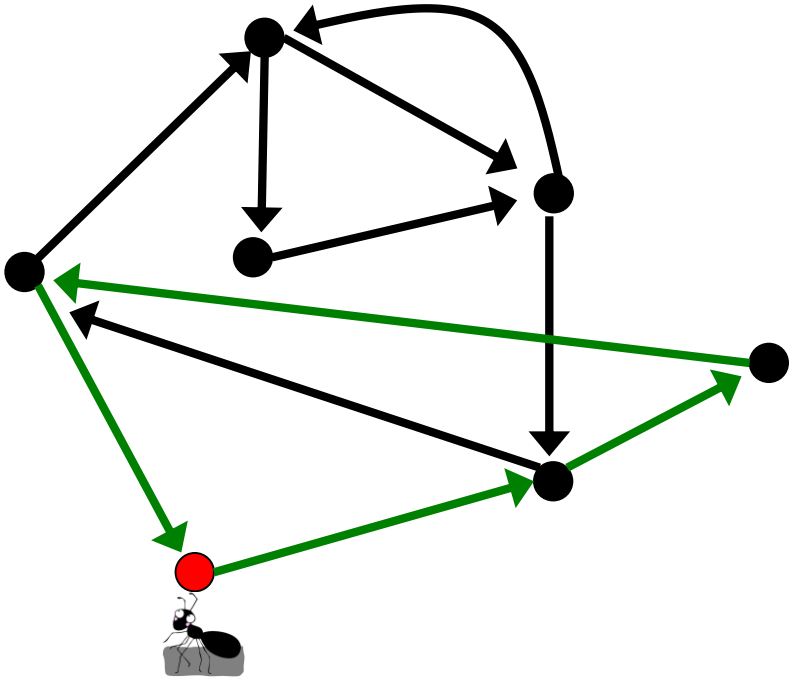


... obilazi ...

Da li može da se zaglavi? **U kom čvoru?**

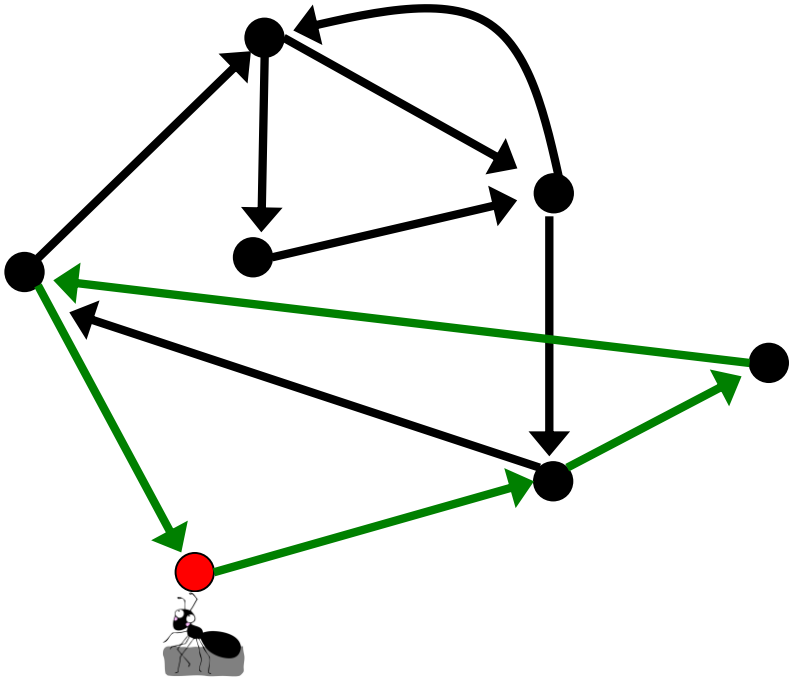


Mrav može da se zaglavi samo u čvoru iz kog je počeo obilazak



Mrav je kreirao ciklus ali još nije dokazao Ojlerovu teoremu

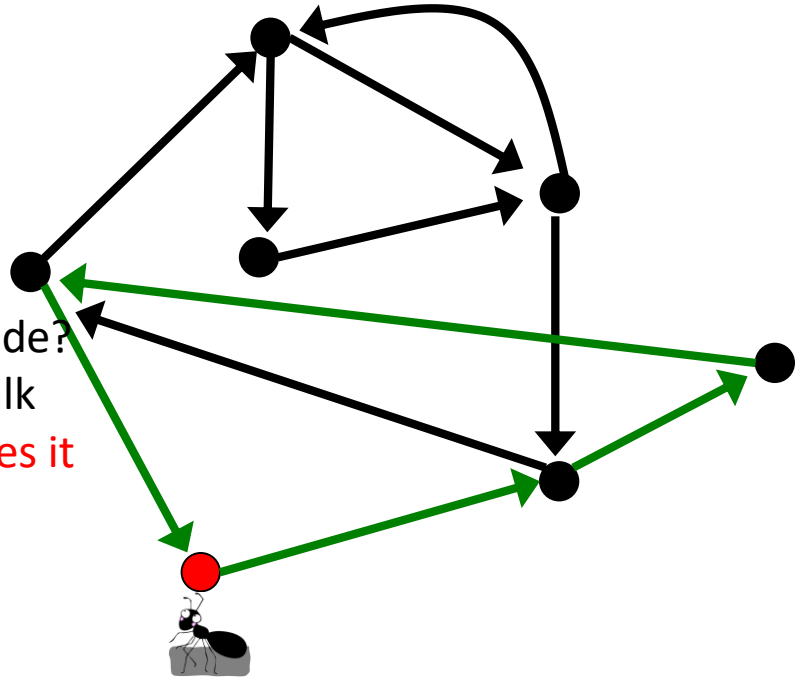
Konstruisani ciklus nije Ojlerov. **Možemo li da ga uvećamo?**



Hajde da započnemo obilazak u nekom drugom čvoru iz zelenog ciklusa

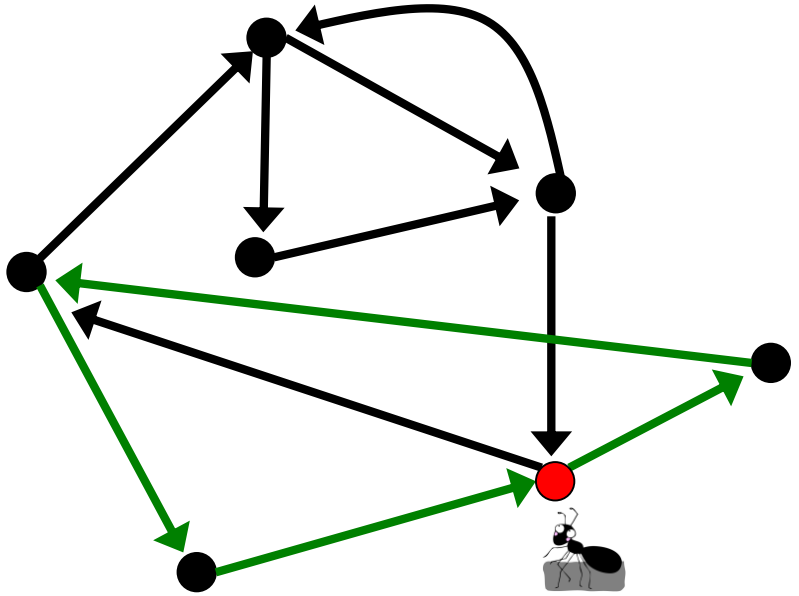
U kom? U onom koji ima neposećene grane.

"Why should I start at a different node? Backtracking? I'm not evolved to walk backwards! And what difference does it make???"



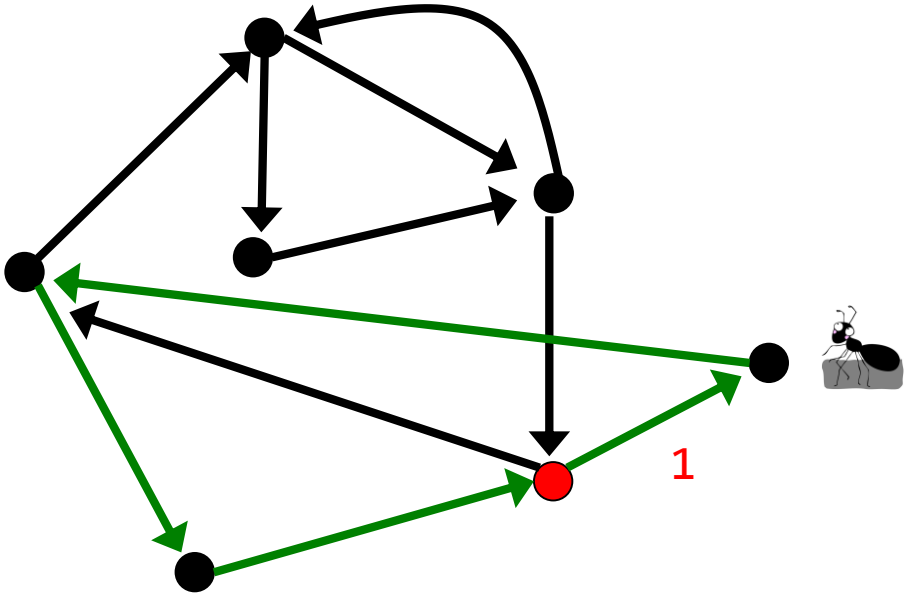
Nove instrukcije za mrava:

Počni od čvora koji ima neposećenu granu, obiđi već konstruisani zeleni ciklus i vrati se u početni čvor



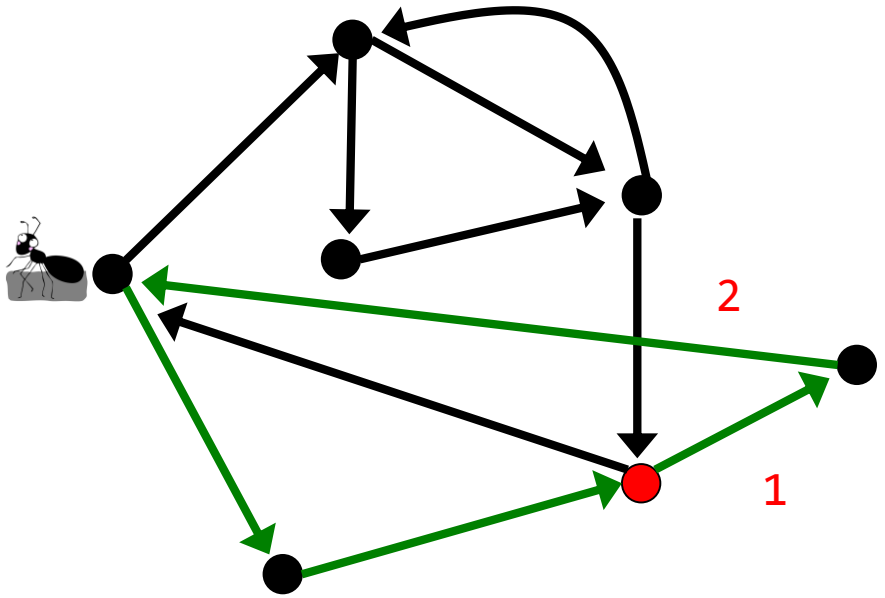
Mrav obilazi prethodno konstruisani ciklus

Počni od čvora koji ima neposećenu granu, obiđi već konstruisani zeleni ciklus i vrati se u početni čvor



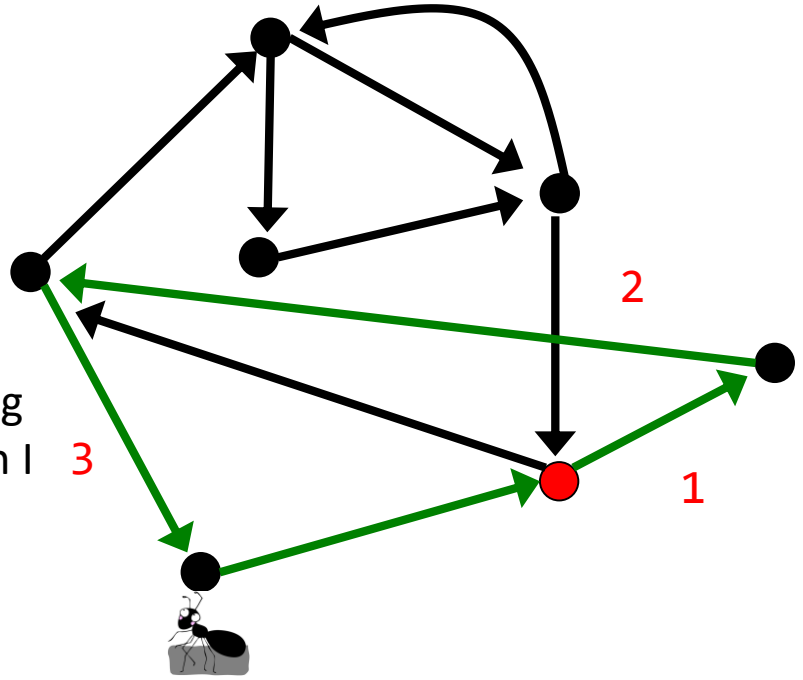
Mrav obilazi prethodno konstruisani ciklus

Počni od čvora koji ima neposećenu granu, obiđi već konstruisani zeleni ciklus i vrati se u početni čvor



Mrav obilazi prethodno konstruisani ciklus

Počni od čvora koji ima neposećenu granu, obiđi već konstruisani zeleni ciklus i vrati se u početni čvor

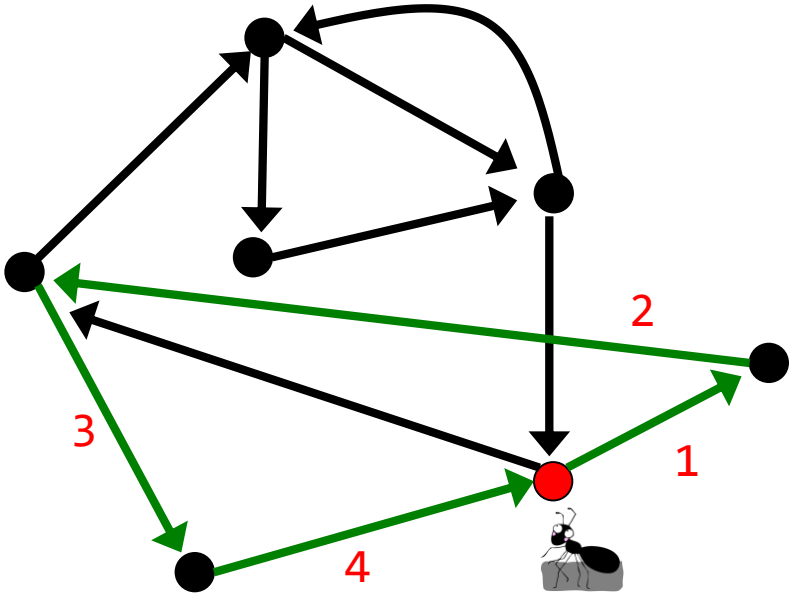


"Why do I have to walk along the same cycle again??? Can I see something new?"

