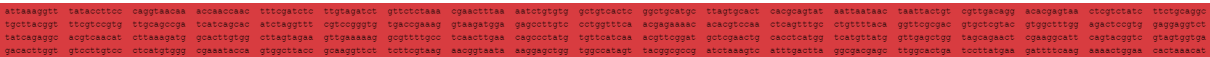
Asembliranje genoma

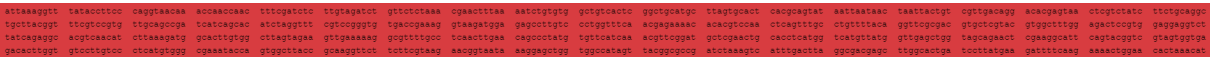
Vladimir Mijić

27. mart 2024.



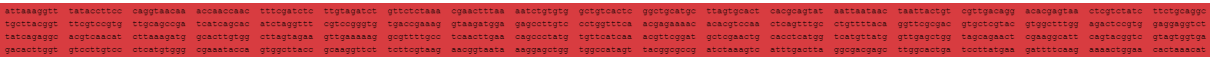
Sekvenciranje genoma - osnovni pojmovi

- Genom - genetski materijal nekog organizma
  - genom se nalazi u DNK ili RNK molekulima



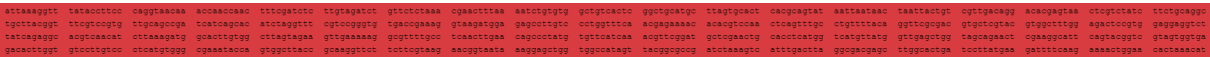
Sekvenciranje genoma - osnovni pojmovi

- Genom - genetski materijal nekog organizma
  - genom se nalazi u DNK ili RNK molekulima
- Nukleotid - osnovni gradivni molekul DNK i RNK
  - adenin, citozin, guanin, timin (DNK) ili uracil (RNK)



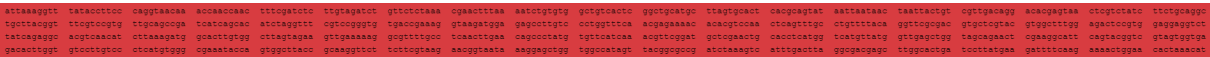
Sekvenciranje genoma - osnovni pojmovi

- Genom - genetski materijal nekog organizma
  - genom se nalazi u DNK ili RNK molekulima
- Nukleotid - osnovni gradivni molekul DNK i RNK
  - adenin, citozin, guanin, timin (DNK) ili uracil (RNK)
- Ljudski genom sadrži oko 3 milijarde nukleotida



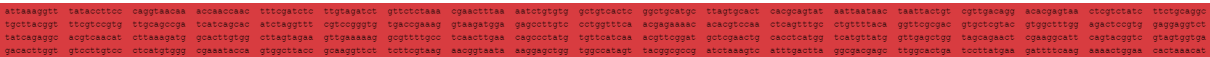
Sekvenciranje genoma - osnovni pojmovi

- Genom - genetski materijal nekog organizma
  - genom se nalazi u DNK ili RNK molekulima
- Nukleotid - osnovni gradivni molekul DNK i RNK
  - adenin, citozin, guanin, timin (DNK) ili uracil (RNK)
- Ljudski genom sadrži oko 3 milijarde nukleotida
- Sekvenciranje genoma predstavlja proces otkrivanja sastava genoma



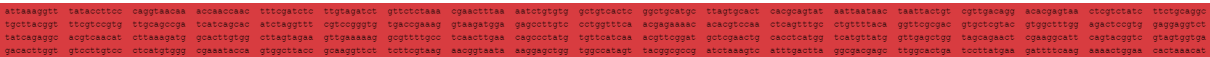
Sekvenciranje genoma - istorijski pregled

- 1977. - Prve metode za sekvenciranja DNK
  - **Sangerova metoda** - Frederick Sanger
  - **Maxam - Gilbert metoda** - Allan Maxam i Walter Gilbert



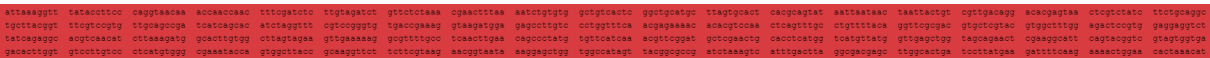
Sekvenciranje genoma - istorijski pregled

- 1977. - Prve metode za sekvenciranja DNK
  - **Sangerova metoda** - Frederick Sanger
  - **Maxam - Gilbert metoda** - Allan Maxam i Walter Gilbert
- Problem?



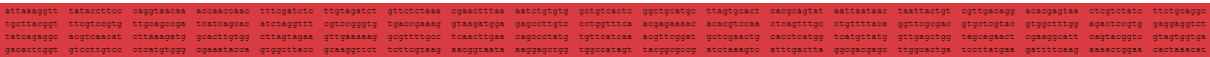
Sekvenciranje genoma - istorijski pregled

- 1977. - Prve metode za sekvenciranja DNK
  - **Sangerova metoda** - Frederick Sanger
  - **Maxam - Gilbert metoda** - Allan Maxam i Walter Gilbert
- Problem?
  - Predložene metode su bile izuzetno skupe - $3 milijarde dolara za sekvenciranje ljudskog genoma

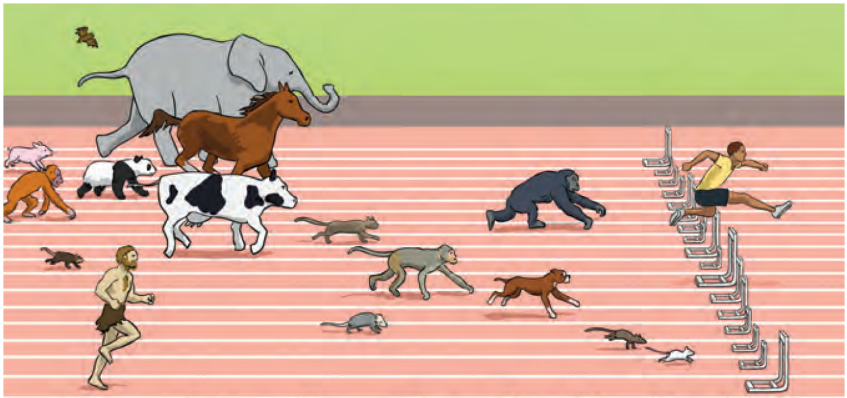


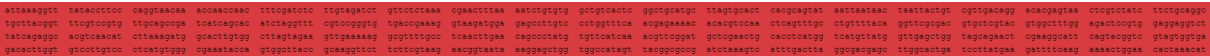
Sekvenciranje genoma - istorijski pregled

- 1977. - Prve metode za sekvenciranja DNK
  - **Sangerova metoda** - Frederick Sanger
  - **Maxam - Gilbert metoda** - Allan Maxam i Walter Gilbert
- Problem?
  - Predložene metode su bile izuzetno skupe - $3 milijarde dolara za sekvenciranje ljudskog genoma
- 1990. - Početak projekta **Human Genome Project**
  - Cilj: sekvenciranje ljudskog genoma
  - Prvi rezultati objavljeni 2001. godine
  - Završen 2003. godine