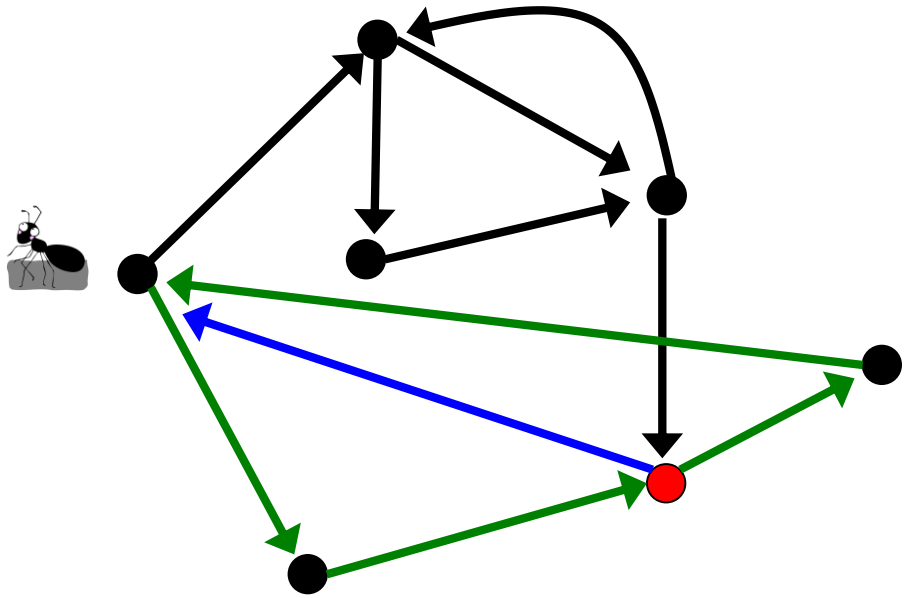
Mrav se vratio nazad ali može da nastavi da obilazi!

Počni od čvora koji ima neposećenu granu, obiđi već konstruisani zeleni ciklus i vrati se u početni čvor

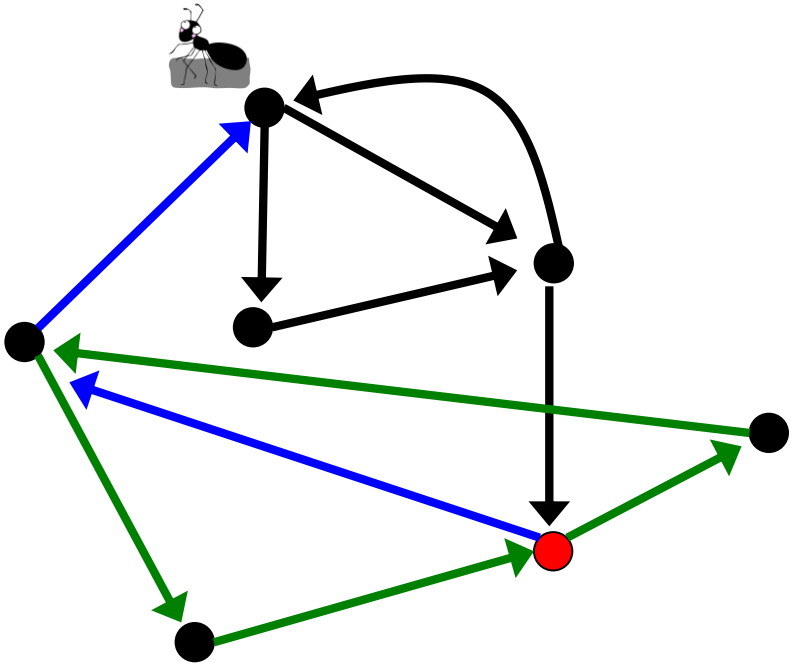
Nakon završenog ciklusa, nastavi obilazak tako što ćeš posetiti neku granu koja nije posećena ranije. Ako ih ima više, odaberi jednu na slučajan način.



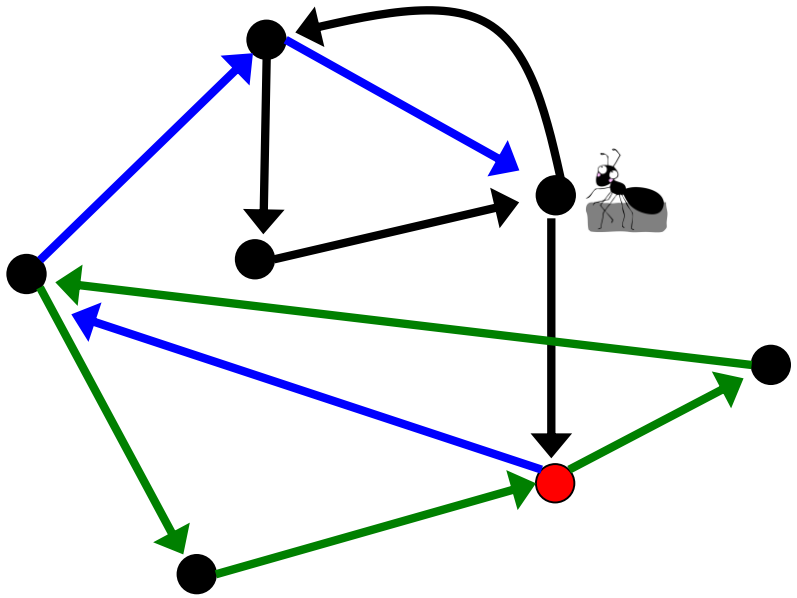
Uvećavamo prethodno konstruisani ciklus



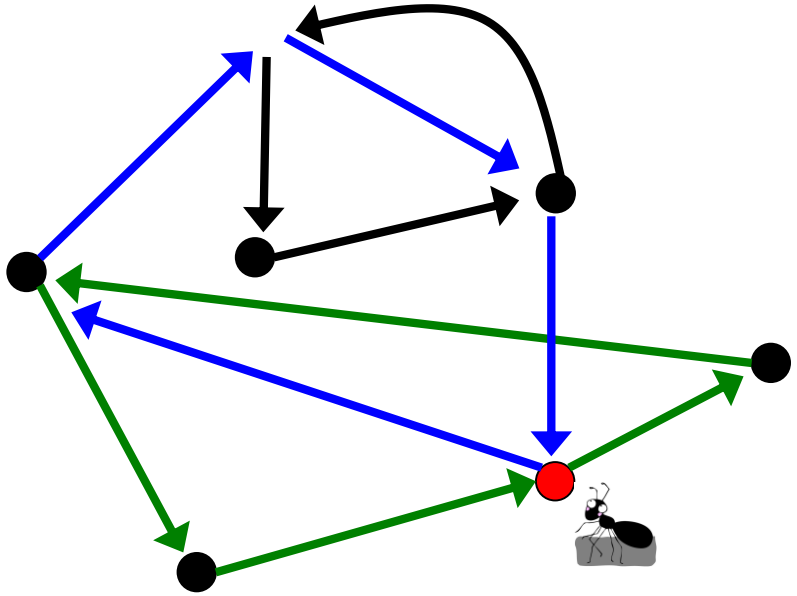
Uvećavamo prethodno konstruisani ciklus



Uvećavamo prethodno konstruisani ciklus



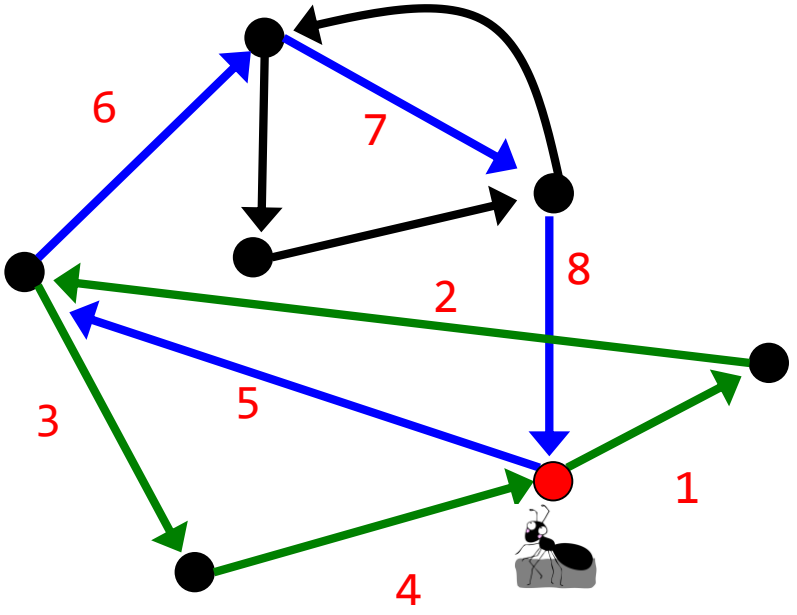
Uvećavamo prethodno konstruisani ciklus



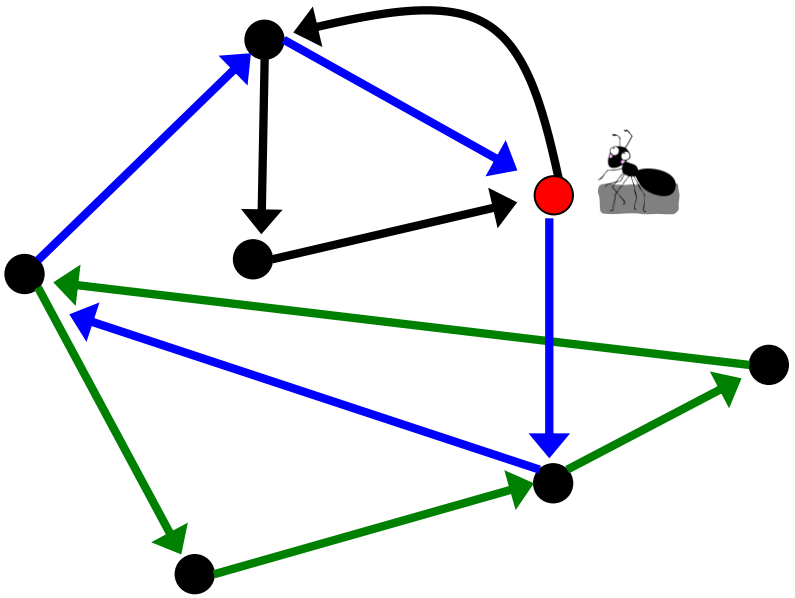
Mrav se ponovo zaglavio!

Konstruisani zeleno-plavi ciklus i dalje nije Ojlerov. Da li možemo da ga uvećamo?

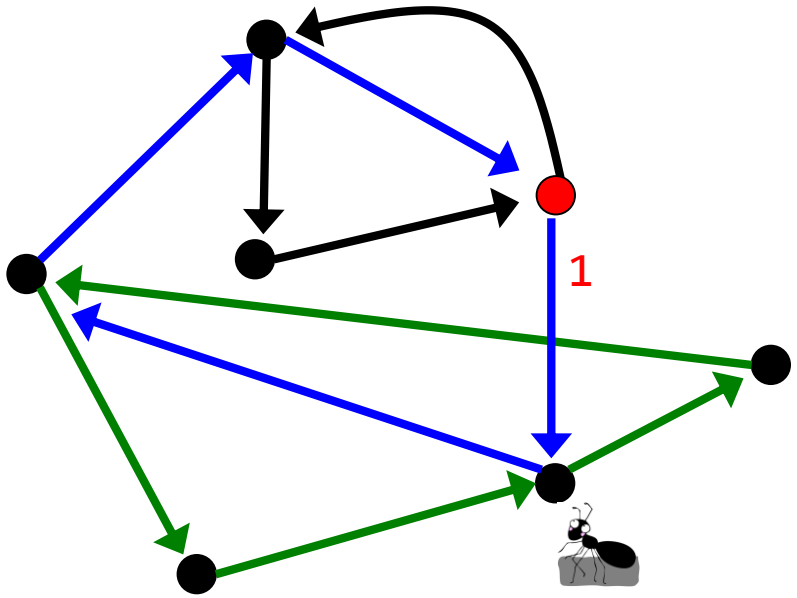
Mrav treba da obiđe konstruisani ciklus počev od drugog čvora. Od kog?



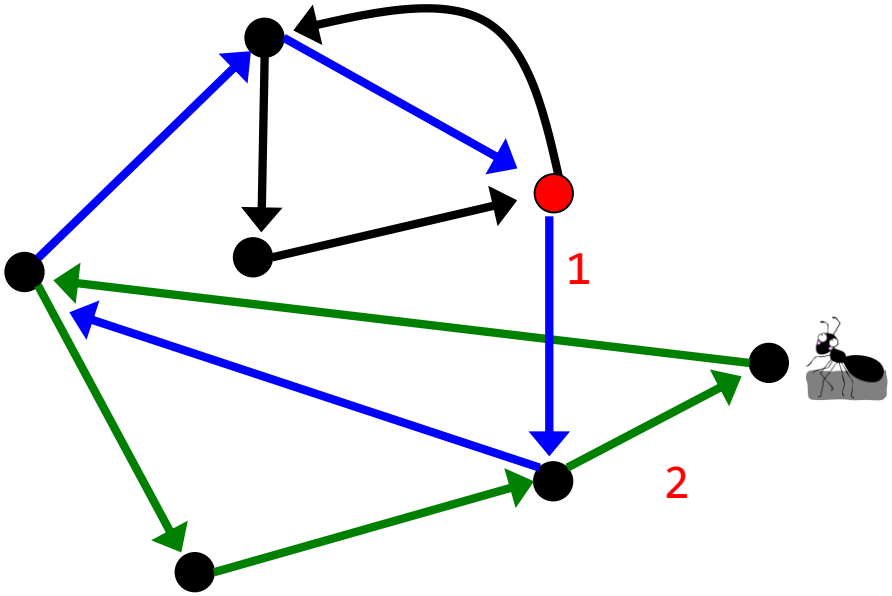
Počinjemo od novog čvora, ponovo…



Obilazimo prethodno konstruisan zeleno-plavi ciklus

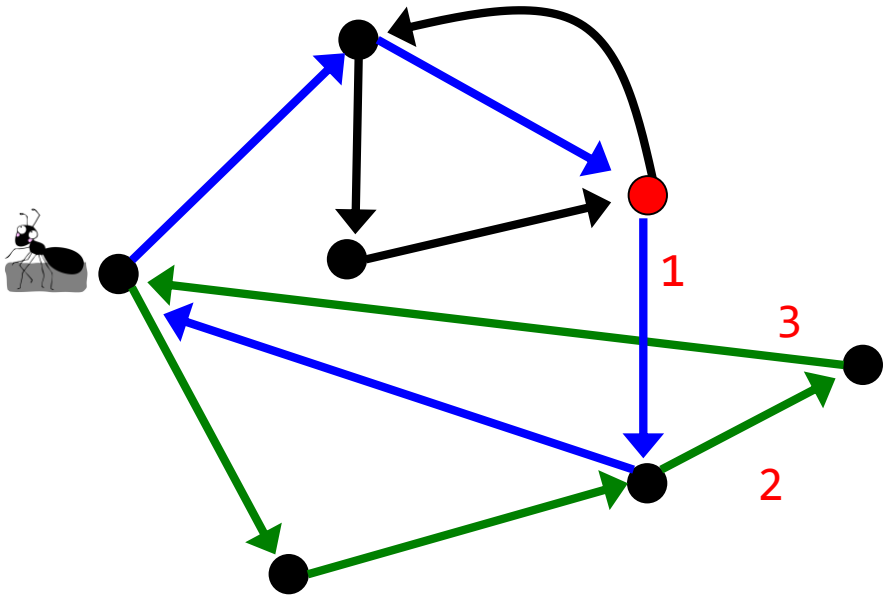


Obilazimo prethodno konstruisan zeleno-plavi ciklus

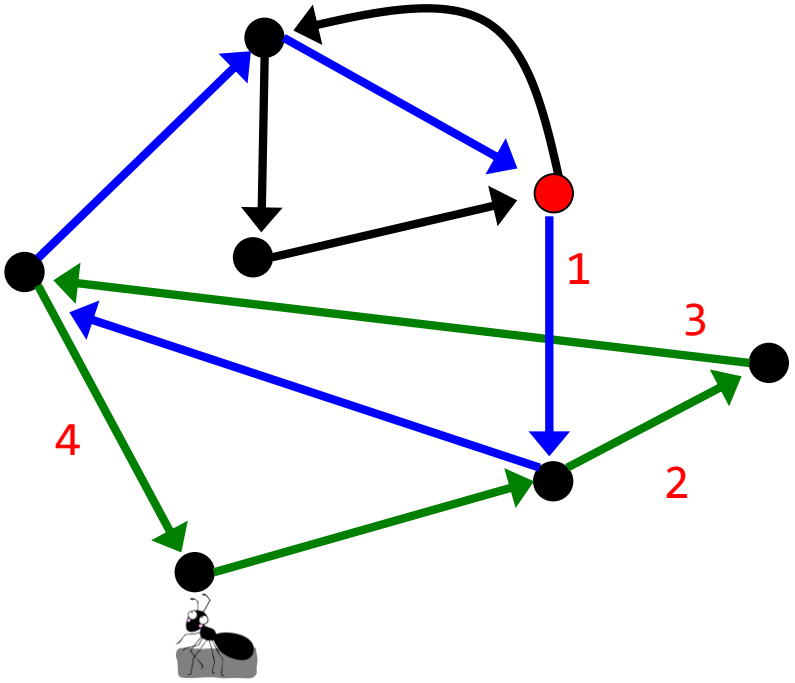


Obilazimo prethodno konstruisan zeleno-plavi ciklus

"I hate to traverse the same cycle! What difference does it make where I start my walk???

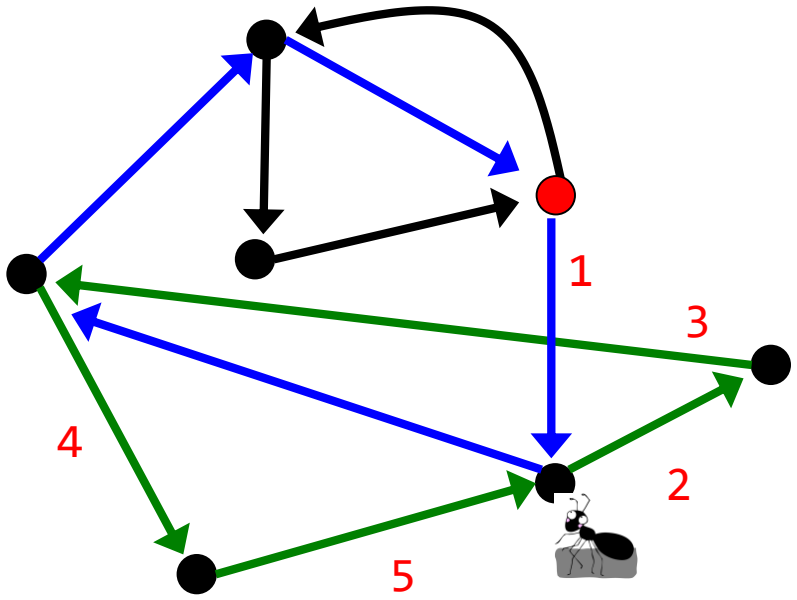


Obilazimo prethodno konstruisan zeleno-plavi ciklus

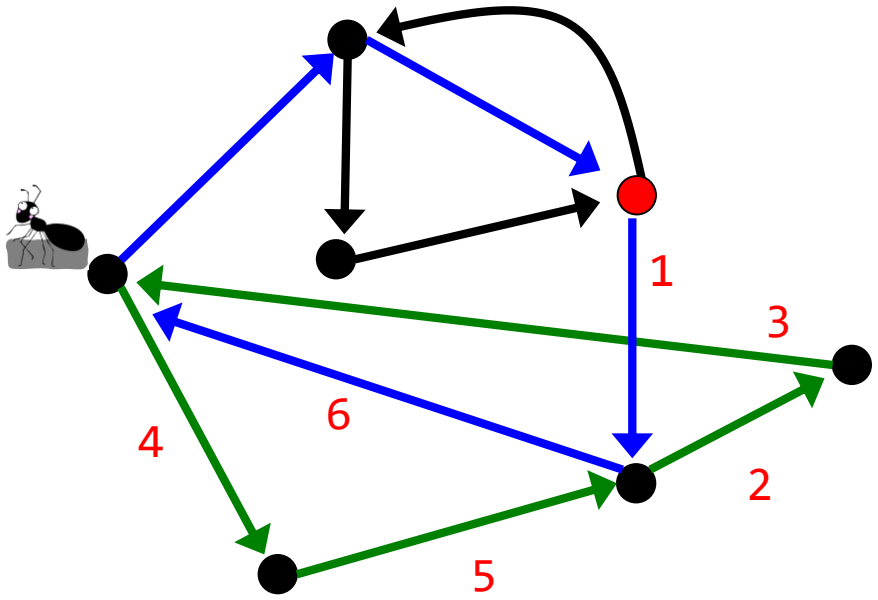


"These instructions are stupid…"

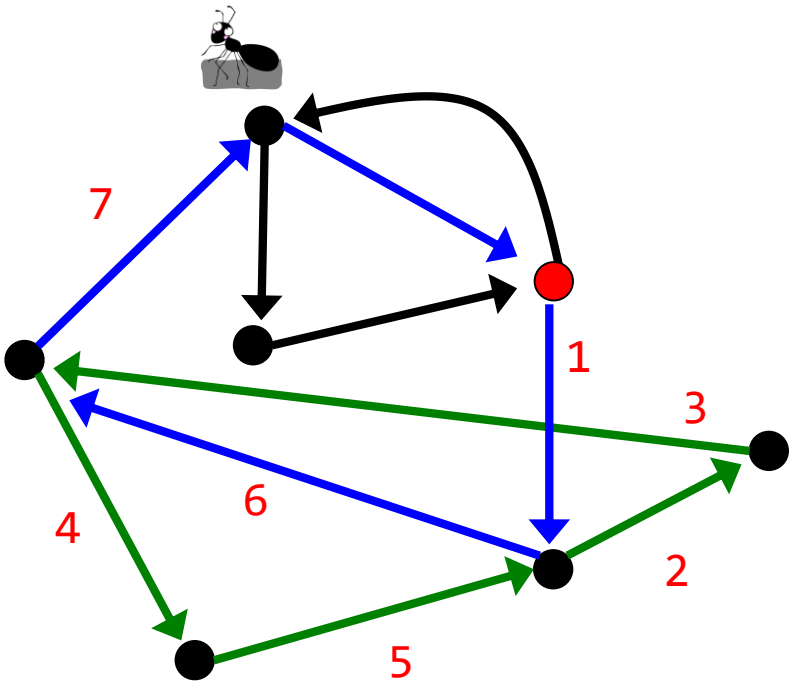
Obilazimo prethodno konstruisan zeleno-plavi ciklus



Obilazimo prethodno konstruisan zeleno-plavi ciklus

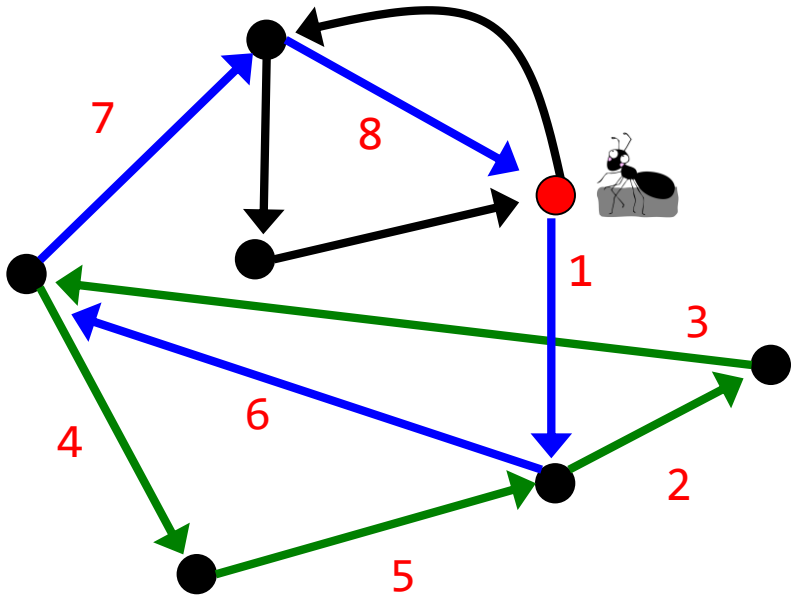


Obilazimo prethodno konstruisan zeleno-plavi ciklus

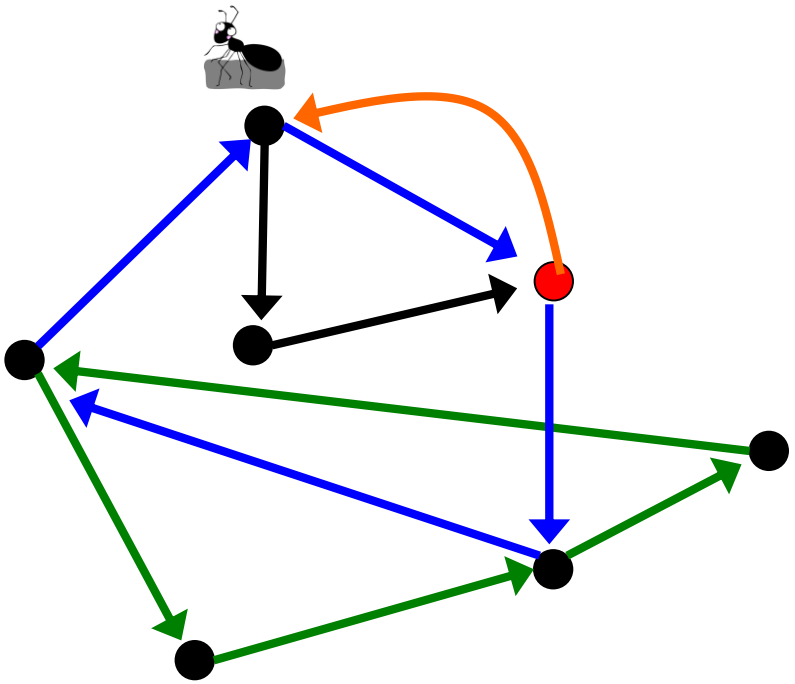


Mrav se vratio nazad ali može da nastavi da obilazi!

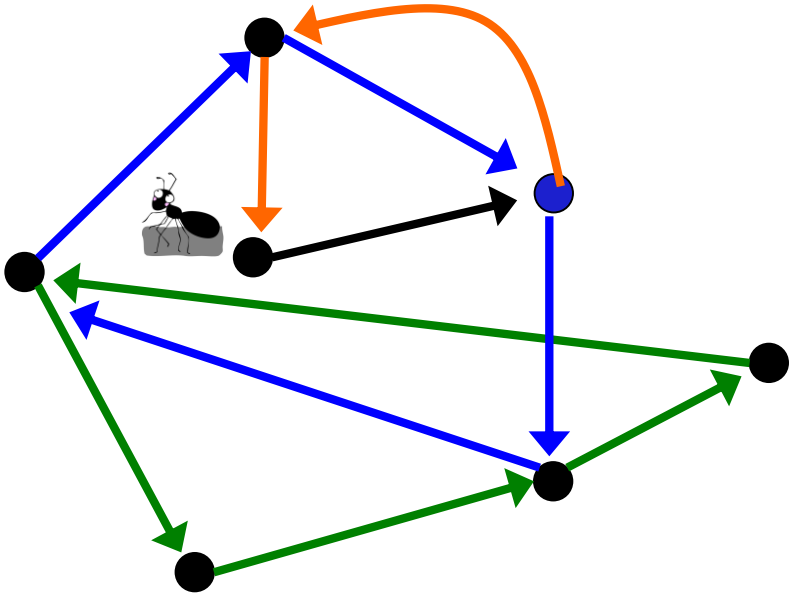
"Hmm, maybe these
instructions were no
that stupid…"