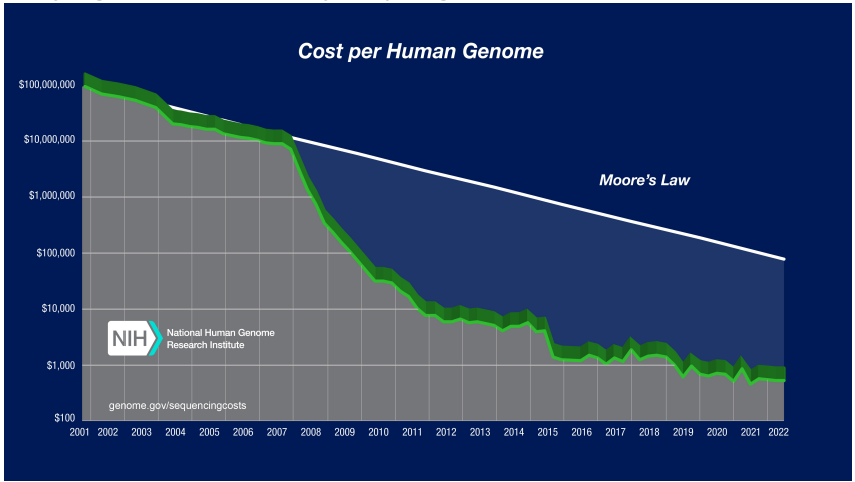


Sekvenciranje genoma - istorijski pregled

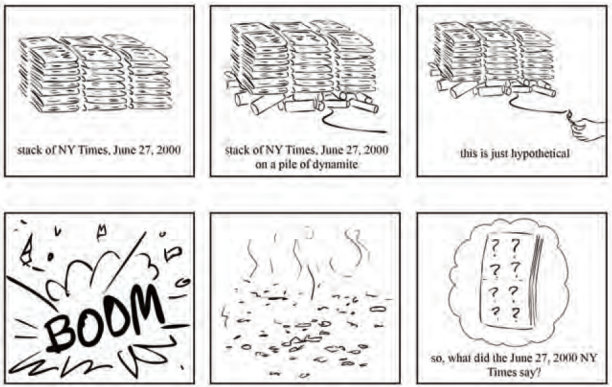




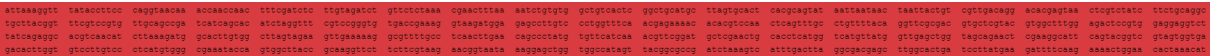
Sekvenciranje genoma - istorijski pregled



Analogija - problem eksplozije novina

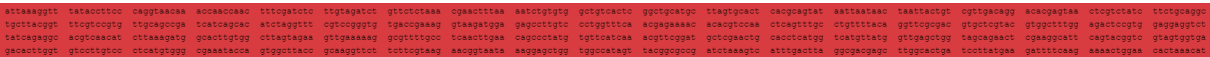


Slika: Problem eksplozije novina će poslužiti u razumijevanju problema asembliranja genoma



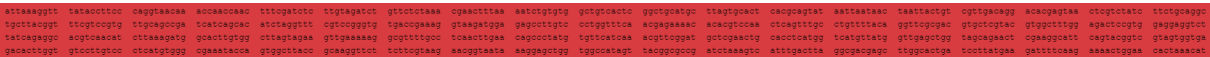
Analogija - problem eksplozije novina

Slika: Spajanje dijelova koji se preklapaju



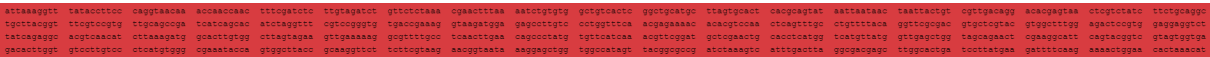
Analogija - problem eksplozije novina

- Kopije novina - odgovaraju uzorcima DNK



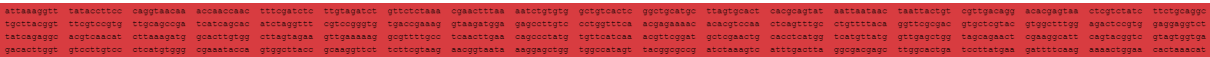
Analogija - problem eksplozije novina

- Kopije novina - odgovaraju uzorcima DNK
- Dijelovi novina - odgovaraju očitanjima sekvenciranja



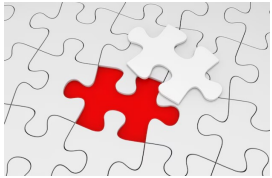
Analogija - problem eksplozije novina

- Kopije novina - odgovaraju uzorcima DNK
- Dijelovi novina - odgovaraju očitanjima sekvenciranja
  - Sekvencer *sječe* genom na manje dijelove koji se nazivaju **očitavanja** (eng. *reads*)
  - Neki dijelovi će biti izgubljeni

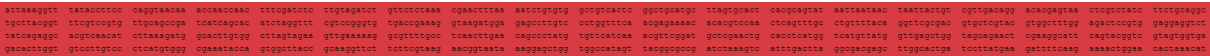


Analogija - problem eksplozije novina

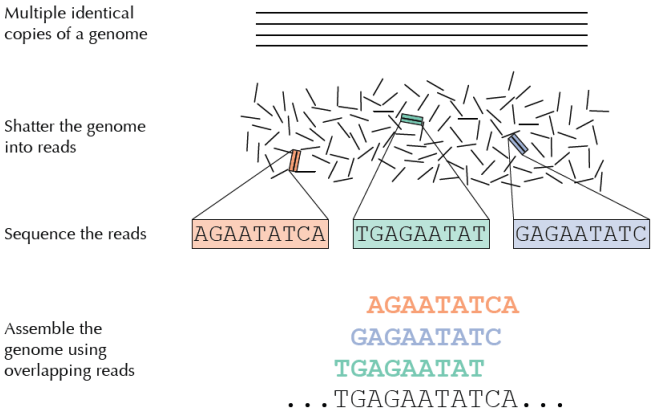
- Kopije novina - odgovaraju uzorcima DNK
- Dijelovi novina - odgovaraju očitanjima sekvenciranja
  - Sekvencer *sječe* genom na manje dijelove koji se nazivaju **očitavanja** (eng. *reads*)
  - Neki dijelovi će biti izgubljeni
- Cilj - spojiti dijelove u jednu cjelinu



Slika: Nije isto kao slaganje slagalice!



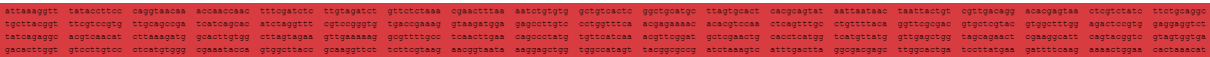
Ilustracija problema



Problem rekonstrukcije niske

Asembliranje genoma je težak problem.
Prije uvođenja formulisanja računarskog problema, navešćemo neke ključne karakteristike problema:

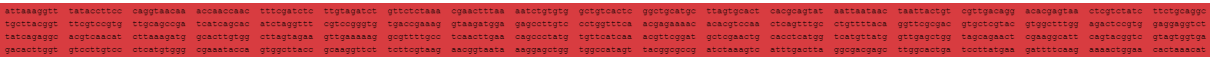
- DNK je dvostruki lanac nukleotida - nije poznato iz kog lanca proizilazi koje očitanje



Problem rekonstrukcije niske

Asembliranje genoma je težak problem.
Prije uvođenja formulisanja računarskog problema, navešćemo neke ključne karakteristike problema:

- DNK je dvostruki lanac nukleotida - nije poznato iz kog lanca proizilazi koje očitanje
- Očitanja mogu sadržati greške

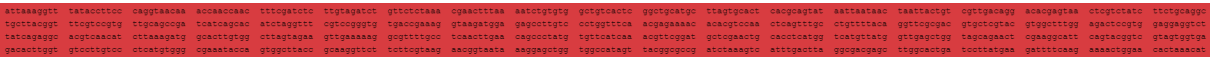


Problem rekonstrukcije niske

Asembliranje genoma je težak problem.