Uvećavamo zeleno-plavi ciklus

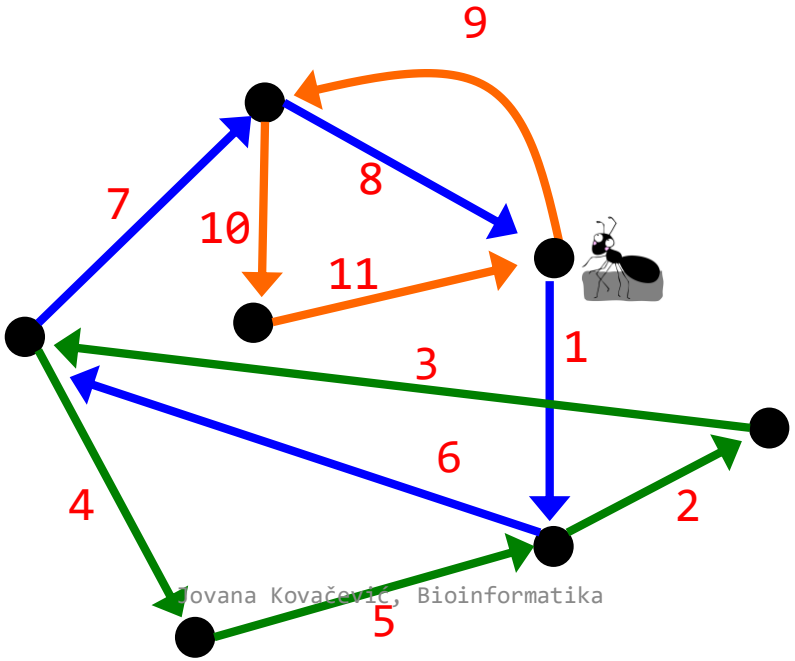


Uvećavamo zeleno-plavi ciklus



Ojlerova teorema je dokazana

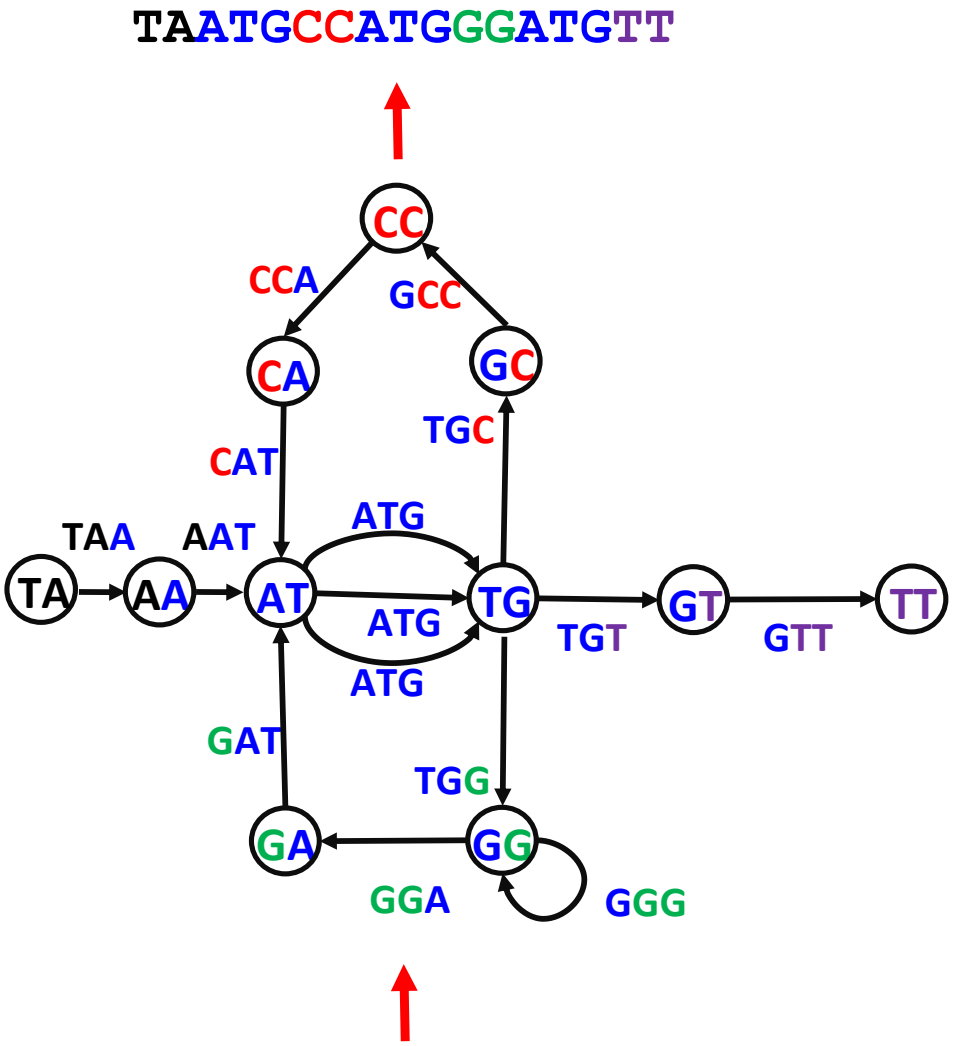
```
EulerianCycle(BalancedGraph)
   form a Cycle by randomly walking in BalancedGraph (avoiding
already visited edges)
      while Cycle is not Eulerian
         select a node newStart  in Cycle with still
unexplored outgoing edges
         form a Cycle' by traversing Cycle from newStart and
randomly walking afterwards
         Cycle ← Cycle'
   return Cycle
```



Pregled

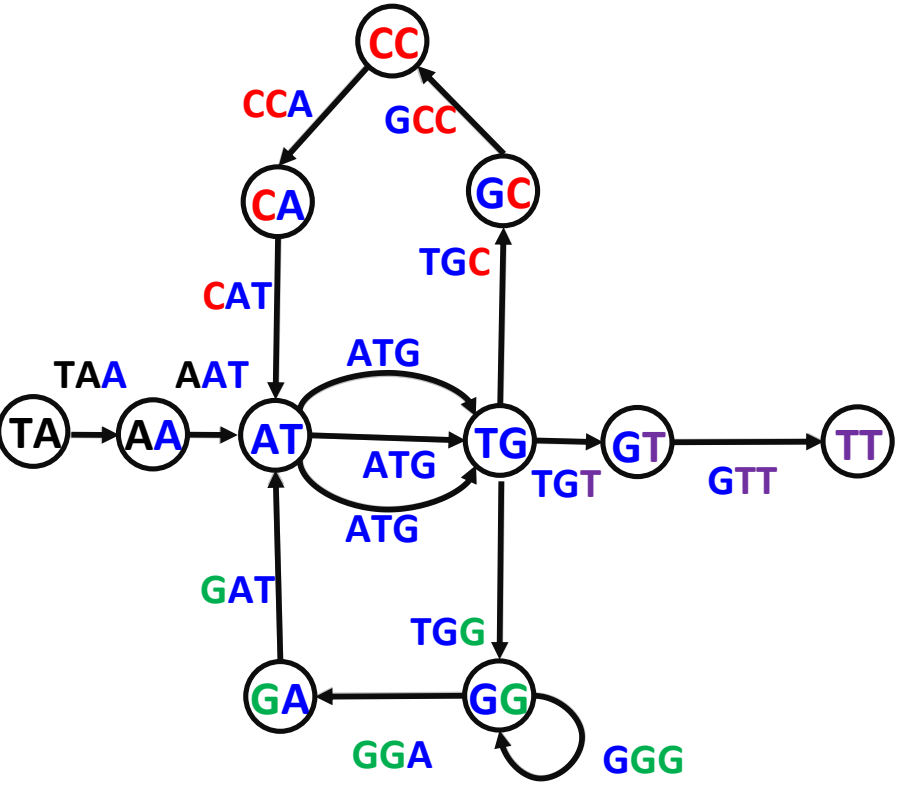
- Šta je sekvencioniranje genoma?
- Eksplozija u štampariji
- Problem rekonstrukcije niske
- Rekonstrukcija niske kao problem Hamiltonove putanje
- Rekonstrukcija niske kao problem Ojlerove putanje
- Slični problemi sa različitim sudbinama?
- De Brujinovi grafovi
- Ojlerova teorema
- **Sastavljanje parova očitavanja**
- De Bruijn Graphs Face Harsh Realities of Assembly

Od očitavanja do De Brujinovog grafa do genoma

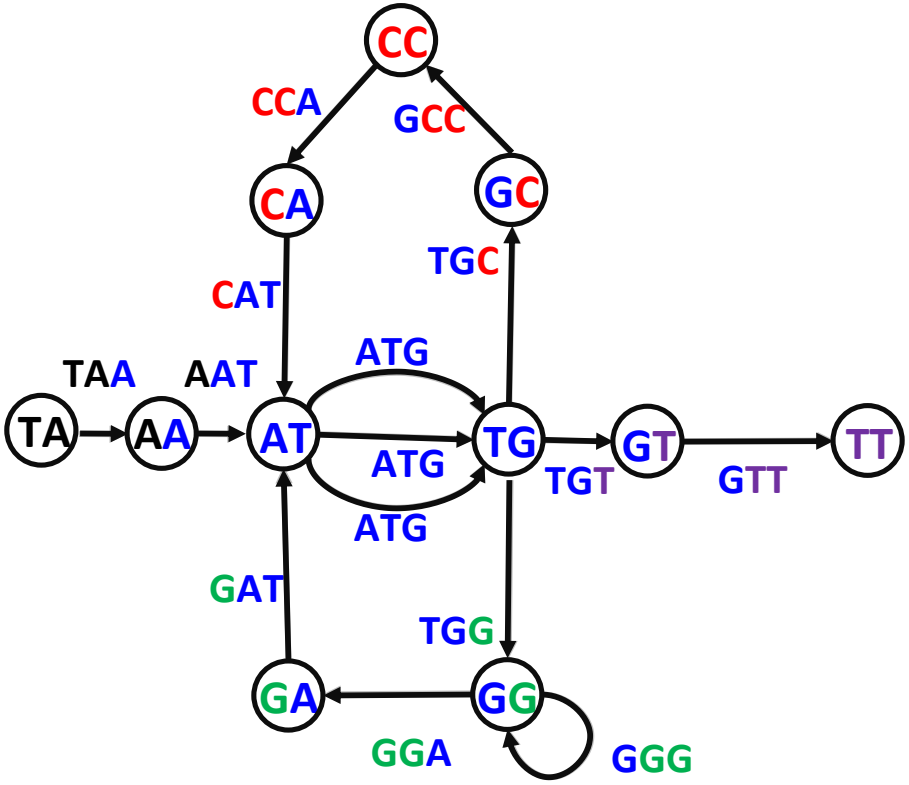


Graf može imati više Ojlerovih putanja

TAATGCCATGGGATGTT



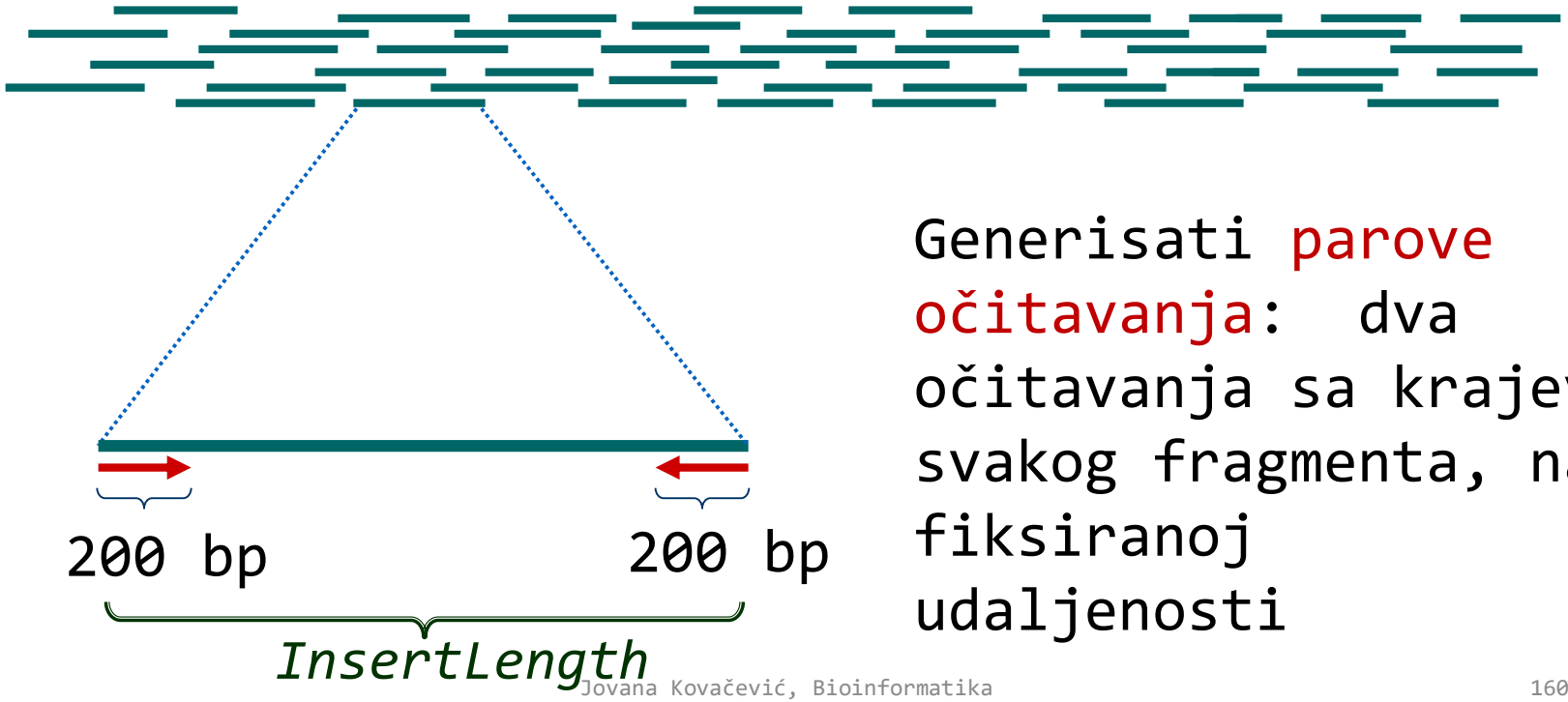
TAATGGGATGCCATGTT



DNK sekvencioniranje sa parovima očitavanja

Više identičnih kopija genoma

Na slučajnim pozicijama iseći genom na fragmente jednake dužine *InsertLength*



Generisati parove očitavanja: dva očitavanja sa krajeva svakog fragmenta, na fiksiranoj udaljenosti

Od *k*-grama do uparenih *k*-grama

Pod uparenim *k*-gramom podrazumevamo par *k*-grama na fiksiranom rastojanju d u genomu. Na primer, TCA i TCC na rastojanju d=11 čine jedan upareni *k*-gram.

Šta je upareni *k*-gramski sastav PairedComposition(**TAATGCCATGGGATGTT**)?

TAA GCC

upareni 3-gram

Šta je upareni *k*-gramski sastav PairedComposition(**TAATGCCATGGGATGTT**)?

TAA GCC
AAT CCA
ATG CAT
TGC ATG
GCC TGG
CCA GGG
CAT GGA
ATG GAT
TGG ATG
GGG TGT
GGA GTT

Predstavimo upareni 3-gram **TAA GCC** na sledeći način:

TAA
GCC

TAA	**AAT**	**ATG**	**TGC**	**GCC**	**CCA**	**CAT**	**ATG**	**TGG**	**GGG**	**GGA**
GCC	**CCA**	**CAT**	**ATG**	**TGG**	**GGG**	**GGA**	**GAT**	**ATG**	**TGT**	**GTT**

PairedComposition(TAATGCCATGGGATGTT)

TAA GCC
AAT CCA
ATG CAT
TGC ATG
GCC TGG
CCA GGG
CAT GGA
ATG GAT
TGG ATG
GGG TGT
GGA GTT

TAA	AAT	ATG	TGC	GCC	CCA	CAT	ATG	TGG	GGG	GGA
GCC	CCA	CAT	ATG	TGG	GGG	GGA	GAT	ATG	TGT	GTT

AAT	ATG	ATG	CAT	CCA	GCC	GGA	GGG	TAA	TGC	TGG
CCA	CAT	GAT	GGA	GGG	TGG	GTT	TGT	GCC	ATG	ATG

Leksikografski poredak kolekcije PairedComposition

Problem rekonstrukcije niske na osnovu parova očitavanja

Problem rekonstrukcije niske na osnovu parova očitavanja.
Rekontruisati nisku na osnovu njenih uparenih *k*-grama.

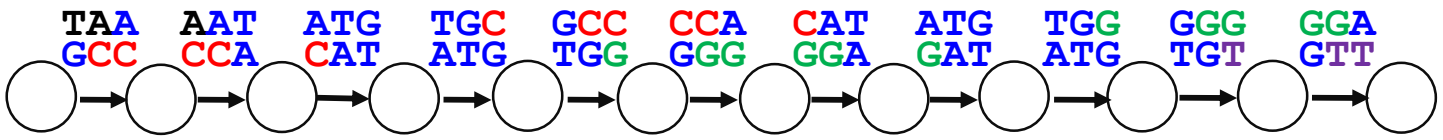
- **Ulaz.** Kolekcija uparenih *k*-grama.
- **Izlaz.** Niska *Text* takva da je *PairedComposition*(*Text*) jednak kolekciji uparenih *k*-grama.

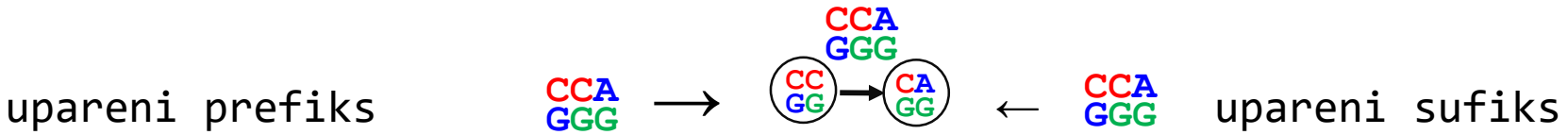
Kako konstruisati **upareni** De Brujinov graf na osnovu uparenog *k*-gramskog sastava?

Pretpostavimo da je **dat genom** (niska *Genome*). Posmatrajmo genom kao putanju u grafu obeleženom na osnovu njegovog uparenog *k*-gramskog sastava

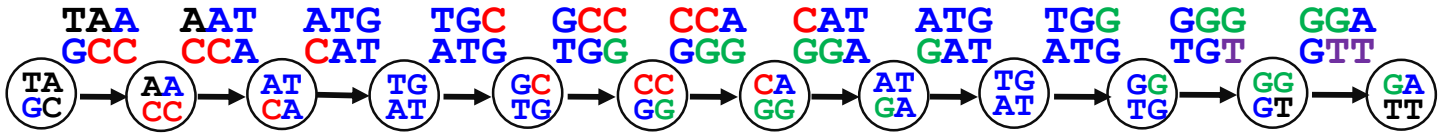
Predstavimo genom TAATGCCATGGGATGTT kao putanju

TAA GCC
AAT CCA
ATG CAT
TGC ATG
GCC TGG
CCA GGG
CAT GGA
ATG GAT
TGG ATG
GGG TGT
GGA GTT

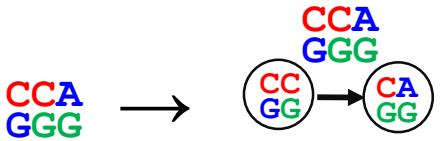




Obeležimo čvorove uparenim prefiksima i sufiksima

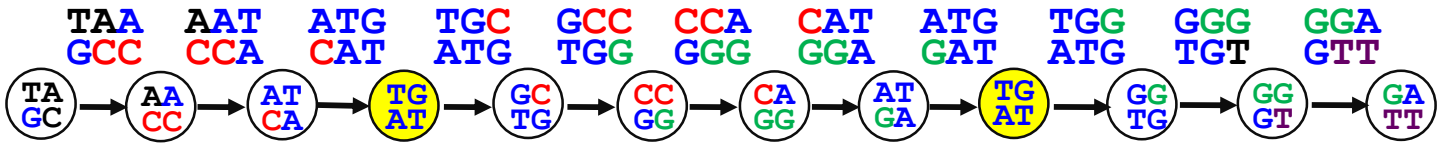


upareni prefiks

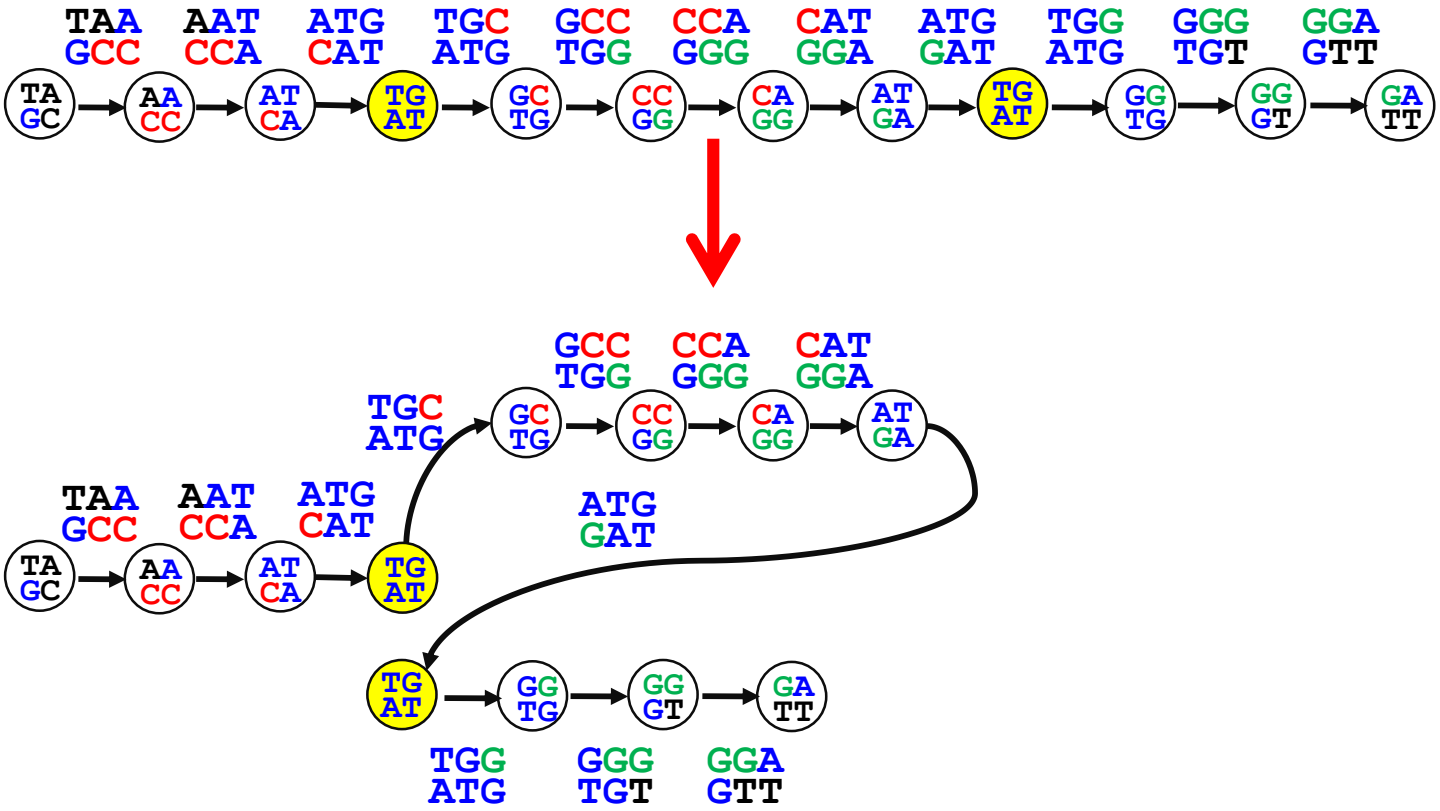


upareni sufiks

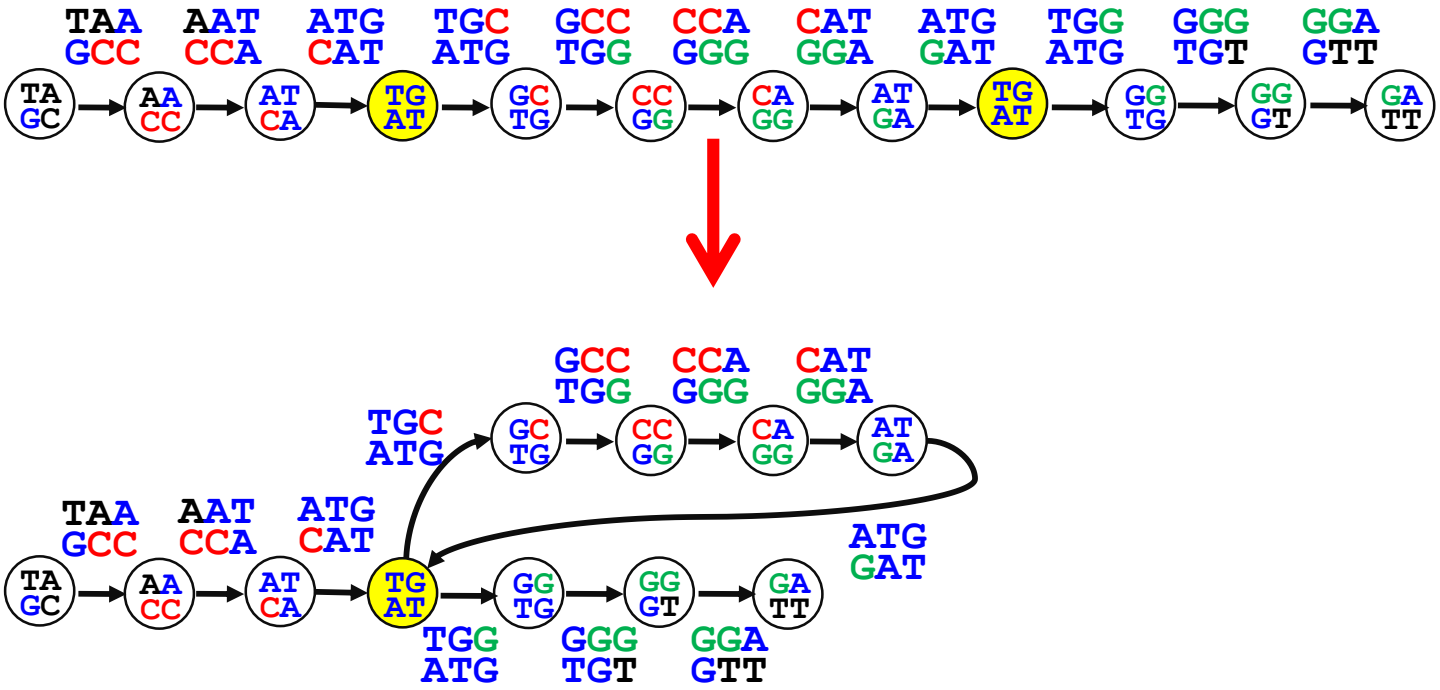
Zalepimo čvorove za identičnim oznakama



Zalepimo čvorove za identičnim oznakama



Zalepimo čvorove za identičnim oznakama

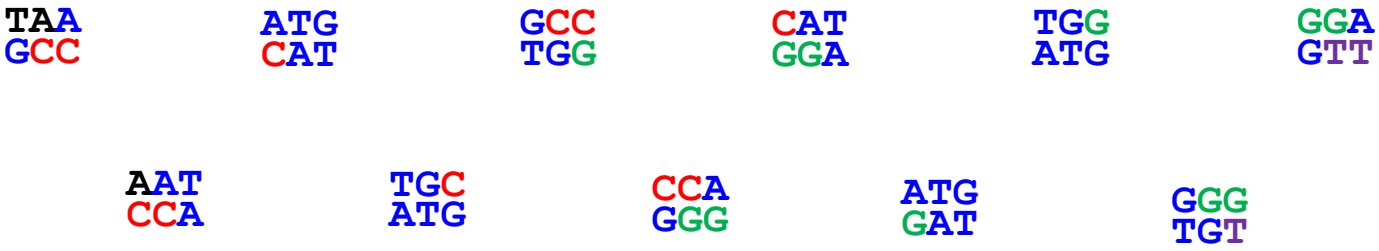


Upareni De Brujinov graf na osnovu datog genoma

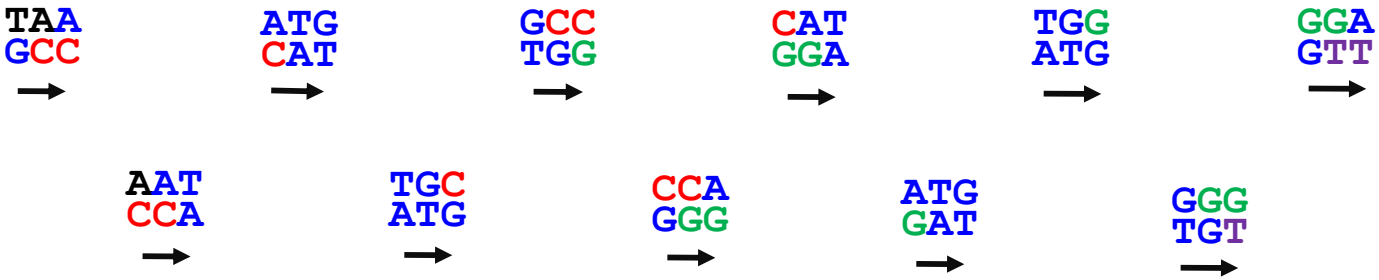
Kako konstruisati de **upareni** de Brujinov graf na osnovu uparenog *k*-gramskog sastava?