Prije uvođenja formulisanja računarskog problema, navešćemo neke ključne karakteristike problema:

- DNK je dvostruki lanac nukleotida - nije poznato iz kog lanca proizilazi koje očitanje
- Očitanja mogu sadržati greške
- Očitanja ne pokrivaju cijeli genom



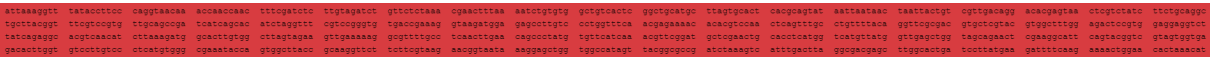
Problem rekonstrukcije niske

Asembliranje genoma je težak problem.

Prije uvođenja formulisanja računarskog problema, navešćemo neke ključne karakteristike problema:

- DNK je dvostruki lanac nukleotida - nije poznato iz kog lanca proizilazi koje očitanje
- Očitanja mogu sadržati greške
- Očitanja ne pokrivaju cijeli genom

Očitavanja su iste dužine, možemo pretpostaviti da su očitavanja svi k-grami (eng. k-mer, podniska dužine k) za neko k.

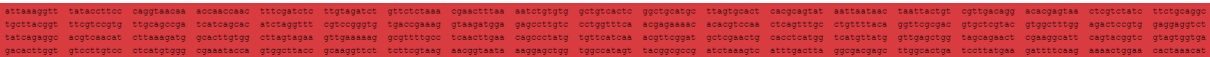


Problem rekonstrukcije niske

Računarska formulacija problema asembliranja genoma:
Za dati string *Text*, njegova k-gram kompozicija $COMPOSITION_3\ (Text)$ je kolekcija svih k-gram podniski od *Text*. Primjer:

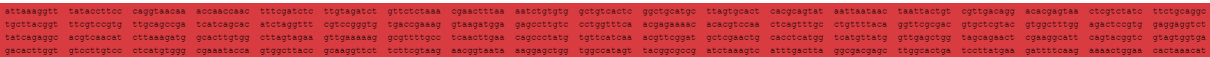
$$COMPOSITION_3\ (TATGGGGTGC) = \{ATG, GGG, GGG, GGT, GTG, TAT, TGC, TGG\}$$

k-grami su prikazani u leksikografskom poretku, zato što tačan redoslijed očitavanja nije poznat kada se generišu.



Problem rekonstrukcije niske

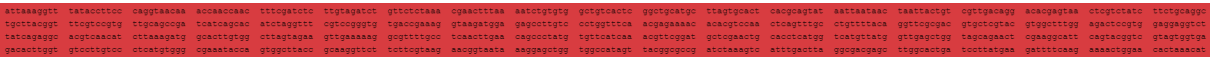
Rekonstruisati nisku iz njenih k-grama.
Ulaz: Vrijednost *k* i kolekcija k-grama *Patterns*.
Izlaz: Niska *Text* koja sadrži sve k-grame iz *Patterns*, ako takva niska postoji.



Problem rekonstrukcije niske - primjer 1.

Neka je zadata 3-gram kompozicija

AAT ATG GTT TAA TGT

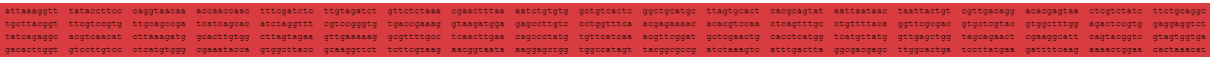


Problem rekonstrukcije niske - primjer 1.

Neka je zadata 3-gram kompozicija

AAT ATG GTT TAA TGT

Zadatak: povežimo parove k-grama ako se preklapaju u k-1 simbola.

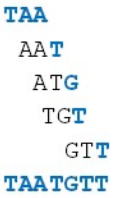


Problem rekonstrukcije niske - primjer 1.

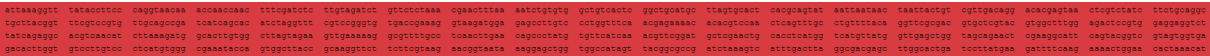
Neka je zadata 3-gram kompozicija

AAT ATG GTT TAA TGT

Zadatak: povežimo parove k-grama ako se preklapaju u k-1 simbola.



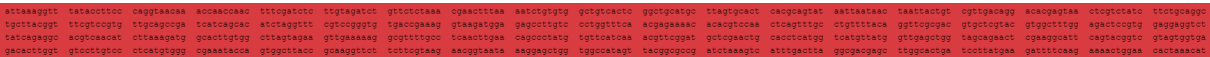
Slika: Rekonstruisana niska je *TAATGTT*



Problem rekonstrukcije niske - primjer 2.

Neka je zadata 3-gram kompozicija

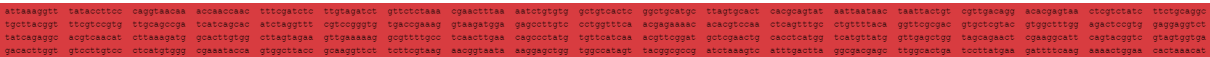
AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT



Problem rekonstrukcije niske - primjer 2.

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

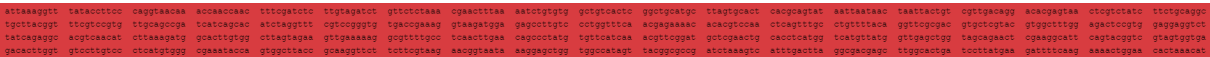
Slika: ATG može biti povezan sa TGC, ili TGG, ili TGT



Problem rekonstrukcije niske - primjer 2.

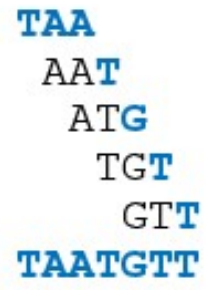
AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

Slika: Kako smo izabrali TGT, jedini sljedeći izbor je GTT

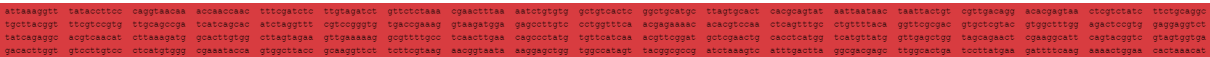


Problem rekonstrukcije niske - primjer 2.

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



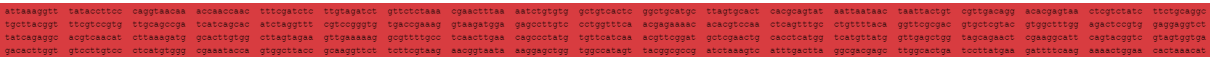
Slika: Ne možemo nastaviti dalje jer nema k-grama koji počinju sa TT



Problem rekonstrukcije niske - primjer 2.

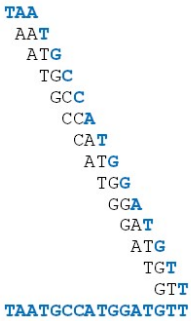
AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

Slika: Vraćamo se na k-gram ATG i nastavljamo dalje sa TGC

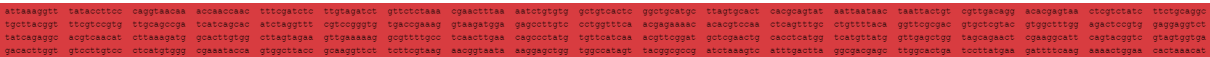


Problem rekonstrukcije niske - primjer 2.

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

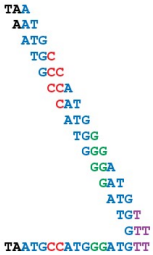


Slika: Nastavljajući dalje, dobili smo *TAATGCCATGGATGTT* ali nismo iskoristili k-gram GGG

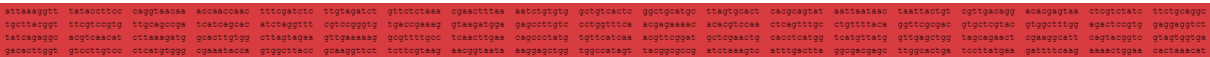


Problem rekonstrukcije niske - primjer 2.

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

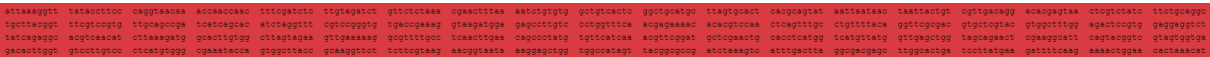


Slika: Ovo je jedno od mogućih rješenja, korištene su različite boje za interval između pojavljivanja ATG



Problem rekonstrukcije niske - grafovska reprezentacija

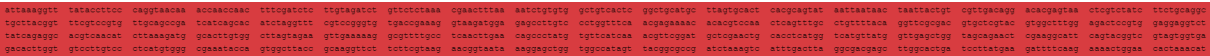
- Čvorovi grafa su k-grami



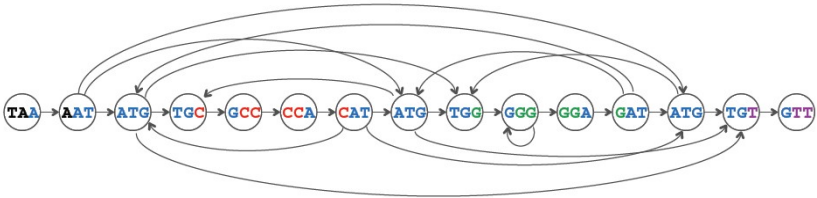
Problem rekonstrukcije niske - grafovska reprezentacija

- Čvorovi grafa su k-grami
- Grane su usmjerene, od čvora *a* do čvora *b* postoji grana ako:
  - *suffix(**a**) == prefix(**b**)*

Slika: Genomska putanja TAATGCCATGGGATGTT

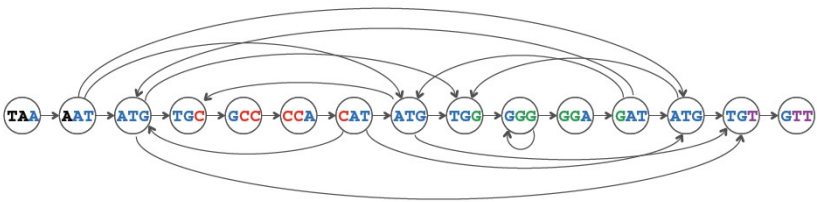


Problem rekonstrukcije niske - grafovska reprezentacija



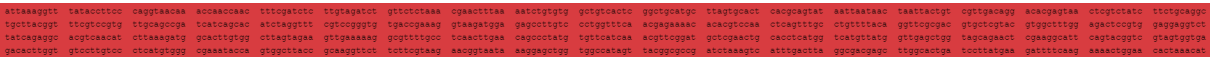
Slika: Graf na osnovu k-gramskog sastava

Problem rekonstrukcije niske - grafovska reprezentacija

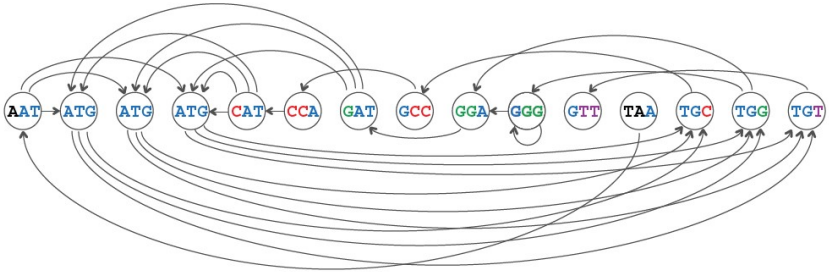


Slika: Graf na osnovu k-gramskog sastava

Možemo li pronaći genomsku putanju u ovom grafu?

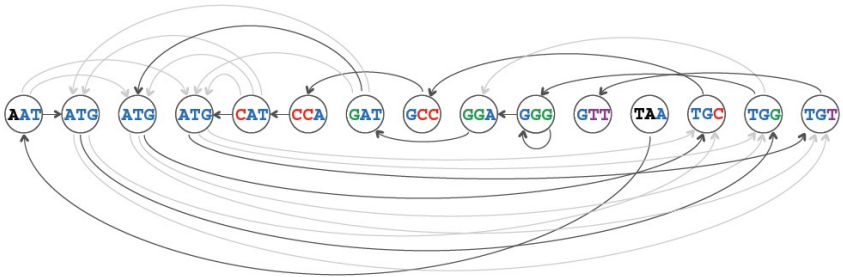


Problem rekonstrukcije niske - grafovska reprezentacija



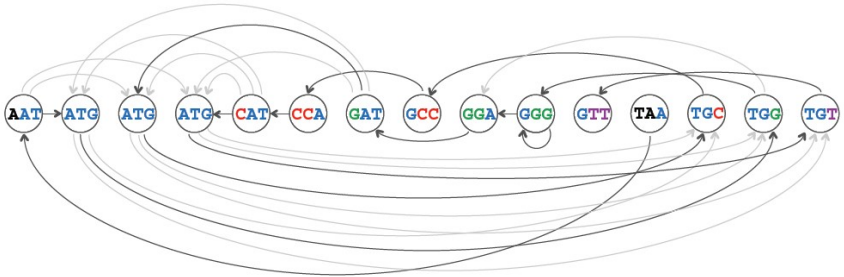
Slika: Graf na osnovu k-gramskog sastava, sortiran po leksikografskom poretku

Problem rekonstrukcije niske - grafovska reprezentacija



Slika: Genomska putanja TAATGCCATGGGATGTT

Problem rekonstrukcije niske - grafovska reprezentacija

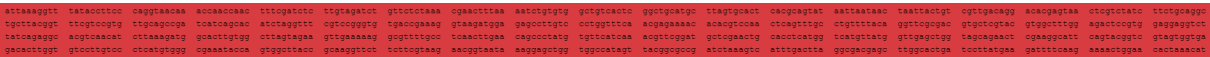


Slika: Genomska putanja TAATGCCATGGGATGTT

Da li je ovo jedino rješenje?

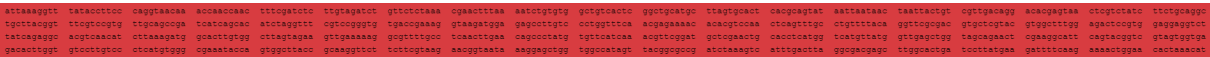
Problem rekonstrukcije niske - problem Hamiltonove i Ojlerove putanje

Hamiltonova putanja



Problem rekonstrukcije niske - problem Hamiltonove i Ojlerove putanje

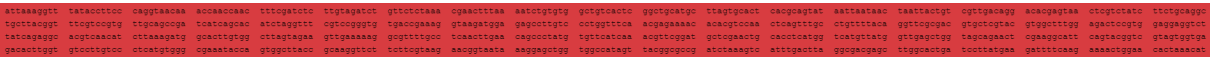
Hamiltonova putanja - putanja koja prolazi kroz sve **čvorove** grafa tačno jednom.