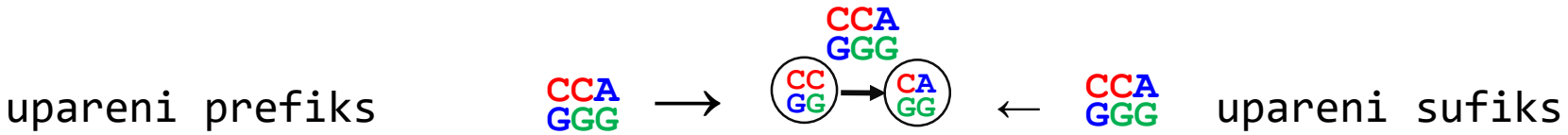
- Pretpostavili smo da je dat genom (niska *Genome*). Posmatrali smo genom kao putanju u grafu obeleženom na osnovu njegovog uparenog *k*-gramskog sastava
- Sada pretpostavimo da **nije dat** genom već samo upareni *k*-gramski sastav

Konstrukcija uparenog De Brujinovog grafa na osnovu uparenih *k*-grama

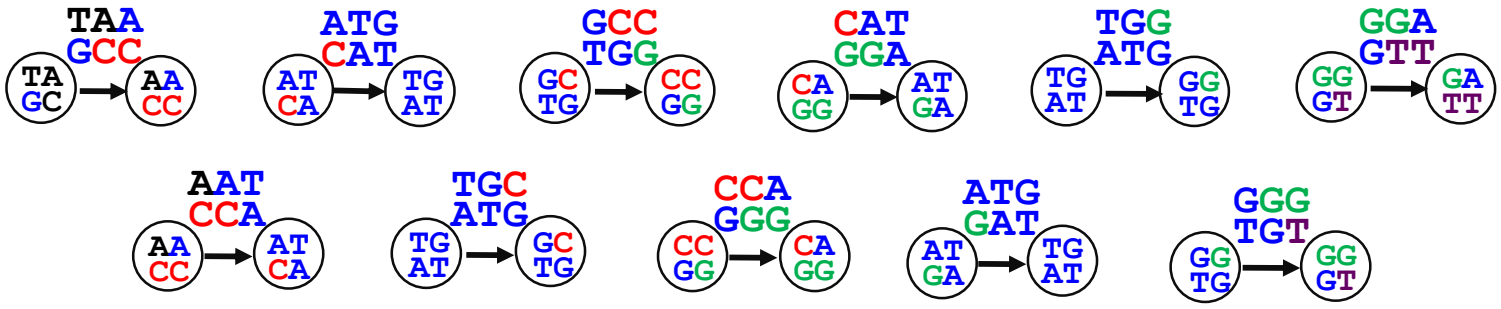


Konstrukcija uparenog De Brujinovog grafa na osnovu uparenih *k*-grama

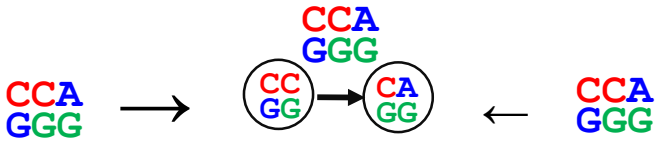




Konstrukcija uparenog De Brujinovog grafa na osnovu uparenih *k*-grama

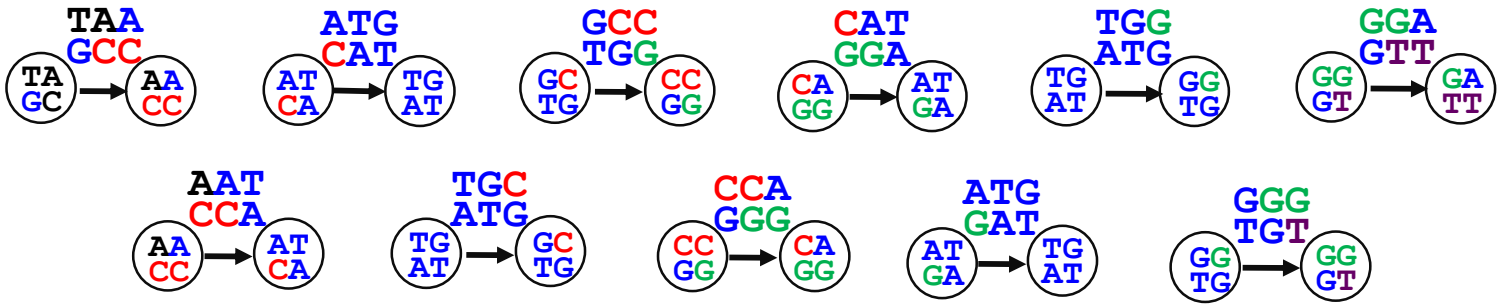


upareni prefiks

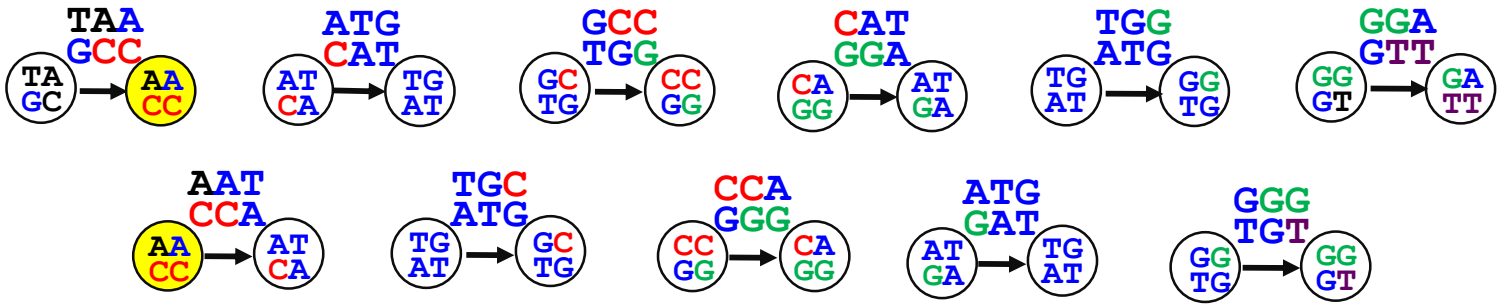


upareni sufiks

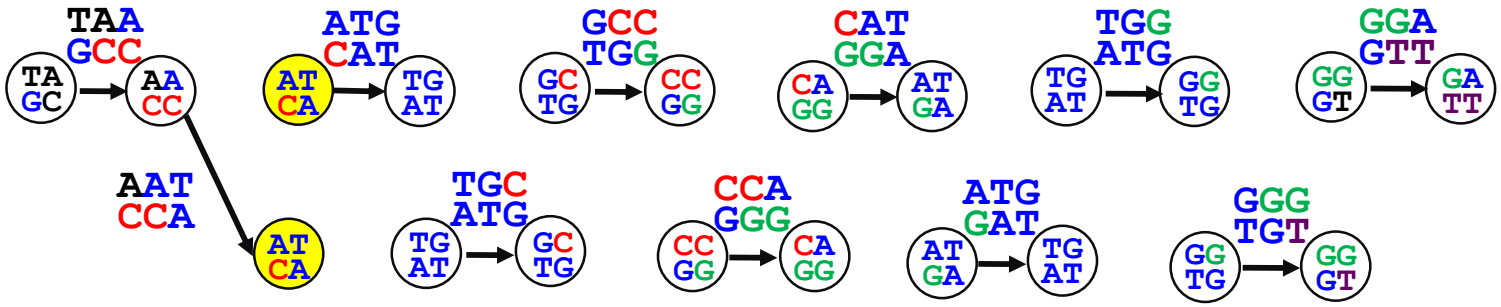
Konstrukcija uparenog De Brujinovog grafa na osnovu uparenih *k*-grama



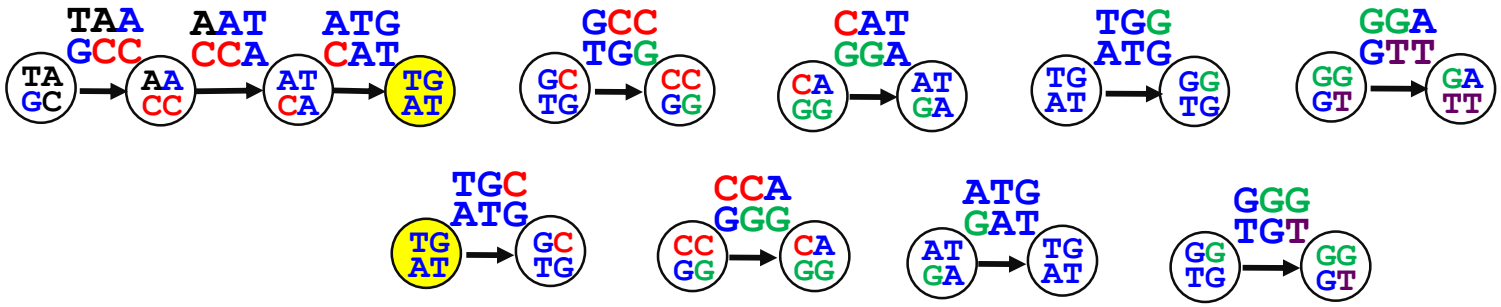
Konstrukcija uparenog De Brujinovog grafa



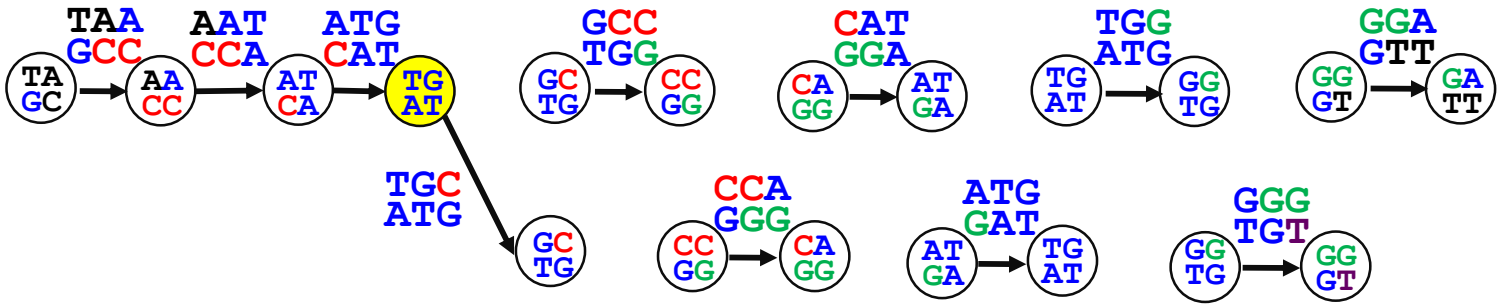
Konstrukcija uparenog De Brujinovog grafa



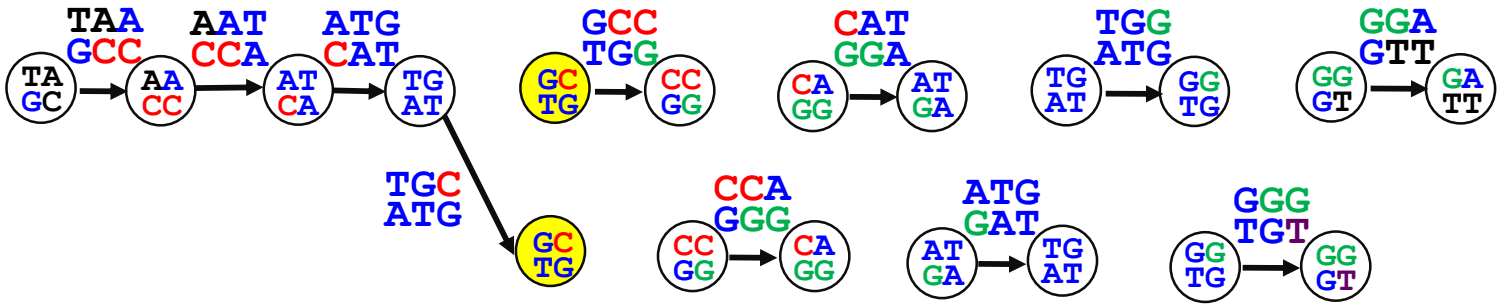
Konstrukcija uparenog De Brujinovog grafa



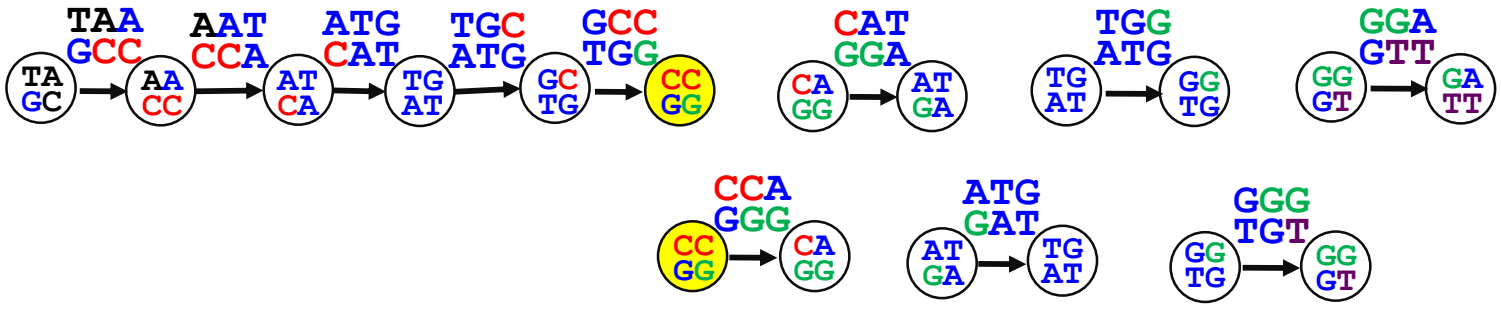
Konstrukcija uparenog De Brujinovog grafa



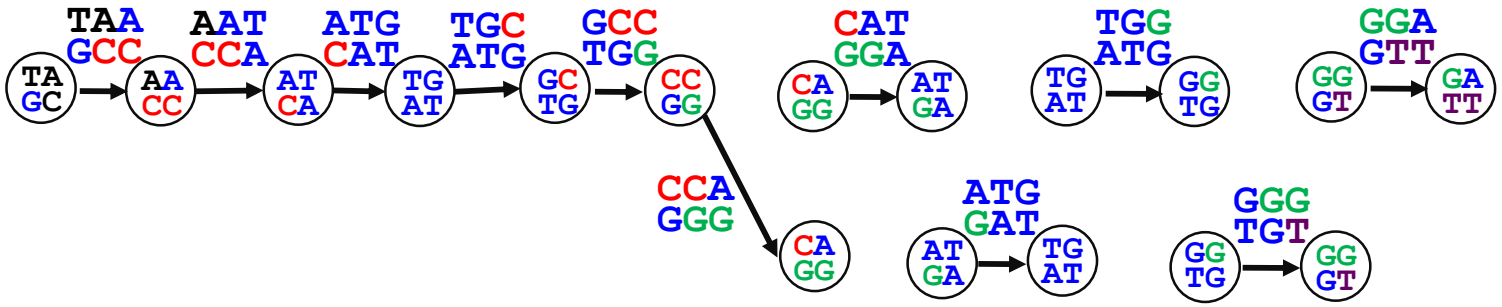
Konstrukcija uparenog De Brujinovog grafa



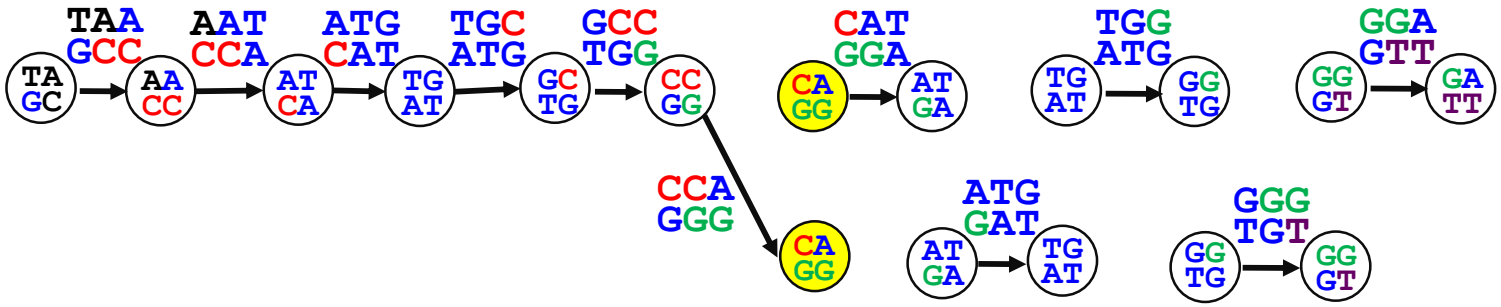
Konstrukcija uparenog De Brujinovog grafa