Problem rekonstrukcije niske - problem Hamiltonove i Ojlerove putanje

Hamiltonova putanja - putanja koja prolazi kroz sve **čvorove** grafa tačno jednom.

- Problem pronalaska Hamiltonove putanje je NP-težak - nema efikasnog algoritma za rješavanje

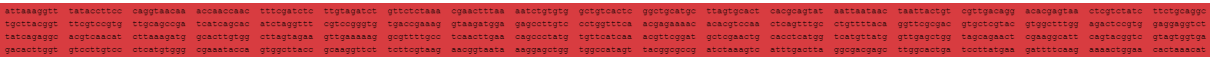


Problem rekonstrukcije niske - problem Hamiltonove i Ojlerove putanje

Hamiltonova putanja - putanja koja prolazi kroz sve **čvorove** grafa tačno jednom.

- Problem pronalaska Hamiltonove putanje je NP-težak - nema efikasnog algoritma za rješavanje

Ojlerova putanja

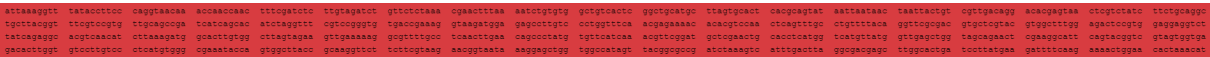


Problem rekonstrukcije niske - problem Hamiltonove i Ojlerove putanje

Hamiltonova putanja - putanja koja prolazi kroz sve **čvorove** grafa tačno jednom.

- Problem pronalaska Hamiltonove putanje je NP-težak - nema efikasnog algoritma za rješavanje

Ojlerova putanja - putanja koja prolazi kroz sve **grane** grafa tačno jednom.



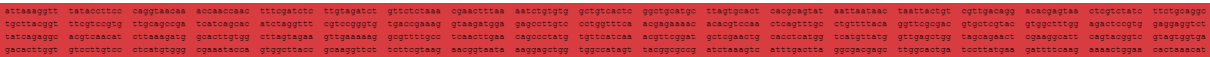
Problem rekonstrukcije niske - problem Hamiltonove i Ojlerove putanje

Hamiltonova putanja - putanja koja prolazi kroz sve **čvorove** grafa tačno jednom.

- Problem pronalaska Hamiltonove putanje je NP-težak - nema efikasnog algoritma za rješavanje

Ojlerova putanja - putanja koja prolazi kroz sve **grane** grafa tačno jednom.

- Problem pronalaska Ojlerove putanje nije NP-težak - postoji efikasan algoritam za rješavanje



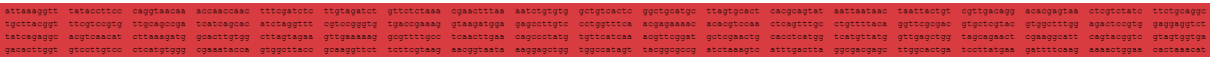
Problem rekonstrukcije niske - problem Hamiltonove i Ojlerove putanje

Hamiltonova putanja - putanja koja prolazi kroz sve **čvorove** grafa tačno jednom.

- Problem pronalaska Hamiltonove putanje je NP-težak - nema efikasnog algoritma za rješavanje

Ojlerova putanja - putanja koja prolazi kroz sve **grane** grafa tačno jednom.

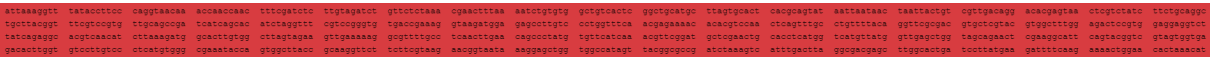
- Problem pronalaska Ojlerove putanje nije NP-težak - postoji efikasan algoritam za rješavanje
- Potrebno je drugačije definisati graf



Problem rekonstrukcije niske - problem Ojlerove putanje

Predstavljanje k-gramskog sastava na drugačiji način:

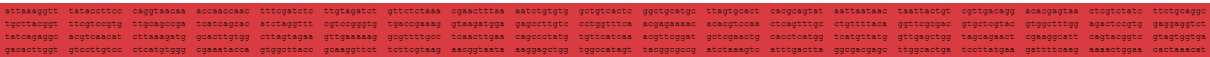
- Čvorovi grafa su k-1-grami
  - Izlazni čvor će biti prefiks k-grama
  - Ulazni čvor će biti sufiks k-grama



Problem rekonstrukcije niske - problem Ojlerove putanje

Predstavljanje k-gramskog sastava na drugačiji način:

- Čvorovi grafa su k-1-grami
  - Izlazni čvor će biti prefiks k-grama
  - Ulazni čvor će biti sufiks k-grama
- Grane su označene k-gramom



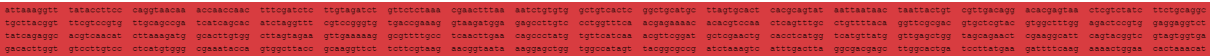
Problem rekonstrukcije niske - problem Ojlerove putanje

Predstavljanje k-gramskog sastava na drugačiji način:

- Čvorovi grafa su k-1-grami
  - Izlazni čvor će biti prefiks k-grama
  - Ulazni čvor će biti sufiks k-grama
- Grane su označene k-gramom

Slika: Grafovska reprezentacija za TAATGCCATGGGATGTT

3-grami su grane, 2-grami su čvorovi



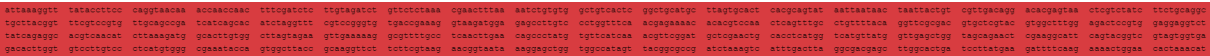
Problem rekonstrukcije niske - problem Ojlerove putanje

Predstavljanje k-gramskog sastava na drugačiji način:

- Čvorovi grafa su k-1-grami
  - Izlazni čvor će biti prefiks k-grama
  - Ulazni čvor će biti sufiks k-grama
- Grane su označene k-gramom

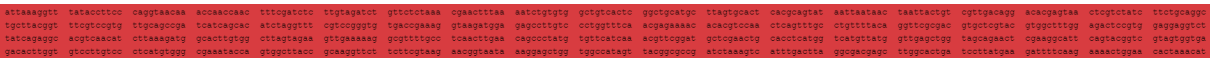
Slika: Grafovska reprezentacija za TAATGCCATGGGATGTT

Neki čvorovi imaju istu oznaku, možemo li ih spojiti?



Problem rekonstrukcije niske - problem Ojlerove putanje

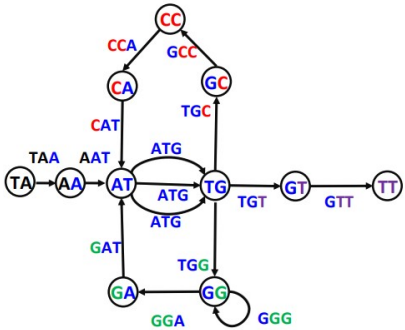
Sve čvorove koji imaju istu oznaku je moguće spojiti u jedan čvor, zadržavajući sve grane. Takav graf se naziva **De Bruijnov graf**.



Problem rekonstrukcije niske - problem Ojlerove putanje

Sve čvorove koji imaju istu oznaku je moguće spojiti u jedan čvor, zadržavajući sve grane.

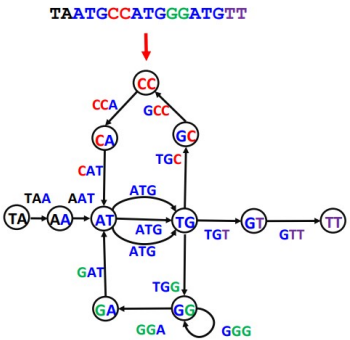
Takav graf se naziva **De Bruijnov graf**.



Slika: De Bruijnov graf za TAATGCCATGGGATGTT

Problem rekonstrukcije niske - De Brujinov graf od k-grama

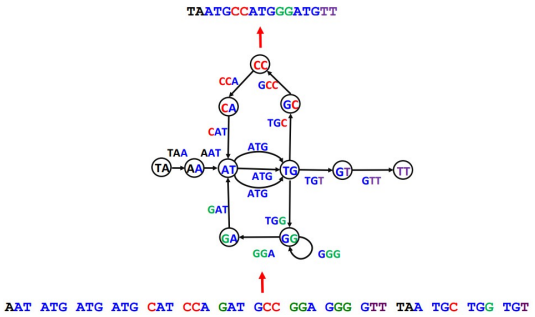
Vidjeli smo kako od zadate niske (genoma) možemo konstruisati De Brujinov graf. U realnim primjenama, potrebno je konstruisati graf od k-grama (očitavanja).



Slika: Genom $\rightarrow$ De Bruijnov graf

atttaaaggtt tataccttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc ttctgcaggc
tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt cgttccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac acacgtccaa ctcagtttgc ctgttttaca ggttcgcgac gtgctcgtac gtggctttgg agactccgtg gaggaggtct
tatcgaggcc acgtcaacat cttaaagatg gcacttgtgg cttagtagaa gttgaaaaag gcgttttgcc tcaacttgaa cagcctatg tgttcatcaa acgttcggat gctcgaactg cacctcatgg tcatgttatg gttgagctgg tagcagaact cgaaggcatt cagtacggtc gtagtggtga
gacacttggt gtccttgtcc ctcatgtggg cgaaatacca gtggcttacc gcaaggttct tcttcgtaag aacggtaata aaggagctgg tggccatagt tacggcgccg atctaaagtc atttgacttta ggcgacgagc ttggcactga tccttatgaa gatttccaag aaaactggaa cactaaacat

Problem rekonstrukcije niske - De Brujinov graf od k-grama

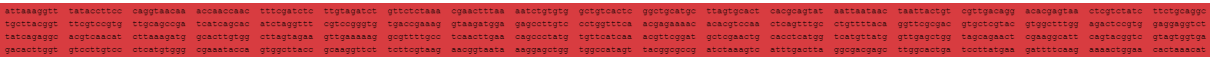


Slika: Očitavanja → De Bruijnov graf → genom

Problem rekonstrukcije niske - De Brujinov graf od k-grama

Postupak:

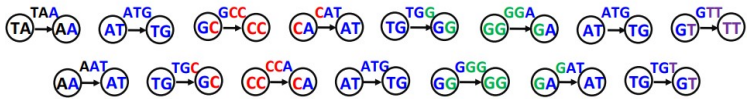
- Za svaki k-gram kreiramo dva čvora i jednu granu, jedan čvor predstavlja prefiks, a drugi sufiks k-grama



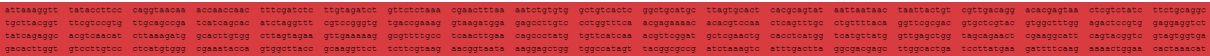
Problem rekonstrukcije niske - De Brujinov graf od k-grama

Postupak:

- Za svaki k-gram kreiramo dva čvora i jednu granu, jedan čvor predstavlja prefiks, a drugi sufiks k-grama



- Čvorove sa istom oznakom spajamo u jedan dok ne dobijemo graf sa jedinstvenim čvorovima



Problem rekonstrukcije niske - De Brujinov graf od k-grama

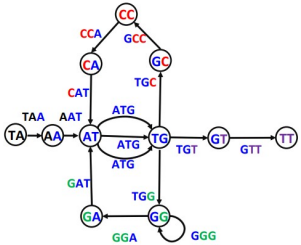
Postupak:

- Za svaki k-gram kreiramo dva čvora i jednu granu, jedan čvor predstavlja prefiks, a drugi sufiks k-grama
- Čvorove sa istom oznakom spajamo u jedan dok ne dobijemo graf sa jedinstvenim čvorovima

Problem rekonstrukcije niske - De Brujinov graf od k-grama

Postupak:

- Za svaki k-gram kreiramo dva čvora i jednu granu, jedan čvor predstavlja prefiks, a drugi sufiks k-grama
- Čvorove sa istom oznakom spajamo u jedan dok ne dobijemo graf sa jedinstvenim čvorovima



Slika: De Bruijnov graf za TAATGCCATGGGATGTT

Ojlerova teorema

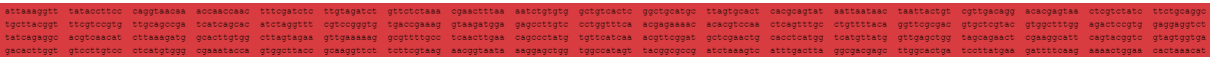
Svaki Ojlerov graf je balansiran.

Svaki povezan graf i balansiran graf je Ojlerov.

Povezan graf je graf u kojem postoji put između svaka dva čvora.

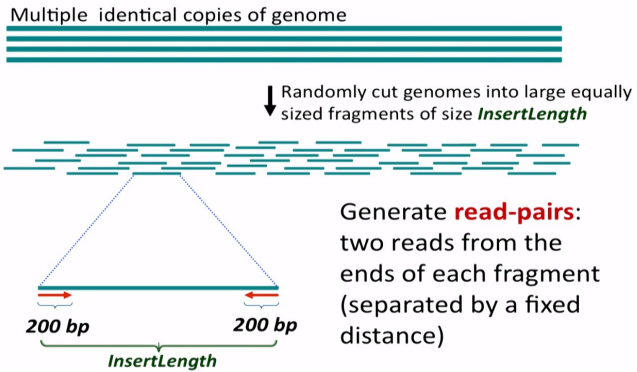
```
EULERIANCYCLE(Graph)
    form a cycle Cycle by randomly walking in Graph (don't visit the same edge twice!)
    while there are unexplored edges in Graph
        select a node newStart in Cycle with still unexplored edges
        form Cycle' by traversing Cycle (starting at newStart) and then randomly walking
        Cycle ← Cycle'
    return Cycle
```

Slika: Pseudokod algoritma EulerianCycle



Sastavljanje parova očitavanja

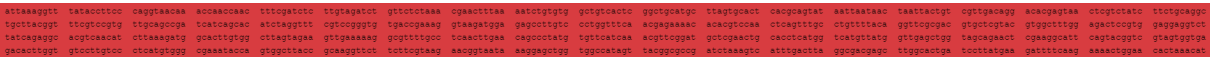
- **Problem:** Postojanje više Ojlerovih putanja u De Brujinovom grafu
- **Rješenje:** Sekvenciranje sa parovima očitavanja



Sastavljanje parova očitavanja

Upareni k-gram predstavlja par k-grama na fiksiranom rastojanju d u genomu.