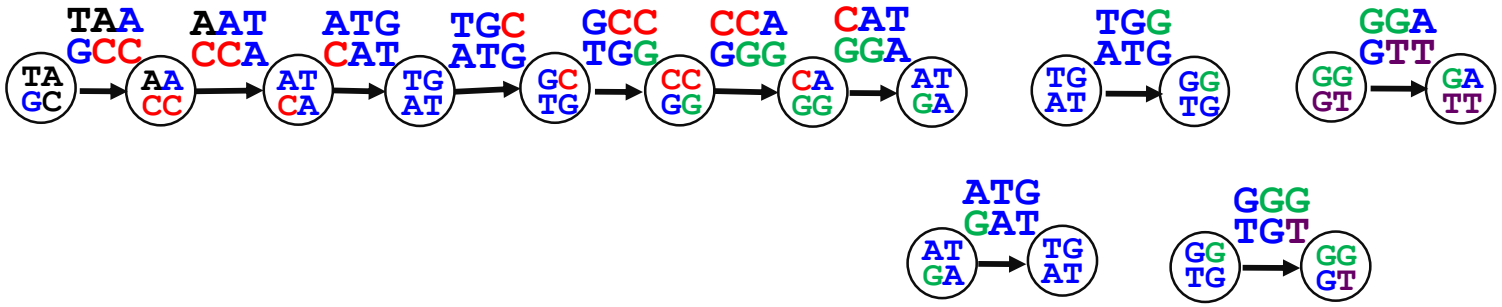
Konstrukcija uparenog De Brujinovog grafa



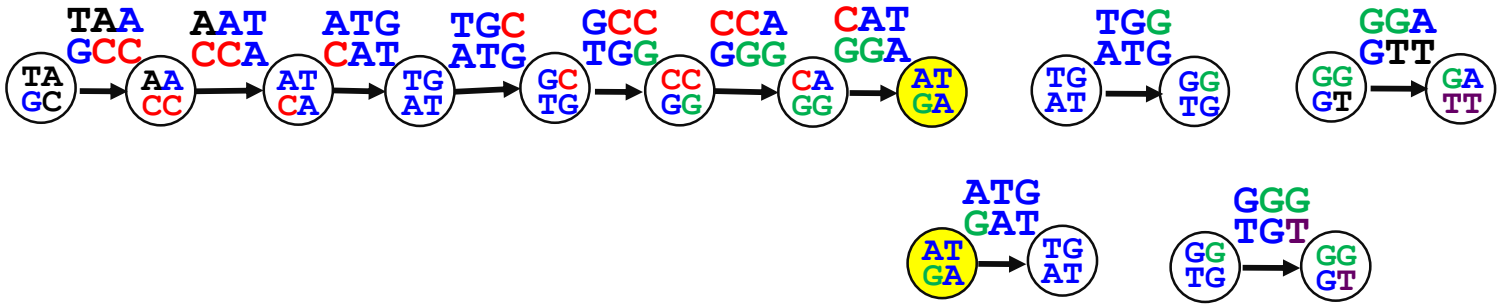
Konstrukcija uparenog De Brujinovog grafa



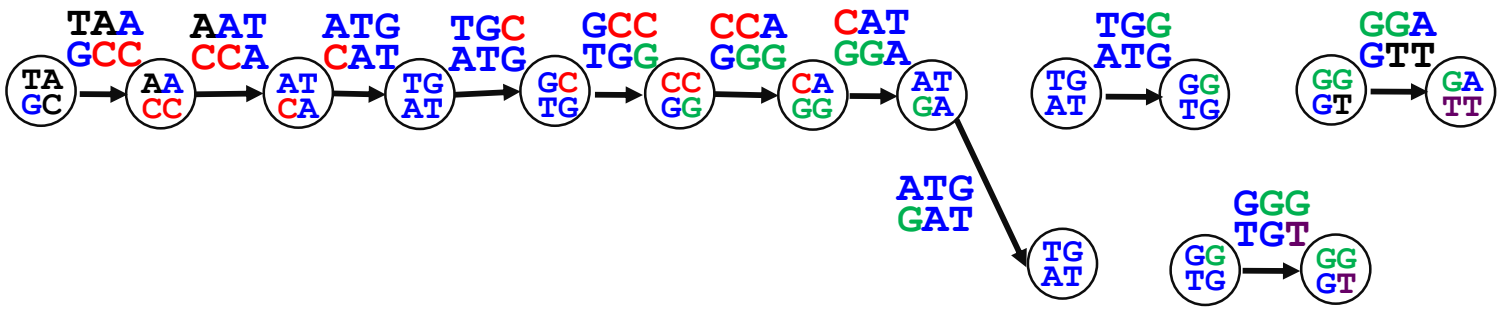
Konstrukcija uparenog De Brujinovog grafa



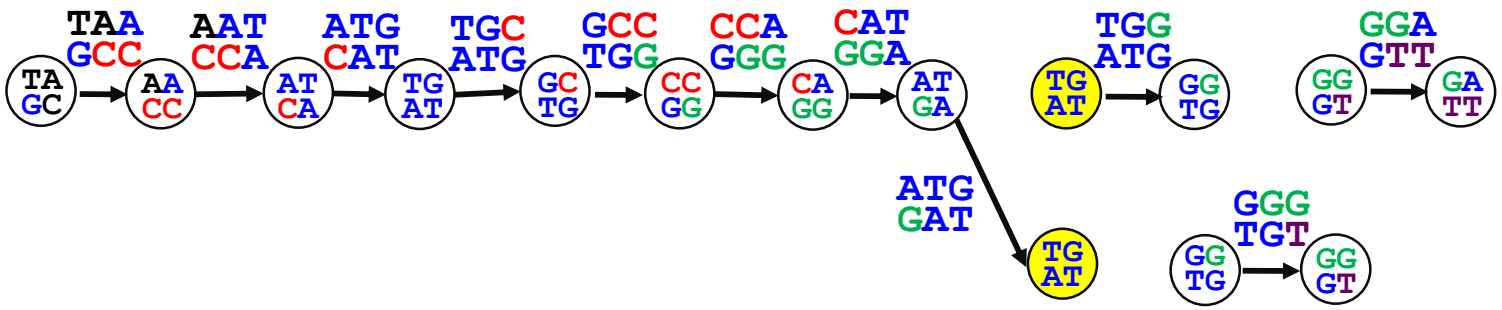
Konstrukcija uparenog De Brujinovog grafa



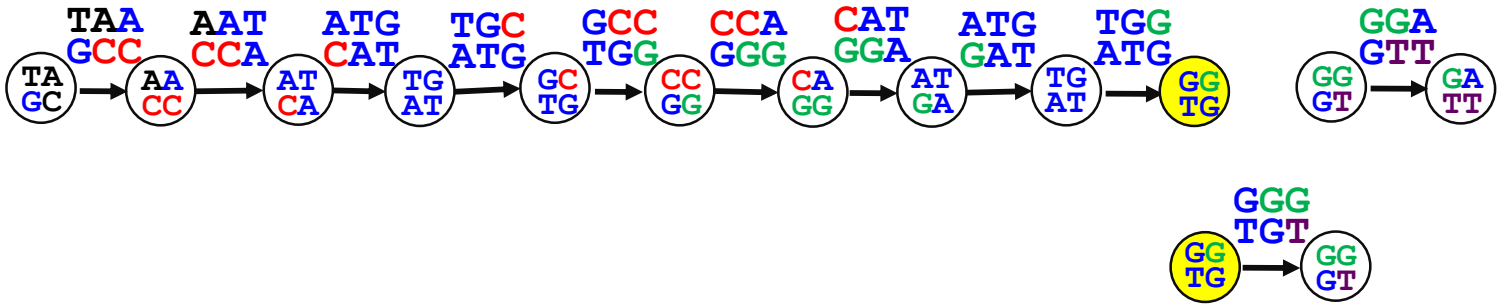
Konstrukcija uparenog De Brujinovog grafa



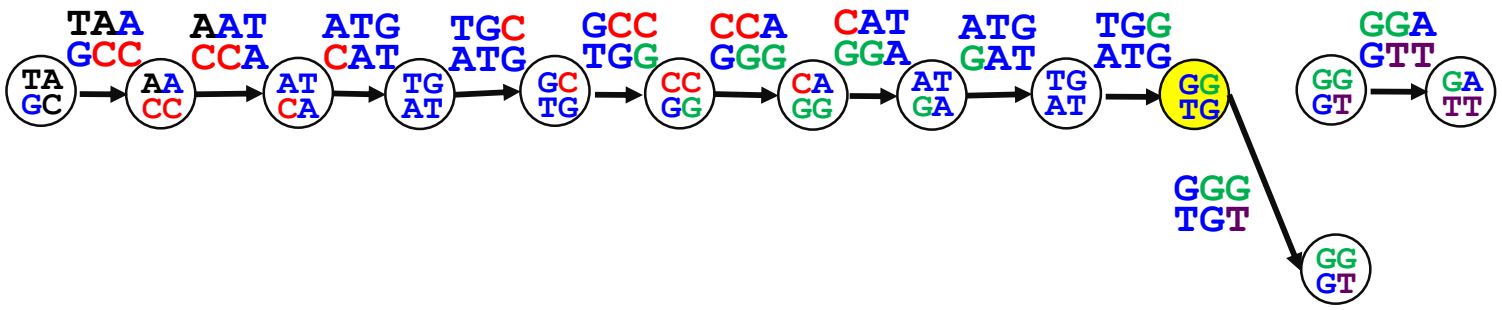
Konstrukcija uparenog De Brujinovog grafa



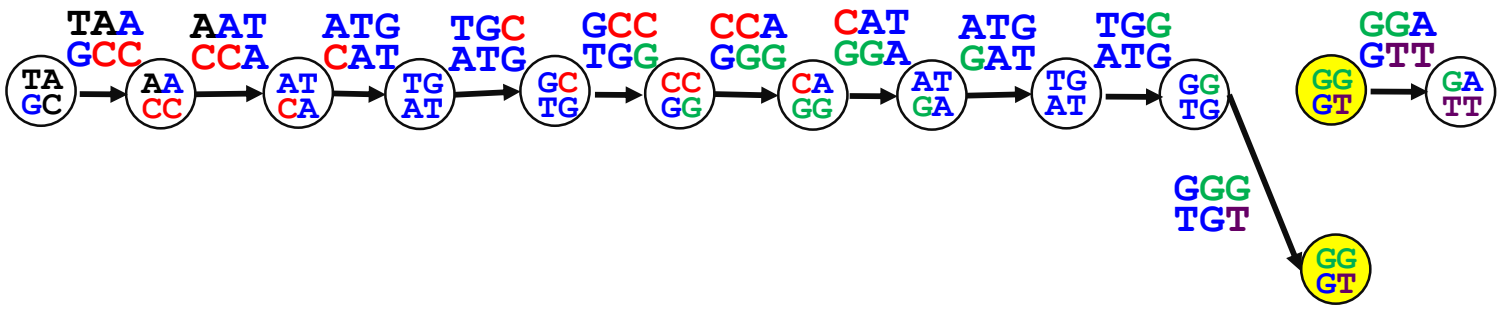
Konstrukcija uparenog De Brujinovog grafa



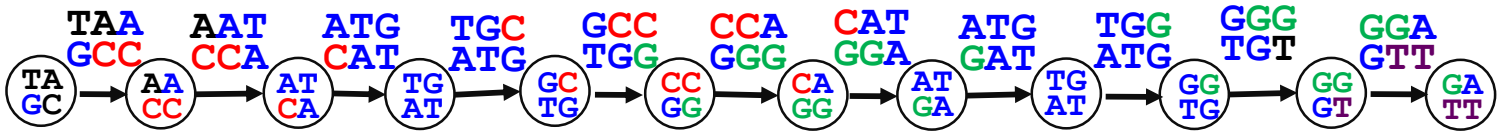
Konstrukcija uparenog De Brujinovog grafa



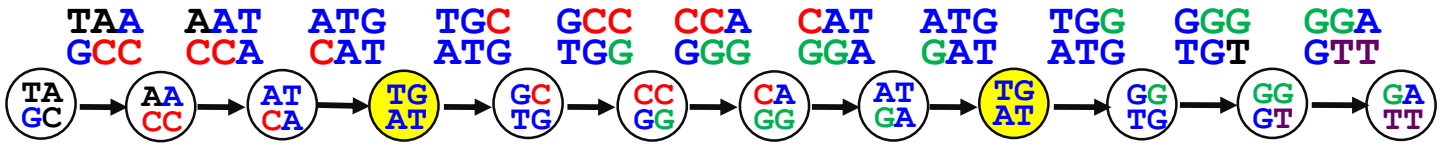
Konstrukcija uparenog De Brujinovog grafa



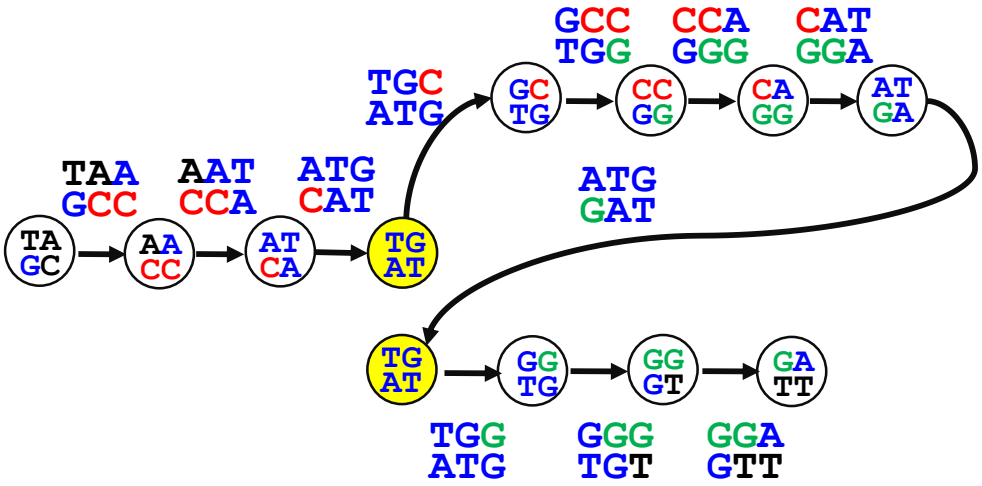
Konstrukcija uparenog De Brujinovog grafa



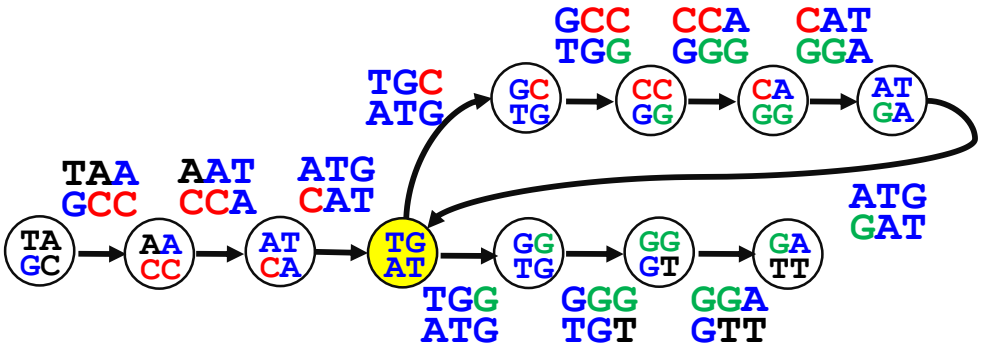
Lepljenje čvorova sa identičnom oznakom



Konstrukcija uparenog De Brujinovog grafa



Konstrukcija uparenog De Brujinovog grafa



Upareni De Brujinov graf na osnovu parova očitavanja

Upareni De Brujinov graf

Upareni De Brujinov graf na osnovu kolekcije uparenih *k*-grama:

- Svaka grana je označena jednim uparenim *k*-gramom
- Svaki čvor je označen prefiksima/sufiksima izlazne/ulazne grane
- Zalepljeni su svi čvorovi sa identičnim oznakama.