Slika: TCA i TCC na rastojanju d=11 čine jedan upareni 3-gram



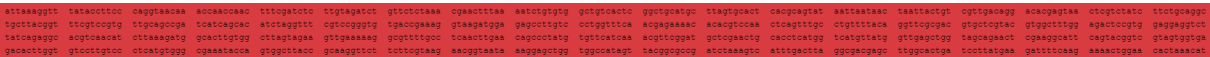
Sastavljanje parova očitavanja

Upareni k-gramski sastav *PairedComposition$_d$ (Text)* se sastoji od svih k-grama niske *Text* i njihovih parova.

PairedComposition(**TAATGCCATGGGATGTT**)

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

Slika: PairedComposition$_3$ (TAATGCCATGGGATGTT)

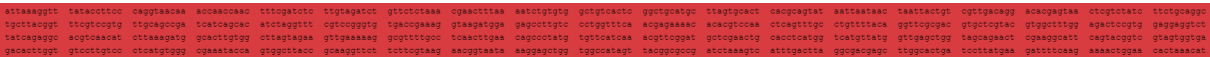


Sastavljanje parova očitavanja

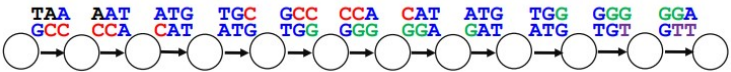
Slika: Leksikografski poredak kolekcije PairedComposition$_3$ niske TAATGCCATGGGATGTT

Problem rekonstrukcije niske na osnovu parova očitavanja:

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

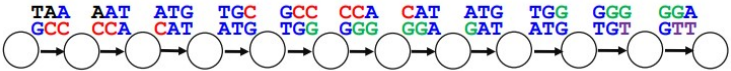


Sastavljanje parova očitavanja

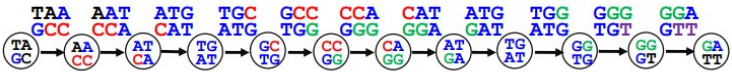


Slika: Genom TAATGCCATGGGATGTT predstavljen kao putanja

Sastavljanje parova očitavanja

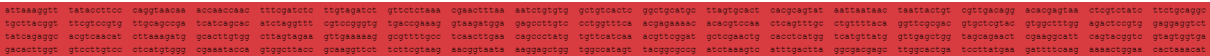


Slika: Genom TAATGCCATGGGATGTT predstavljen kao putanja

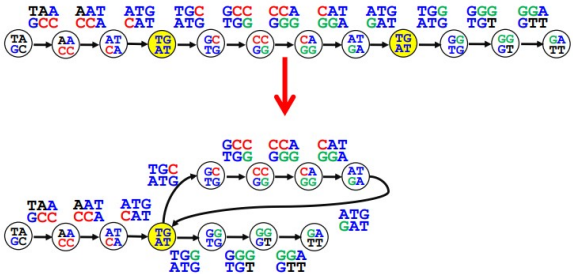


Slika: Graf koji odgovara uparenom 3-gramskom sastavu TAATGCCATGGGATGTT

Grana - upareni 3-gram, čvor - prefiks i sufiks uparenog 3-grama



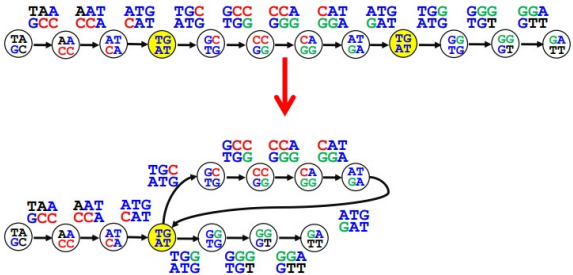
Sastavljanje parova očitavanja



Slika: Upareni De Brujinov graf za TAATGCCATGGGATGTT

Dva čvora (TG AT) su spojeni u jedan čvor

Sastavljanje parova očitavanja



Slika: Upareni De Brujinov graf za TAATGCCATGGGATGTT

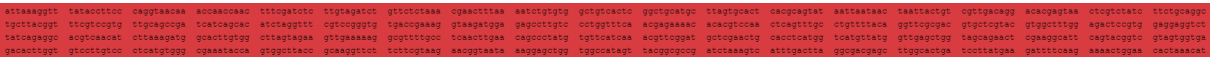
Dva čvora (TG AT) su spojeni u jedan čvor **Kako konstruisati upareni De Brujinov graf na osnovu uparenih k-grama?**

atttaaaggtt tatacctttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc ttctgcaggc
tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac acacgtccaa ctcagtttgc ctgttttaca ggttcgcgac gtgctcgtac gtggctttgg agactccgtg gaggaggtct
tatcgaggcc acgtcaacat cttaaagatg gccttgtggg cttagtagaa gttgaaaaag gcgttttgcc tcaacttgaa cagcoctatg tgttcatcaa acgttcggat gctcgaactg cacctcatgg tcatgttatg gttgagctgg tagcagaact cgaaggcatt cagtacggtc gtagtggtga
gacacttggt gtccttgtcc ctcatgtggg cgaaatacca gtggcttacc gcaaggttct tcttcgtaag aacggtaata aaggagctgg tggccatagt tacggccgcg atctaaagtc atttgacttta ggcgacgagc ttggcactga tccttatgaa gattttcaag aaaactggaa cactaaacat

Sastavljanje parova očitavanja

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

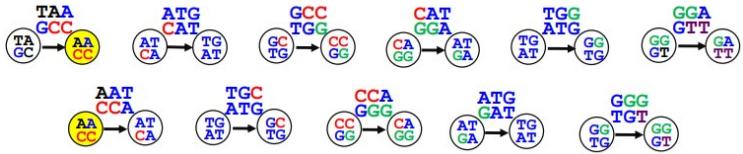
- Svaka grana je označena jednim uparenim k-gramom
- Svaki čvor je označen prefiksom/sufiksom izlazne/ulazne grane
- Čvorovi sa istom oznakom se spajaju u jedan čvor



Sastavljanje parova očitavanja

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

- Svaka grana je označena jednim uparenim k-gramom
- Svaki čvor je označen prefiksom/sufiksom izlazne/ulazne grane
- Čvorovi sa istom oznakom se spajaju u jedan čvor



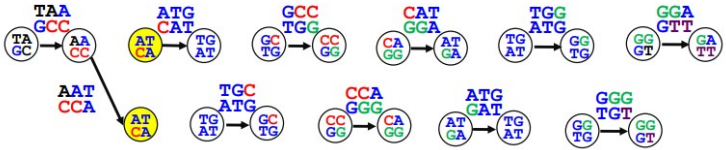
Slika: Konstrukcija uparenog De Brujinovog grafa na osnovu uparenih k-grama

atttaaaggtt tataccttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc ttctgcaggc
tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac acacgtccaa ctcagtttgc ctgttttaca ggttcgcgac gtgctcgtac gtggctttgg agactccgtg gaggaggtct
tatcgaggcc acgtcaacat cttaaagatg gcacttgtgg cttagtagaa gttgaaaaag gcgttttgcc tcaacttgaa cagcoctatg tgttcatcaa acgttcggat gctcgaactg cacctcatgg tcatgttatg gttgagctgg tagcagaact cgaaggcatt cagtacggtc gtagtggtga
gacacttggt gtccttgtcc ctcatgtggg cgaaatacca gtggcttacc gcaaggttct tcttcgtaag aacggtaata aaggagctgg tggccatagt tacggcgccg atctaaagtc atttgacttа ggcgacgagc ttggcactga tccttatgaa gattttcaag aaaactggaa cactaaacat