Koji graf je bolja reprezentacija?

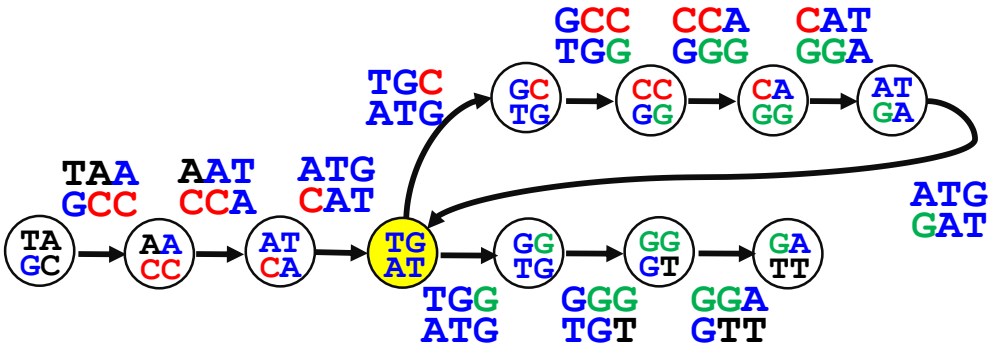
Jedinstvena rekonstrukcija genoma

Višestruka rekonstrukcija genoma

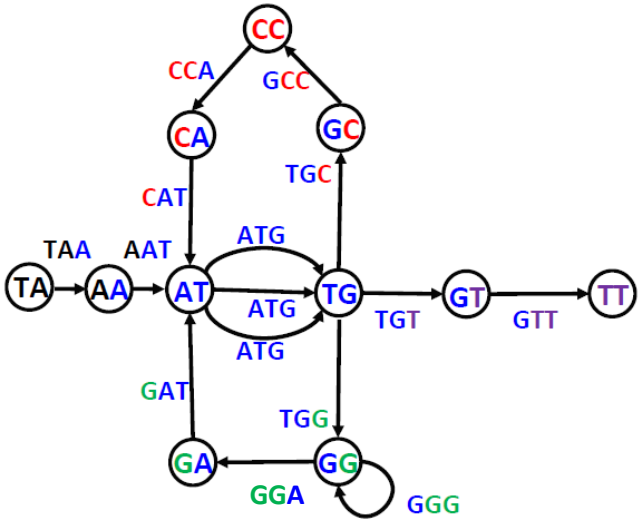
TAATGCCATGGGATGTT

TAATGCCATGGGATGTT

TAATGGGATGCCATGTT



Upareni De Brujinov graf



De Brujinov graf

Pregled

- Šta je sekvencioniranje genoma?
- Eksplozija u štampariji
- Problem rekonstrukcije niske
- Rekonstrukcija niske kao problem Hamiltonove putanje
- Rekonstrukcija niske kao problem Ojlerove putanje
- Slični problemi sa različitim sudbinama?
- De Brujinovi grafovi
- Ojlerova teorema
- Sastavljanje parova očitavanja
- **U realnosti**

Nerealne pretpostavke

- Savršena pokrivenost genoma očitavanjima (svaki k-gram iz genoma je očitan)
- Očitavanja ne sadrže greške
- Rastojanja između očitavanja u okviru parova očitavanja su egzaktna

~~Ne~~realne pretpostavke

- **Nesavršena** pokrivenost genoma očitavanjima (svaki k-gram iz genoma je očitan)
- Očitavanja ~~ne~~ **sadrže** greške
- Rastojanja između očitavanja u okviru parova očitavanja **nisu** egzaktna
- **Itd.**

Prva nerealna pretpostavka: savršena pokrivenost

```
atgccgtatggacaacgact
atgccgtatg
  gccgtatgga
     gtatggacaa
          gacaacgact
```

Očitavanja dužine 250 nukleotida dobijena *Illumina* tehnologijom predstavljaju samo mali deo 250-grama unutar genoma.

Rešenje: razbiti dobijena očitavanja na kraće *k*-grame

```
atgccgtatggacaacgact
atgccgtatg
  gccgtatgga
     gtatggacaa
          gacaacgact
```

```
atgccgtatggacaacgact
atgcc
 tgccg
  gccgt
   ccgta
    cgtat
     gtatg
      tatgg
       atgga
        tggac
         ggaca
          gacaa
           acaac
            caacg
             aacga
              acgac
               cgact
```

Druga nerealna pretpostavka: očitavanja ne sadrže greške

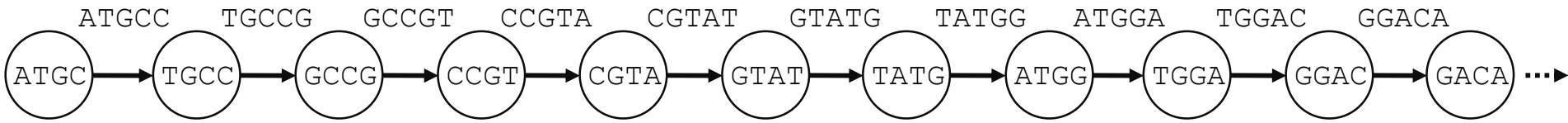
atgccgtatggacaacgact
atgccgtatg
gccgtatgga
gtatggacaa
gacaacgact
cgtaCggaca

Očitavanje sa greškom (promena t u C)

atgccgtatggacaacgact
atgcc
tgccg
gccgt
ccgta
cgtat
gtatg
tatgg
atgga
tggac
ggaca
gacaa
acaac
caacg
aacga
acgac
cgact
cgtaC
gtaCg
taCgg
aCgga
Cggac

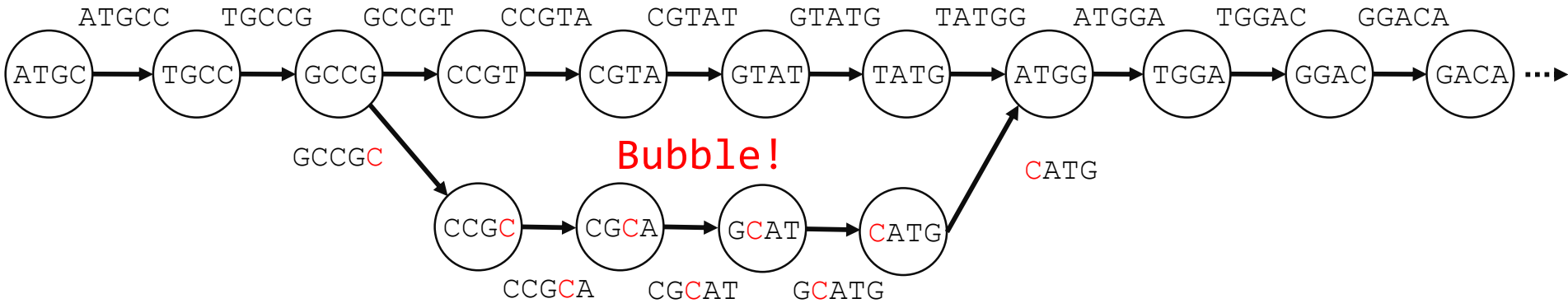
De Brujinov graf genoma ATGGCGTGCAATG… kostruisan na osnovu očitavanja koja ne sadrže greške

.

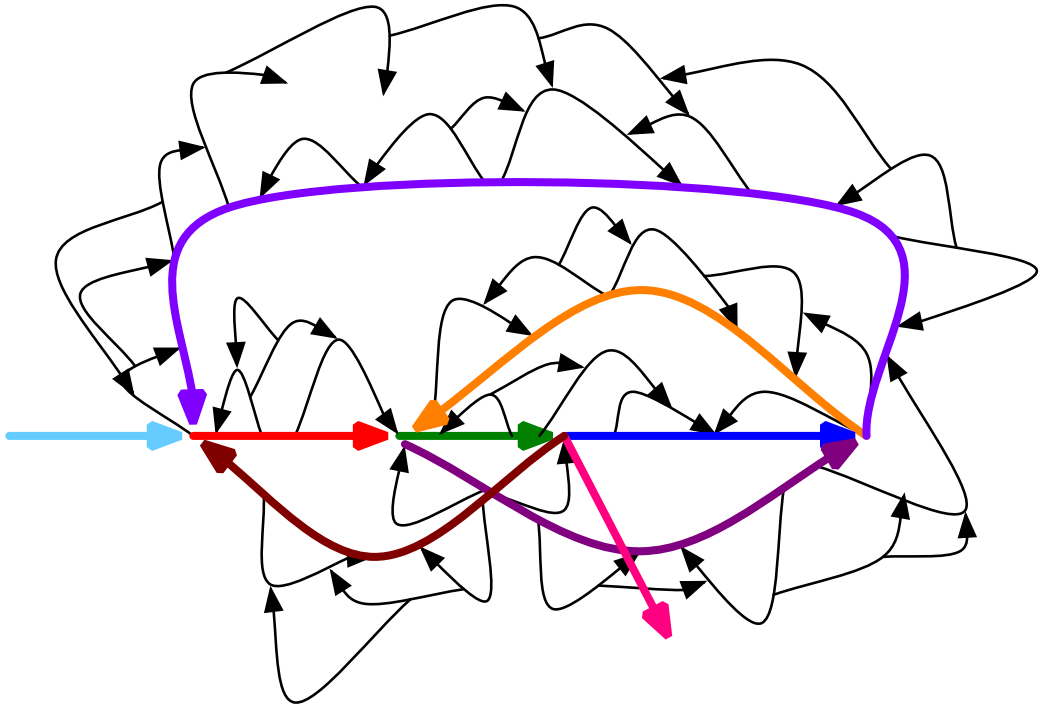


Greške u očitavanjima vode do stvaranja **balončića** u De Brujinovom grafu

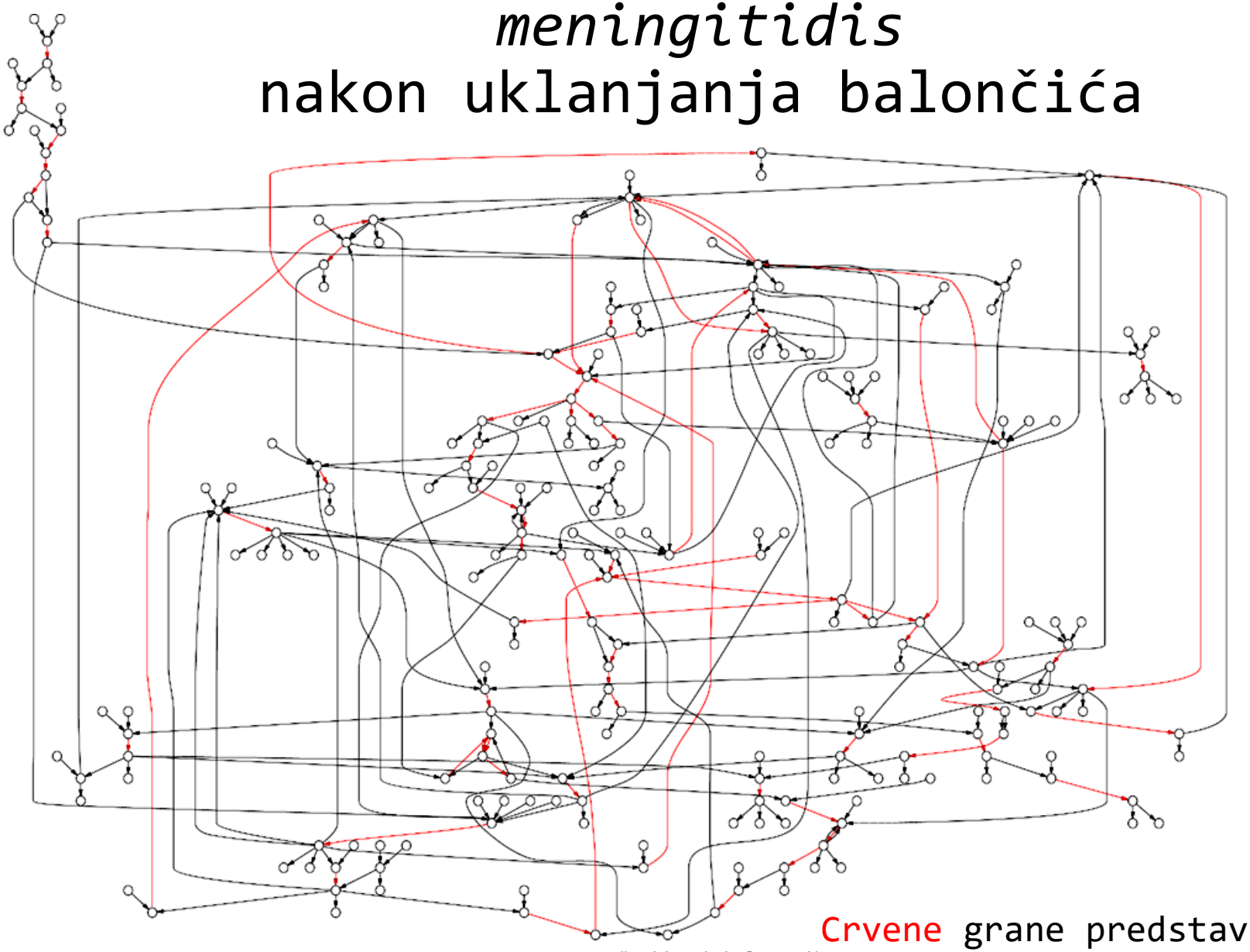
.



Eksplozija balončića



De Brujinov graf genoma *N. meningitidis* nakon uklanjanja balončića



Crvene grane predstavljaju ponavljanja

- Slajdovi pokrivaju poglavlje 3 knjige *Bioinformatics Algorithms: an Active Learning Approach*
- Sadržaj slajdova je preuzet sa zvaničnih prezentacija autora i dodatno prilagođen