Sastavljanje parova očitavanja

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

- Svaka grana je označena jednim uparenim k-gramom
- Svaki čvor je označen prefiksom/sufiksom izlazne/ulazne grane
- Čvorovi sa istom oznakom se spajaju u jedan čvor



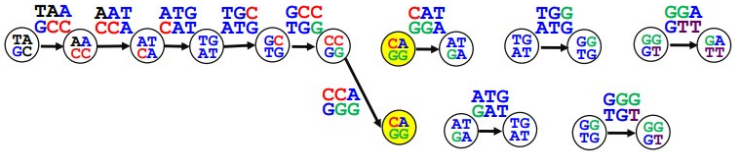
Slika: Konstrukcija uparenog De Brujinovog grafa na osnovu uparenih k-grama

atttaaaggtt tataccttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc ttctgcaggc
tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac acacgtccaa ctcagtttgc ctgttttaca ggttcgcgac gtgctcgtac gtggctttgg agactccgtg gaggaggtct
tatcgaggcc acgtcaacat cttaaagatg gcacttgtgg cttagtagaa gttgaaaaag gcgttttgcc tcaacttgaa cagcctatg tgttcatcaa acgttcggat gctcgaactg cacctcatgg tcatgttatg gttgagctgg tagcagaact cgaaggcatt cagtacggtc gtagtggtga
gacacttggt gtccttgtcc ctcatgtggg cgaaatacca gtggcttacc gcaaggttct tcttcgtaag aacggtaata aaggagctgg tggccatagt tacggcgccg atctaaagtc atttgacttta ggcgacgagc ttggcactga tccttatgaa gattttcaag aaaactggaa cactaaacat

Sastavljanje parova očitavanja

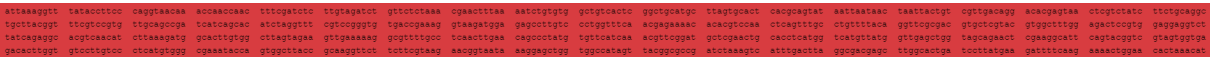
Upareni De Brujinov graf na osnovu kolekcije uparenih k-grama:

- Svaka grana je označena jednim uparenim k-gramom
- Svaki čvor je označen prefiksom/sufiksom izlazne/ulazne grane
- Čvorovi sa istom oznakom se spajaju u jedan čvor

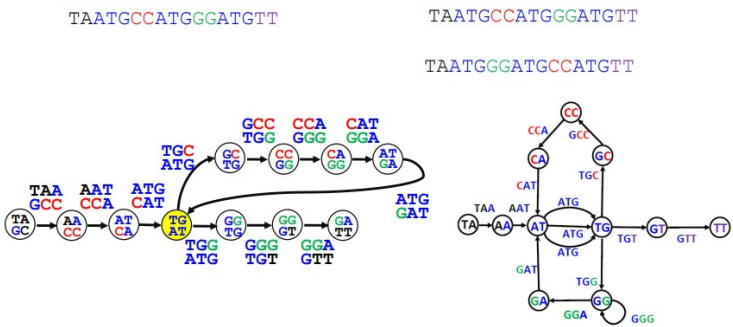


Slika: Konstrukcija uparenog De Brujinovog grafa na osnovu uparenih k-grama

Iterativnim spajanjem čvorova dobijamo upareni De Brujinov graf



Sastavljanje parova očitavanja

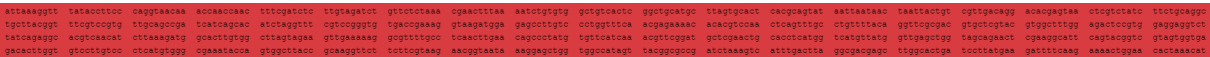


Slika: Poređenje uparenog De Brujinovog grafa i običnog De Brujinovog grafa

Lijevo - jedinstvena rekonstrukcija genoma, desno - više mogućih rekonstrukcija genoma

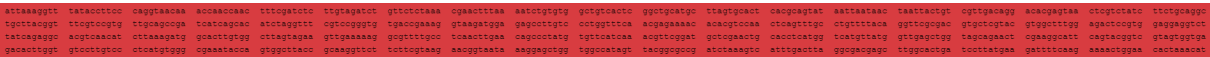
U realnosti

Do sada smo pretpostavljali da su očitavanja savršena i da ne sadrže greške. Ali, u ralnosti…



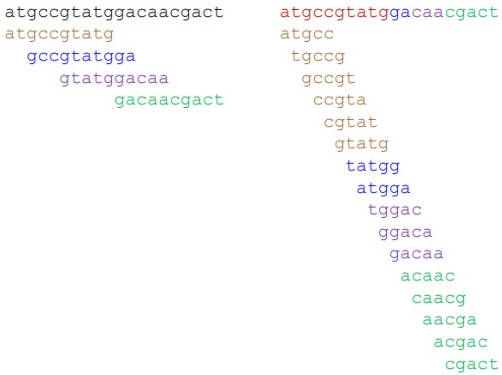
U realnosti

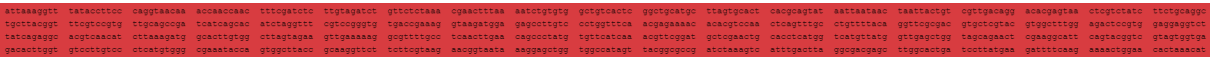
- Očitavanja **sadrže** greške



U realnosti

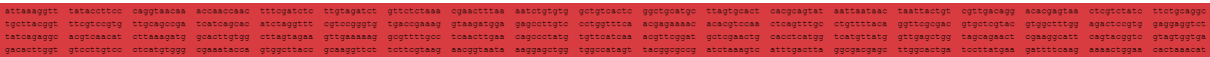
- Očitavanja **sadrže** greške





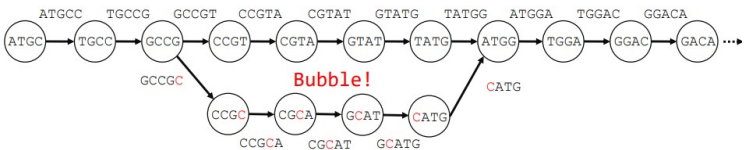
U realnosti

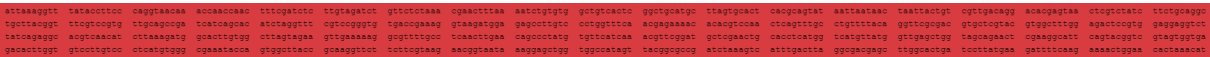
- Očitavanja **sadrže** greške
- Rastojanja između očitavanja u parovima **nisu tačna**



U realnosti

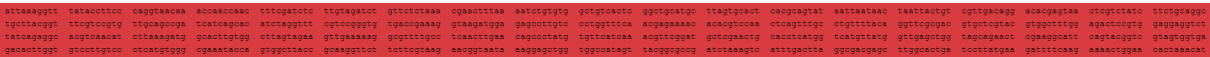
- Očitavanja **sadrže** greške
- Rastojanja između očitavanja u parovima **nisu tačna**





U realnosti

- Očitavanja **sadrže** greške
- Rastojanja između očitavanja u parovima **nisu tačna**
- **Nesavršena** pokrivenost genoma očitavanjima



Hvala na pažnji!

Reference

- Phillip Compeau, Pavel Pevzner, Bioinformatics Algorithms 2nd Edition, 2015.
- Bioinformatics Algorithms video predavanja
- Predavanja iz kursa Uvod u bioinformatiku, prof. dr Jovana Kovačević - MATF